THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

The Hong Kong Polytechnic University

Department of Computing

Web-based Data Mining and Discovery of Useful Herbal Ingredients (WD^2UHI)

Jackei Ho Kei WONG

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

December 2009

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Jackei Ho Kei WONG

ABSTRACT

This PhD thesis is in the TCM (Traditional Chinese Medicine) area, focusing on discovering trusted and useful herbal ingredients from the enterprise angle. In the research, a conceptual framework of six essential elements is proposed, namely: i) enterprise TCM ontology; ii) automation of ontology-based system generation, directly from the given iconic specification by adhering to the Meta-Interface (MI) concept; iii) concept of "living ontology" and its "reversible" implementation support; iv) text mining of open data sources (e.g. the open web and other public knowledge repertoires) in an on-line manner; v) techniques to define the associations/relevance among various ontological entities, namely, automatic semantic aliasing and neural network; and iv) system trustworthiness attained by achieving cross-layer semantic transitivity. Therefore, we call the trusted conceptual framework, which represents the aim of this research, the *WD*^{2}*UHI* (***Web-based Data Mining and Discovery of Useful Herbal Ingredients***) platform. In fact, the trustworthiness of the platform is ensured throughout the research process because all the prototypes at different stages are verified in the clinical environments, involving physicians and treatment of patients whenever appropriate. The meaning of trustworthiness or "being trusted" adheres to spirit of the RFC 2828 – Internet Security Glossary. From the above, the two objectives of this project have become transparent: i) 1st objective – the proposal and development of the trusted *WD*^{2}*UHI* platform, and ii) 2nd objective – the proposal of novel methods to discover herbal ingredients correctly and meaningfully.

The trusted WD^2UHI platform in the 1st objective involves details in two areas: i) reliable client/server communication over the mobile Internet; and ii) meaningful herbal information discovery, which must adhere to IT (Information Technology) formalisms and globally accepted TCM formal principles. Finding suitable and efficacious methods to guarantee reliable client/server communication over the mobile Internet and defining TCM formalisms for meaningful TCM discoveries require a colossal amount of work, which would exceed the time/effort constraints imposed on this PhD research. For this reason, we had to make a decision from the experience of my serious and extensive preliminary explorations. As a result, it was decided that the rest of the research energy should be focused mainly on the second objective, which itself requires a substantial amount of effort as manifested by the scale of the essential elements defined for the proposed conceptual framework.

As is mentioned above, this research covers two domains of formalisms – IT and TCM. Since my TCM knowledge is limited, I had to consult and discuss with different TCM experts (e.g. physicians including those of the YOT (Yan Oi Tong) mobile clinics that treat hundreds of patients daily in the Hong Kong SAR, and also pharmacologists from other parties), in light of applicable TCM formalisms, continuously. The research activities are organized as a *fast prototyping process*, which feeds the current useful experience to the next stage incrementally to re-orient the research direction when necessary.

The TCM formalism identified for this research is the SIMILARITY/SAME (i.e. “同”) principle (or “同病異治, 異病同治” in

classical TCM terminology). If the three different sets of prescriptions for Illness (A, a) (i.e. illness A for region a), Illness (A, b) (i.e. the same illness name for region b) and Illness (A, c) (i.e. the same illness name for region c) are PAa , PAb , and PAc respectively, by the SIMILARITY/SAME principle the total/common set of usable prescriptions for treating the three illnesses should be $P_{all} = PAa \cup PAb \cup PAc$. The $\cup$ operation (i.e. union) associates the three different sets of prescriptions into a single pool (i.e. common set P_{all}) by their common attributes/factors. In fact, the three illnesses are defined by some additional attributes on top of the common set, due to geographical and epidemiological differences.

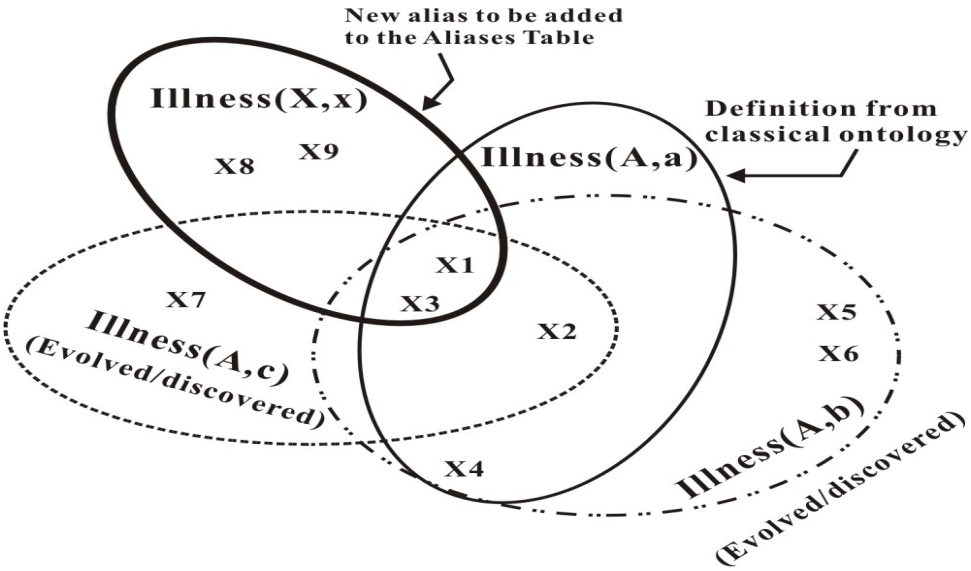


Figure a.1 Illness (X, x) is a new alias for the referential context (RC) Illness (A, a)

The generation of testing prototypes in this research is automatic and adheres to the meta-interface (MI) philosophy that was originally proposed by the Nong's Company Limited, which also allows my prototypes to be

generated/customized from its production enterprise TCM ontology core (onto-core) for real-life mobile-clinic operations. The original MI paradigm proposed by Nong's is only a conceptual "shell" containing insufficient details for implementation, but the Nong's enterprise TCM onto-core was already a production version when I started my PhD research. It is part of the PhD research endeavour to make the "shell" MI paradigm work.

The Nong's enterprise TCM onto-core (ontology core) is skeletally built from classical information already enshrined in canons, treatises, and case histories since the ancient past, via a consensus certification process. Since its resultant consensus-certified onto-core does not evolve automatically, it risks the danger of stagnating with old knowledge. The OCOE&CID (On-line Continuous Ontology Evolution and Clinical Intelligence Discovery) paradigm proposed in my PhD research is actually the advanced, implementable version of the "shell" Nong's MI philosophy. It neutralizes the danger of knowledge stagnation by opening up the closed skeletal TCM onto-core with the help of continuous on-line text mining and automatic semantic aliasing (ASA). The ASA weights the similarity between two terms (e.g. Ter_1 and Ter_1). $Ter_1 = Ter_2$ means that the two terms are synonyms or logically the same. In the logical expression, which is an IT formalism, $P(Ter_1 \cup Ter_2) = P(Ter_1) + P(Ter_2) - P(Ter_1 \cap Ter_2)$, the symbols $\cup$ and $\cap$ stand for union and intersection respectively. If Ter_1 and Ter_1 are only similar, $Ter_1 \neq Ter_2$ is logically true; they are then aliases (not synonyms). $P(Ter_1 \cap Ter_2)$ represents the degree of similarity (probability) between Ter_1

and Ter_1. $P(Ter_1)$ and $P(Ter_2)$ are probabilities for the multi-representations (other meanings).

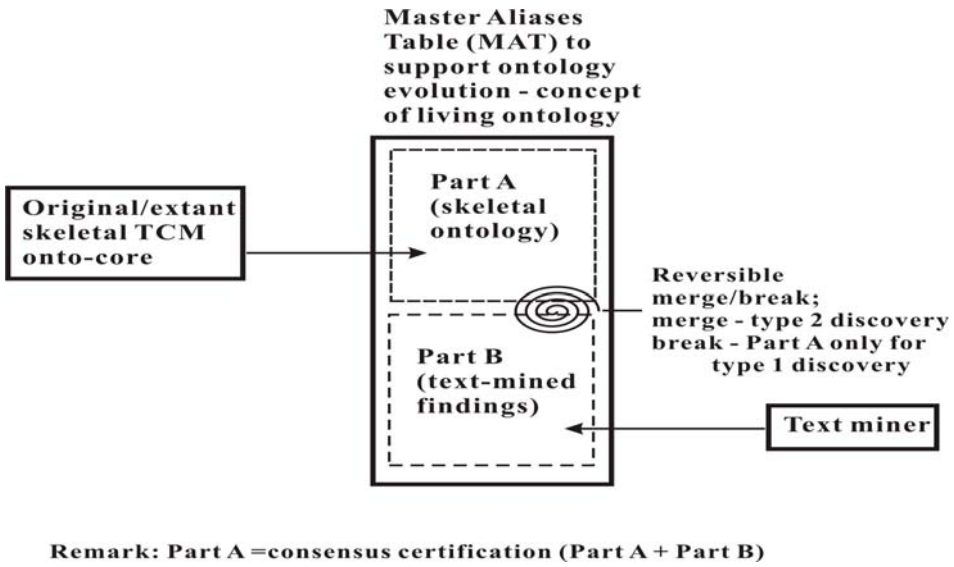


Figure a. 2 Type 1 and Type 2 discoveries

In this research there are two types of discoveries conceptually: i) Type 1 – if the discovery is ***not within*** the context of the extant skeletal ontology (i.e. Part A in Figure a.2); and ii) Type 2 – if the discovery is within the current context of the working ontology (i.e. "Part A and Part B" together). Type 1 is considered as "high-level" and Type 2 "low-level". In fact, any discovery is determined on the relevance index, which can be computed by the ASA mechanism or the NN (neural network) backpropagation approach. The NN approach is particularly suitable for Type 2 discovery of individual herbal ingredients. Since the NN named module is trained only with the prescribed dataset, training is ***considered completed*** in the context of Type 2 discovery, as long as the NN has learned all the "***current knowledge***" intertwined and

embedded in the current training dataset. It produces potential discoveries, which should be later decided upon by TCM domain experts.

The solutions proposed in my PhD research have contributed to 16 publications so far. All the stated PhD research objectives have been achieved. The research has also uncovered many relevant problems, which should be resolved in the future work including: i) reducing the ANT value defined by $ANT = \sum_{j=1}^{k\to\infty} jP_j \approx \frac{1}{(1-\delta)}$, where δ is the channel error probability, in order to parallelize a very large data base (VLDB), such as a sizeable TCM ontology, for fast system response, and ii) identifying other formal TCM principles to facilitate more effective discovery of herbal ingredients, with the help of practicing TCM physicians and domain experts.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I thank my Ph.D. thesis Chief Supervisor Dr. Allan Kang Ying WONG and Industrial Co-supervisor Dr. Wilfred Wan Kei LIN for their incessant encouragement and guidance, and Prof. Tharam S. DILLON for his enlightenment for the research. I heartedly appreciate the frequent clerical help by Ms. Miu TAI of the Department of Computing (General Office). I am grateful to the Visiting Scholar Appointment by the Digital Ecosystems and Business Intelligence Institute (DEBII) of the Curtin University of Technology, from 13th February 2009 to 13th March 2009. During the DEBII visit I had the chance to discuss the importance of my Ph.D. research with international scholars. Last but not least, I thank PuraPharm's formal agreement (APPENDIX I) for me to submit this thesis and its funding of this TCS Ph.D. project (APPENDIX XI).

Jackei Ho Kei WONG

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

WD^2UHI – Web-based Data Mining and Discovery of Useful Herbal Ingredients, the name of this PhD research project.

***MI** – Meta-interface, a conceptual prelude to the implemented EOD-ISD paradigm; it was first proposed by the Nong's without any implementations or verifications. It was proposed with the intention to cut down the amount of software errors introduced inadvertently in the Waterfall process.*

***CTSS** – Customized Telemedicine Software System is a product of the automatic software system generation approach (e.g. MI).*

***WTS** – Web-based Telemedicine System, for which the Nong's D/P system deployed on the YOT (Yan Oi Tong) mobile clinics is a real-life example that treats hundreds of patients daily in the Hong Kong SAR.*

***EOD-ISD** – Enterprise Ontology Driven Information System Development, a novel software engineering approach proposed and implemented in this research.*

***D/P system** – Diagnosis/Prescription system.*

***XML** – Extensible Mark-up Language - a member of the XML, RDF and OWL family of metadata systems proposed by the W3C.*

RDF *– Resource Description Framework (a W3C metadata model).*

OWL *– Web Ontology Language (a W3C metadata model).*

W3C *– World Wide Web Consortium.*

STV *– Semantic Transitivity Visualizer, which is a novel means proposed in this thesis to verify the cross-layer semantic transitivity in a TCM ontology.*

OCOE&CID *– On-line Continuous Ontology Evolution and Clinical Intelligence Discovery.*

ASA *– Automatic Semantic Aliasing, which is a novel mechanism proposed in this thesis to support Type 1 discovery as the primary objective.*

ESE *– Enterprise Standard Extension.*

RI *– Relevance Index; this indicates the relevance/similarity of an entity to the referential host/entity.*

MAT *– Master Aliases Table, which is the main data structure to support real-time ontological evolution, and the important elements include different referential entities for which each has a list of similar elements/entities that their degree of similarity is indicated by the specific RI.*

Trusted communication *– According to RFC 2828 – Internet Security Glossary, 2000, ftp.isi.edu/in-notes/rfc2828.txt, system trust is defined by many parameters depending on the domain. For example, in e-business, communication dependability and fast response are two factors of public trust of a system – trusted communication.*

Trust/trustworthiness *– Under the same umbrella as RFC 2828.*

Ontology *– It is the explicit specification of a conceptualization; 3 conceptual levels; 1st top level – pure philosophy of a domain and not 100% implementable in every case; 2nd local enterprise level – facts in a domain (a subset of the 1st level) that suit the operation of an enterprise and is therefore implementable; 3rd local-of-the-local level – variants customized from a predefined enterprise ontology.*

Ontology categories *– Two main categories: i) consultative - for user to familiarize with the domain including facts and conjectures; ii) practical (e.g. TCM clinical practice) – practice with adherence to facts only.*

Ontology implementation *– 3-layer architecture with cross-layer semantic transitivity: top layer for human manipulation - the query system to reflect exactly the middle layer; middle layer for machine understanding and execution (parsing) - the semantic net to exactly reflect the bottom semantic ontology; bottom layer for representing the domain knowledge – concepts, lexicon*

(vocabulary) and their relationships are represented in a subsumption hierarchy in an axiomatic manner.

Semantic transitivity *– If any item is picked from any layer of the 3-layer architecture of an ontology system, its matched forms in the other two layers will surface correctly and consistently.*

Axioms *– Rules upon which a logical system is defined (e.g. probability axioms).*

IT axioms *– Rules to be applied on the IT domain; for example, defining the degree of similarity between* P_1 *and* P_2 *by* $P_r(P_1\ P_2)$ *in* $P_r(P_1 \cup P_2) = P_r(P_1) + P_r(P_2) - P_r(P_1P_2)$.

TCM axioms *– Rules/principle to be applied on the TCM domain; for example the SIMILARITY/SAME principle.*

SAME principle *– It is a formal diagnostic principle enshrined in the TCM classics: "同病異治, 異病同治"; SAME and "同" are translational synonyms. In English this principle is defined as: "If the symptoms are the same or similar, different conditions could be treated in the same way medically, independent of whether they come from the same illness or different ones [WHO07]."*

RA *– Retrieval Algorithm, which is working by predicate logic - if "a" is true then "b".*

Type 1 discovery – *If set P is not part of the knowledge base K (i.e.* $P \notin K$ *) but it does possess all the attributes of one of classes in the predefined class set CL for K, then logically it is "*$< P \in CL > \wedge < P \notin K >$*", where* $\wedge$ *is the logical "AND". This means that P is a new occurrence (Type 1 discovery) with respect to the extant K. The intent of Type 1 discovery is dealing with open knowledge sources such as the open web.*

Type 2 discovery – *For the population* $\Omega = \{s_i\}$ *and* $i = 1,2,..., j,..,k,...,n$ *, the following are true logically:* $CL_j = (s_j)$ *;* $CL_k = (s_k)$ *;* $CL_j \neq CL_k$ *; and* $< CL_j, CL_k > \in K$ *. The expression* $CL_j = (s_j)$ *says that the class* CL_j *is defined by the set of attributes included in the set* s *, where the subscript* j *(i.e.* s_j*) marks the particular element in* s *. The* $P_r(CL_j CL_k) = \Theta$ *expression indicates that the two classes* CL_j *and* CL_k *are logically independent. For example, if both of* $\{< X \in CL_j > \wedge < X \in CL_k >\}$ *and* $X \in K$ *conditions are logically (*$\wedge$ *for logical AND) satisfied by the set X, then X is considered a discovery if either or both of the* $< X \in CL_j >$ *or/and* $< X \in CL_k >$ *association(s) was no listed in K previously and explicitly. This discovery is not Type 1, for* $X \in K$ *and only an intrinsically hidden association in K has just been uncovered. If the condition,* $P_r(UV) = P_r(U \cap V) = P_r(U) + P_r(V) - P_r(U \cup V)$ *holds logically for* $U = CL_j = (s_j)$ *and* $V = CL_k = (s_k)$*, conceptually* $P_r(UV)$ *or* $P_r(U \cap V)$ *is the relevance index (RI) that shows the degree of similarity between U and V.*

Trustworthiness *– A system is trusted because it meets the predefined level of performance. According to RFC 2828 – Internet Security Glossary, 2000, ftp.isi.edu/in-notes/rfc2828.txt, system trustworthiness is defined by the set of chosen parameters with respect to the application domain.*

Consensus certification *– It is a process to agree upon a concept, the semantics of an entity, or a set of procedures by a sufficient number of experts from a domain or community. The agreement becomes the standard vocabulary or lexicon of that domain or community. In effect, this vocabulary is the communal ontology that embeds the body of knowledge to be passed on or adhered to during practice.*

Episodes *– They are the number of rounds or iterations in the NN training process, with the same set of input parameters for better convergence.*

Accuracy *– It is the chance of yielding the correct answer by the NN.*

Weighting *– The input parameters to the NN can be weighted differently (and normalized) or have equal weights (i.e. uniform and normalized).*

Supervised training *– The NN is of the form* (x, y)*, where* x *and* y *are the input and the labelled output or class respectively [Callan03].*

Unsupervised training *– It is the training of a NN without a target value or class or label [Coppin04].*

Sensitivity analysis *– It is the study of how the value change in a variable* x *would affect the output of the associated function, namely* $f(x)$*. In real-life applications the resultant change in* $f(x)$ *may invoke some necessary actions, which are defined with respect to the problem domain.*

RMSE *(Root Mean Square Error) - If* O *is the conceptual predicted output,* m_i *the measured output of a process at the* i^{th} *cycle, then* $RMSE = \sqrt{\sum_{i=1}^{N} (O - m_i)^2}$ *for* $i \leq N$*.*

Key References

[Callan03] R. Callan, Artificial intelligence, Palgrave Macmillan, 2003

[Coppin04] B. Coppin, Artificial Intelligence Illuminated, Jones and Bartlett Publishers, Incorporated, 2004

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

Thesis Preamble

This research project is entitled "***Web-based Data Mining and Discovery of Useful Herbal Ingredients (*** *WD^2UHI* ***)***" and is applied to the area of Traditional Chinese Medicine (TCM). The aim is to discover useful TCM herbal ingredients as well as prescriptions. Therefore the objectives include:

a) Propose a conceptual framework to support trusted and meaningful web-based data mining.
b) Propose methods to achieve two types of discoveries, namely Type 1 and Type 2.
c) Propose principle(s) to which the discoveries can adhere and stand the scrutiny in the process of consensus certification.

A system is trusted (or its trustworthiness) only it has met the predefined performance indicators in a consistent manner. This notion is indicated clearly by the RFC 2828 – Internet Security Glossary, 2000, (ftp.isi.edu/in-notes/rfc2828.txt). If the conceptual framework is not a trusted entity, then the discoveries made within is not trustworthy.

Web-based data mining is to plough the open sources, including the web, for desired information. The term "web-based" in this project has the following connotations:

i) The telemedicine clinical environment (i.e. the deployed Nong's D/P (diagnosis/prescription) system for mobile clinics) in which I would carry out my verifications is web-based

ii) The MI paradigm is a novel concept whereby ontology-based (i.e. semantic) D/P systems should be automatically generated/customized and installed in the specified websites (the Internet web) in a remote manner by the clients

iii) The "living ontology" mechanism (LOM) is web-based

iv) The text mining process, which is an essential part of the LOM, is web-based

Other technical notions outside the web-based scope defined above are not essential part of this research at the present (e.g. data cleaning and aggregation).

The intent of the web-based data mining is to enrich the extant knowledge base so that it can keep abreast of contemporary scientific findings world wide. Therefore, it is an important step in the evolution of a knowledge base. Yet, a meaningful TCM (Traditional Chinese Medicine) knowledge base should be the domain ontology or at least the sub-ontology. This means that the knowledge contained within this base is construed as standard vocabulary by TCM practitioners/experts across the globe. Every ontology or sub-ontology is created by a rigorous consensus certification process that always involves a sufficient number of experts, who would agree upon the ontological contents by

adhering to classical information, formal scientific reports, treatise and case histories.

In this research there are two types of discoveries conceptually: i) Type 1 – if the discovery is ***not within*** the context of the extant ontology; and Type 2 – if the discovery is within the current context of the extant ontology. Type 1 is considered as "high-level" and Type 2 "low-level".

A discovery is valid in the domain context only if it satisfies the predefined principle(s) in the domain. Yet, if this discovery can be incorporated into the domain ontology, it is the decision of the consensus certification process. In this sense, the discoveries made in the WD^2UHI are only syllogistically valid because their eventual inclusions into the original ontology are outside the scope of my PhD research. Such inclusions make the original ontology evolve into a new version. Yet, within the proposed conceptual framework WD^2UHI , new discoveries are recorded and they may be referenced in a real-time manner. In this way the new discovery records provide the necessary data in the subsequent consensus certification process that would push the ontology forward in an evolutionary sense.

In reality every prescription contains several herbal ingredients in different capacities/roles: *principal*, *courtier*, *assistant* and *messenger*. The discoveries are useful and meaningful only if the following issues are successfully addressed: i) the TCM knowledge base K is axiomatic and unambiguous; ii) K embeds correct TCM semantics; iii) K will automatically

evolves toward the global conceptualization (or hierarchy) of medical/clinical semantics – the ideal ontology; iv) the software system (i.e. parser) can deal with the ontological hierarchy correctly in the sense that every query to be parsed has the corresponding semantic path in the hierarchy; and v) all the *WD*^{2}*UHI* prototypes (i.e. interim and final) will be verified in the real telemedicine clinical environment.

Therefore, the *WD*^{2}*UHI* platform should be a semantic (ontology-based) web-based system, which is characterized, conceptually, by the following:

a) 3-layer architecture: The bottom layer is the TCM ontology; the middle layer embeds the semantic net, which is the machine-understandable form of the ontology, and includes the parser to process it; the top layer, which is a query system that abstracts the semantic net, for human understanding and manipulation.
b) An unambiguous and axiomatic TCM ontology core (onto-core): This is a local/enterprise view (or subset) carved out of the global TCM ontology, which is made up of all the available TCM classics, treatises, case histories, and new scientific findings, by the process of consensus certification.
c) On-line evolution: The platform should absorb new knowledge from the open sources continuously and in a real-time manner. Yet, the absorbed knowledge is only a temporary extension of the enterprise TCM onto-core of the running system (or prototype), for permanent knowledge

absorption must be vetted by the consensus certification process instituted by the system owner or enterprise afterward.

d) Ontology-based software engineering: This approach would help produce the correct *WD²UHI* prototypes in a single step from the named TCM onto-core given as the basis. As a result, the semantic transitivity among the three system layers (i.e. bottom ontology, middle semantic net and upper query system) in a prototype can be guaranteed. The high-level primordial concept of semantically based or ontology-based software engineering was first proposed by Nong's Company Ltd. (Nong's) without the appropriate details for implementation. The primordial concept is called the MI (meta-interface) paradigm, which argues that the target system should be generated from the given ontology in one step for the given MI specification. Nong's went on to transform its master TCM knowledge base, which had been supporting its deployed mobile-clinic (MC) TCM telemedicine systems successfully, into the proprietary master/enterprise TCM ontology core (onto-core). The preliminary investigation of thesis found that the MI approach, if perfected, could help generate correct prototypes quickly for verifying the methods proposed in this thesis for herbal discoveries. Therefore, research effort was dished out to make the MI paradigm work for this research so that, with permission from Nong's, all my verification experiments could be carried out in the real TCM telemedicine environment. This research effort was focused on providing missing elements from the original primordial MI paradigm proposed by Nong's, including: *cross-layer semantic transitivity,*

automatic software generator, master library of icons, metadata impact on clarity of ontology representation, living ontology (i.e. on-line ontology evolution to support the MI mechanism), standard keyed-in ontological information versus non-standard hand-written information, and support for meaningful discoveries. These elements will be explained later in detail.

This PhD research is a TCS (Teaching Company Scheme) project agreed between the Hong Kong Polytechnic University and the Nong's Company Ltd., which is a subsidiary of the Hong Kong PuraPharm Group. For this reason all the *WD*^{2}*UHI* prototypes can be verified in the Nong's mobile-clinic (MC) environment, with permission. That is, the knowledge bases or ontology cores of these prototypes are derived from the Nong's master/enterprise TCM onto-core that supports all its MC clinical operations. In the verification processes, physicians will be invited to actually use the prototypes to help treat patients.

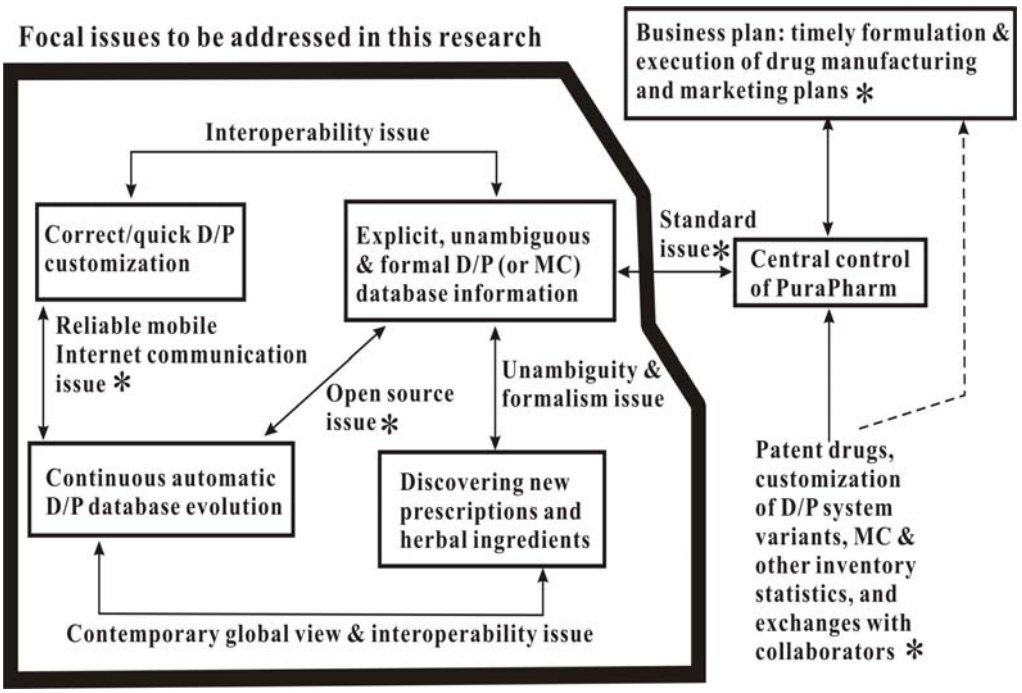


Figure pa.1 Ambit of PuraPharm's e-business platform and focal research issues (* marks issues that require reliable mobile Internet communication support)

The Nong's MC system is a computer-aided diagnosis/prescription (D/P) setup used by physicians. Since its deployment a few years ago it has been successfully used in the treatment of hundreds of patients daily in the Hong Kong SAR. The original system, which is now deployed, however, has the following shortcomings:

1. Its original TCM knowledge base is not "ontology-based" even though it was created as an enterprise telemedicine standard by a process of "*pruning and agreeing*" among a large number of invited domain experts. The knowledge base is derived from time-honoured TCM classics and not equipped to automatically evolve. As a result it

stagnated with the status quo – extant knowledge ingrained by the "*pruning and agreeing*" operation, a form of consensus certification. Yet, this status quo is the enterprise-wide TCM clinical telemedicine "standard".

2. Customization of Nong's D/P variants for individual clients means copying the ***complete*** original MC knowledge (i.e. from the proprietary "master database" as it is usually referred to). Therefore, any customization acts as simply a repeat of the traditional Waterfall software engineering procedure for developing the original MC D/P system in an algorithmic manner. Therefore, any manual changes to extant D/P variants to satisfy new requirements would inevitably be prone to programming errors. An added problem is that it is not easy to debug the D/P software, for the D/P system is pervasively distributed (i.e. with intrinsic parallelism) in nature; it was designed for the mobile Internet.
3. The original design of the *graphical user interface* (GUI) records hand-written diagnostic information only without any computer-aided interpretation or logical manipulations. If the handwritten information is not an element in the Nong's enterprise "standard database", it is recorded as a fault of temporary interest only and treated as disposable garbage. In the MI paradigm, the Nong's management wants to treat such faults as potentially re-absorbable knowledge to enrich the enterprise TCM ontology core instead – a very different philosophy from the original D/P. This can be achieved by the technique of *automatic semantic aliasing*.

4. The hierarchical relationships among different information entities in the database are incompletely defined. They are intrinsically for "the eyes of the beholders", who could have varied, impromptu interpretations. Thus, the same entities and the intertwined associations among them may have different, sometime conflicting, ambiguous interpretations. Such conflicts would lead to ambiguous conceptual representations of the same phenomena, therefore hindering: i) interoperability among D/P variants; ii) their future D/P system improvement and development, and iii) local software patching jobs. In this research we call conflicting semantics for the same entity ***multi-representation***. For this reason the Nong's management proposed the primordial MI concept and transformed the extant knowledge base into the corresponding TCM ontology core. This is the Nong's master/enterprise TCM onto-core for telemedicine clinical practice. This onto-core knowledge is annotated by the XML metadata system for clarity and flexibility.
5. Medical discovery is difficult with the extant Nong's master MC database or its down-sized variants because: i) the hierarchical associations among entities within them are not sufficiently formal; and ii) it does not evolve automatically by absorbing new TCM-related scientific reports. In this research, medical discovery has two main connotations: i) new prescriptions versus the computer-aided MC diagnosis (i.e. illness); and ii) new herbal ingredient(s) versus the MC diagnosis. Diagnosis is the act of concluding an illness for the given set of symptoms. In reality, an herbal ingredient, which treats a specific set of symptoms, could be a new scientific finding. This is logically true if

it has never been used in old prescriptions or recorded in the TCM classics and past case histories. A prescription/herbal discovery can have two views: i) *local view* – the discovered items are absent from the "static enterprise TCM standard database"; and ii) *global view* – the use of discovered items is absent from the TCM classics. Logically the local view is based on the knowledge that has been consensus-certified only within the enterprise – only a subset of the whole knowledge domain. In contrast the global view is based on the whole knowledge domain, which is consensus-certified by the whole world-wide TCM community continuously. In this research, the Nong's master/enterprise TCM onto-core provides the basis for the experiments.

6. Unreliable mobile-Internet-based communication hinders successful execution of the overall PuraPharm business plan. The MC operation, which administers herbal medicine, provides crucial field information to what seasonal drugs or herbs that the company should manufacture to reap the maximum financial benefits from the forces of supply and demand. In fact, the MC deployment and Nong's operation are part of the PuraPharm e-business drive. The MC prescription data provides necessary front-line statistics to help PuraPharm timely orient its drug manufacturing and marketing strategies to outperform competitors. In reality, the PuraPharm's clients include different sales outlets (public ones and company's own setups), hospitals, private clinics, outsourced manufacturing sub-contractors, material suppliers, and other collaborators of varied reasons. PuraPharm's e-business, similar to other industrial sectors, thrives on happy and return customers. One way to

make clients happy is to ensure effective communication and fast service response. This requires the communication for the e-business infrastructure over the mobile Internet to be reliable, even under massive client mobility. In e-business setups, clients can electronically interact with a service provider, anytime and anywhere, via small-form-factor (SFF) device (e.g. mobile phone and PDA). Although the preliminary investigation of the thesis had studied the issue of mobile Internet communication carefully, the subsequent focus of the research has excluded any deeper investigation in this direction. The reason is that the time required for deeper investigation into the issue of trusted communication would push my PhD research beyond the time constraints and resources allowed for my PhD pursuit. ***Therefore, this PhD thesis focuses mainly on semantic herbal discoveries, which should be verified in the clinical environment.***

Figure pa.1 summarizes the whole of PuraPharm's e-business platform, and the focal issues within the ambit of this research are within the box surrounded by the bold line. All the issues marked by the asterisks (*) normally require reliable/trusted communications support over the mobile Internet of mixed wireless and wireline protocols. For this reason, the preliminary investigation in this research did investigate some related issues in light of reliable/trusted client/server communication. The investigation result helps define the aim and achievable objectives for the thesis more clearly. A walkthrough of Figure pa.1 is as follows:

i) *Business plan*: The central PuraPharm management and control collects all the necessary performance indicators to timely formulate the drug manufacturing process and marketing strategy. The aim is to balance the supply and demand forces in an optimal manner. The standard basis for the timely formulations is the proprietary company D/P database master, which links this research and the PuraPharm's e-business platform logically.

(***The following are the focal issues to be addressed in this PhD research***.)

ii) *Explicit, unambiguous and formal D/P or MC database*: The present D/P database is *implicit* in the sense the interpretation of some relationships among the different component entities are subject to "individual beholders" – not sufficiently unambiguous and formal. The database provides a place where data/information, however complex, are stored or retrieved as a transaction. The semantics/meaning of the transacted entity is "*explicit*" only to the eye of the interested beholder but "*implicit*" to the outsiders. That is, its semantics are *not* globally unambiguous. In contrast, an ontology-based knowledge transaction always has globally or communally unambiguous semantics. This pinpoints the need to change the implicit database operation to the explicit ontology-based operation.

iii) *Discovering new prescriptions and herbal ingredients*: The discoveries are meaningful only if the source (i.e. the master D/P database or its downsized versions) is being axiomatic, formal and unambiguous.

Again this brings out the need of having a TCM ontology for the D/P (or MC) clinical operations.

iv) *Continuous automatic D/P database evolution*: The extant D/P database is stagnated with original material input when it was first created. This input was extracted from the TCM classics by a group of domain experts in a selective manner – consensus certification. To keep the information in the database (or ontology) abreast with time, there is a need to update the material in a continuous, automatic manner – an automatic evolutionary process. The evolution provides a contemporary global view that facilitates meaningful "*inter-communal (IC)*" operations. IC operations enable various communities that practice computer-aided TCM in this world to freely and correctly exchange. These communities may operate with databases and/or ontological constructs that are created out of the TCM classics. In this research, interoperability among different D/P variants, which are customized from the same master D/P database/ontology, is the focus – "*intra-communal*" interoperability. The automatic evolutionary process must be able to search and identify new TCM-related scientific findings from open sources such as the open web. Therefore, data/text mining techniques, supported with modern technological support (e.g. mobile agents and reliable communication over the mobile Internet), are useful means.

v) *Correct/quick D/P customization*: The extant D/P system, deployed for MC clinical operation now, was developed by applying the traditional Waterfall software engineering model for algorithmic programming. This makes it hard to customize D/P system variants to satisfy different

clients who operate with specific, distinctive requirements. Customization in the Waterfall model means repeating some software engineering procedures manually. This is intrinsically error-prone and time consuming. Ensuring correct customizations of the D/P system variants naturally guarantees the interoperability among them. Therefore, automatic customization would naturally help customize correct variants from the axiomatic, unambiguous master D/P database/ontology directly; that is, a one-step formula to yield an operational target software system from the given specification by using the master D/P or MC "knowledge/database" as the generation basis. This one-step approach is logically sound for building accurate prototypes for verification in the clinical environment.

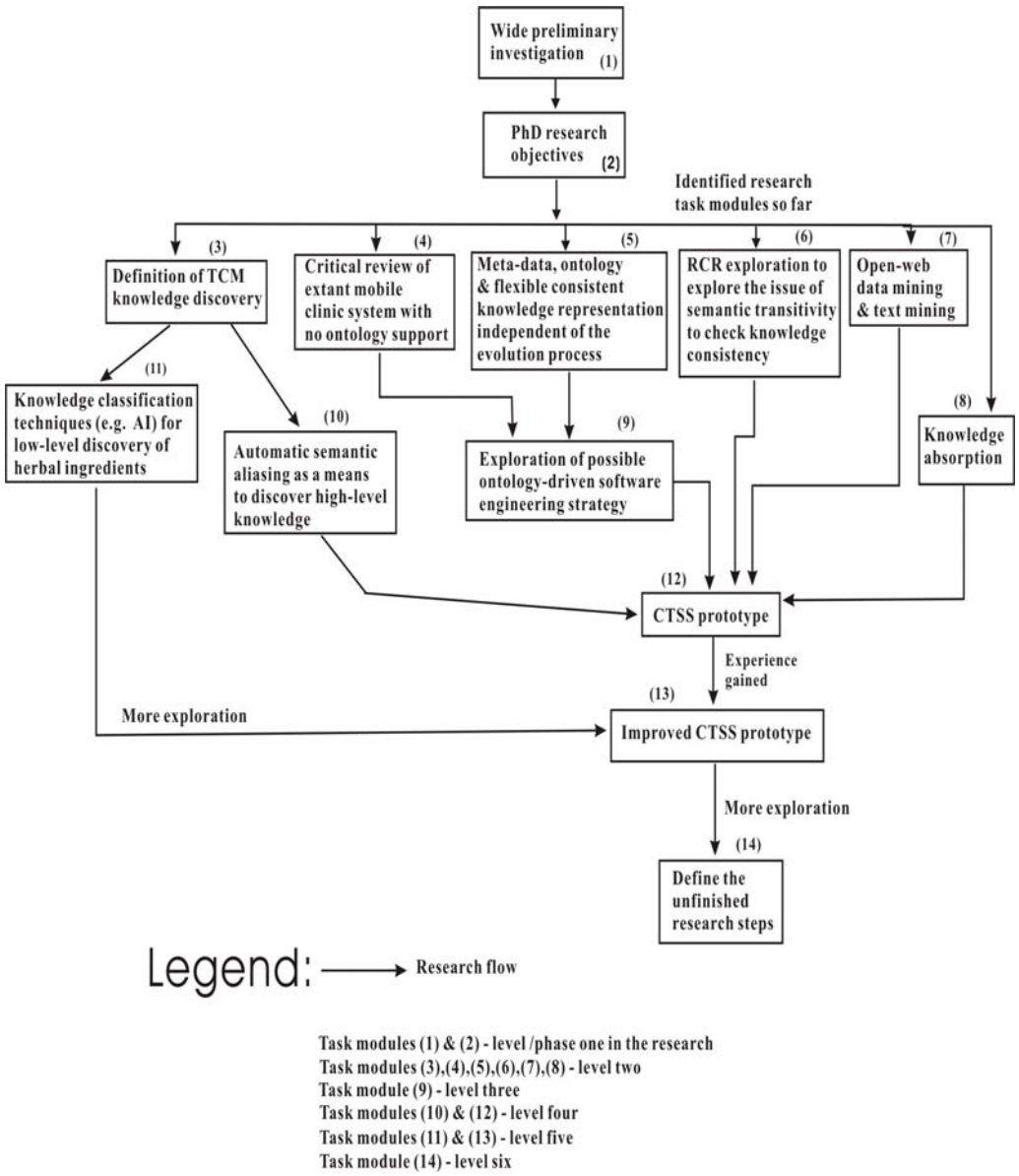


Figure pa.2 Flow of this PhD research of 14 main tasks in 6 levels

Findings from the preliminary investigations of the focal issues (those that are surrounded by the bold line in Figure pa.1) have contributed to 16 publications so far (listed in Chapter 1). These findings help streamline the research and consolidate the formulation of novel models for herbal discoveries in two levels: prescriptions, and individual herbal ingredients. A prescription normally contains several herbal ingredients in the roles of: *principal*, *courtier*,

assistant and *messenger*. Figure pa.2 summarizes concisely the flow of this research in 14 main task modules – the task roadmap. These task modules include: 1) wide preliminary investigation; 2) PhD objectives (condensed from the preliminary investigation results); 3) definition of TCM knowledge discovery (2 levels); 4) critical review of the extant MC system without ontology support (i.e. the problem of implicit semantics); 5) meta-data, ontology and flexible consistent knowledge representation independent of the evolution process (i.e. original knowledge enshrined in the extant database/ontology always remains intact – evolution in this case simply means on-line, temporary knowledge extension); 6) RCR (rehearsal with cross-referencing) exploration to study the issue of semantic transitivity to guarantee knowledge consistency; 7) the issue of open-web data mining & text mining; 8) knowledge absorption; 9) exploration of possible ontology-driven software engineering strategy (so that correct, robust prototypes can be built for experiments); 10) *automatic semantic aliasing* as a means to discover knowledge at the prescription level; 11) classification techniques to aid discovery of individual herbal ingredients; 12) CTSS (customized telemedicine software system(s)) prototype (i.e. results from the previous 11 steps are crystallized into creating the CTSS prototype); 13) improved CTSS prototype (i.e. repeated improvement from experience); 14) define the unfinished research steps (as the future work). The 14 tasks are divided into 6 levels/phases of investigations: level 1 – tasks (1) and (2); level 2 – tasks (3), (4), (5), (6), (7) and (8); level 3 – task (9); level 4 – tasks (10) and (12); level 5 – tasks (11) and (13); and level 6 – task (14).

It is important to point out here that the CTSS provides the platform for both global and local TCM discoveries at two levels: prescription (considered high-level in the context of the thesis) and individual herbal ingredients (low-level). Besides, CTSS is synonymous with MI, WTS and EOD-ISD (refer to the definition of useful terms), which represent the different stages in the course of the primordial MI evolution from being a "shell" in the beginning to a usable software development paradigm finally christened EOD-ISD.

This research, which went through the 14 task modules shown in Figure pa.2, has produced significant findings in 16 publications so far:

Invited Book Chapters:

[2009-p1] Jackei H.K. Wong, Allan K.Y. Wong, Tharam S. Dillon and Wilfred W.K. Lin, Dynamic Cache Size Tuning to Shorten Mobile Business Service to Make E-Shoppers Happy in E-Commerce ed. Aleksandar Lazinica, In-Tech Publications, Vienna, Austria (to appear)

[2008-p2] Wilfred W.K. Lin, Jackei H.K. Wong, ChenYe Zhu, Allan K.Y. Wong, and Tharam S. Dillon, Applying Fuzzy Logic in Dynamic Buffer Tuning to Enhance the Success of Pervasive Medical Consultations, Fuzzy Logic: Theory, Programming and Applications, Nova Science Publishers, Incorporated, 2008 (to appear)

[2008-p3] Jackei H.K. Wong, Tharam S. Dillon, Allan K.Y. Wong and Wilfred W.K. Lin, Text Mining for Real-time Ontology Evolution, Data Mining for Business Applications, Springer, 2008, ISBN: 978-0-387-79419-8, 143-150

Invited sessions (Keynote - Seminar):

[2008-p4] Invited plenary keynote/seminar - Tele-medicine: Application to Traditional Chinese Medicine (TCM), IEEE-DEST2008, Phitsanulok, Thailand, 25-29 February, 2008

Refereed journal papers:

[2009-p5] Jackei H.K. Wong, Allan K.Y. Wong, Wilfred W.K. Lin, A Novel Real-Time Traffic Sensing (RTS) Model to Improve the Performance of Web-based Industrial Ecosystems, IEEE Transactions on Industrial Electronics (TIE), to appear

[2009-p6] Jackei H.K. Wong, Wilfred, W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon, Automatic Enterprise-Ontology-Driven TCM (Traditional Chinese Medicine) Telemedicine System Generation, IEEE Transactions on Industrial Electronics (TIE), to appear

[2009-p7] Jackei H.K. Wong, Allan K.Y. Wong, Wilfred W.K. Lin, and Tharam S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical

Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), to appear

Refereed conference papers:

[2009-p8] Jackei H.K. Wong, Wilfred W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon, TCM (Traditional Chinese Medicine) Telemedicine with Enterprise Ontology Support – a Form of Consensus-Certified Collective Human Intelligence, Proceedings of the International Conference on Industrial Technology (ICIT**),** Monash University, Victoria, Australia, 10-13 February 2009 (**Invited**)

[2008-p9] Jackei H.K. Wong, Wilfred W.K. Lin and Allan K.Y. Wong, Real-Time Enterprise Ontology Evolution to Aid Effective Clinical Telemedicine with Text Mining and Automatic Semantic Aliasing Support, Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, 11-13 November 2008 (**Invited**)

[2008-p10] Jackei H.K. Wong, Wilfred W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon, Dynamic Cache Tuning Aids the Success of Telemedicine, Proceedings of the 11th IEEE International Conference on Computational Science and Engineering – Workshops, Sao Paulo, Brazil, 16-18 July 2008

[2008-p11] Jackei H.K. Wong, Wilfred W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon, Applying Neuro-Fuzzy Dynamic Buffer Tuning to Make

Web-based Telemedicine Successful, Proceedings of the International Conference on E-business and Telecommunication Networks (ICETE), Porto, Portugal, 26-29 July 2008 **(Invited)**

[2008-p12] Jackei H.K. Wong, Wilfred W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon**,** An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008 **(Invited)**

[2008-p13] Jackei H.K. Wong, Allan K.Y. Wong, Wilfred W.K. Lin and Tharam S. Dillon, Dynamic Buffer Tuning: An Ambience-Intelligent Way for Digital Ecosystem Success, Proceedings of the 2nd IEEE International Conference on Digital Ecosystems and Technologies (DEST), Phitsanulok, Thailand, 26-29 February 2008, 184-191

[2008-p14] Wilfred W. K. Lin, Jackei H.K. Wong and Allan K.Y. Wong, Applying Dynamic Buffer Tuning to Help Pervasive Medical Consultation Succeed, Proceedings of the 1st International Workshop on Pervasive Digital Healthcare (PerCare), in conjunction with the 6th Annual IEEE International Conference on Pervasive Computing and Communications, Hong Kong, 17–21 March 2008, 675-679

[2007-p15] Allan K.Y. Wong and Jackei H.K. Wong, Dynamic Cache Size Tuning to Shorten Mobile Business Service Roundtrip Time and Turn E-

Shoppers into Happy Return Customers, Proceedings of the 6th International Conference on the Management of Mobile Business (ICMB), Toronto, Ontario, Canada, 8-11 July 2007 (**Nominated to be an invited book chapter – 2009-p1**)

[2007-p16] Allan K. Y. Wong, Wilfred W. K. Lin, Tharam S. Dillon and Jackei H.K. Wong, A Novel Self-Similar (S^2) Traffic Filter to Enhance E-Business Success by Improving Internet Communication Channel Fault Tolerance, Proceedings of the 9th International Parallel Computing Technologies Conference (PaCT), Pereslavl-Zalessky, 3-7 September 2007, 328-339

The 16 papers listed can be put into 6 different contribution categories, as follows:

a) *Reliable communication over the mobile Internet*: Successful web-based data mining and useful herbal discovery depend on reliable communication to varying degree. Therefore, serious preliminary investigation was carried out in this respect and the findings contributed to the following publications listed above: p1, p2, p5, p8, p9, p11, p12, p13, p14, p15, p16
b) *Unambiguous database design based on the ontology concept*: Meaningful discovery ties in with the unambiguous information in the database. One of the ways to unambiguously organize entities and their relationship is through ontology. As a result, an investigation was conducted on how TCM information can be organized into the

corresponding construct, namely the enterprise ontology construct. The investigation led to the following publications: p4 and p7.

c) *Effective text mining as the technique to extract useful information from the open sources (e.g. the open web)*: The master enterprise ontology is a "closed" system because it cannot evolve automatically. In order to enable the "operational" enterprise ontology keeping itself abreast with contemporary TCM knowledge, the D/P system should be equipped with a mechanism to aid automatic, "selective" evolution. At anytime the mechanism would text-mine the open sources and "value-add" the skeletal master enterprise ontology of the target system. Yet, the information addition is logical because it has not gone through the stringent process of consensus certification by a sufficient number of experts. It is selective because the user can switch off the mechanism for real-time evolution anytime and revert the operation of the target system back to the "original master" ontology only. The investigation contributed to the following publications: p3.

d) *Selective and "extensible" on-line ontology evolution*: It is selective because the user can switch off the mechanism for on-line ontology evolution anytime, anywhere and revert the system operation back to the "original master" ontology only. Any new information found and added to master ontology is of the temporary "extensible" nature because it has not been consensus-certified. The investigation contributed to the following publication: p9.

e) *Undisputable principle for herbal discovery*: Herbal discovery has two basic levels in this research: prescription (a collection of herbal ingredients) and herbal ingredients (single items): p4.

f) *Automatic ontology-based software/system generation*: The information in the database of any D/P can evolve in an "extensible" manner. Besides, open sources of new information should include the "accumulated" experience of a physician. This experience serves as feedback to enrich the database, and this includes the "non-standard" hand-written practices and cases. The argument is that, for the information in the database to be correct and able to match either a part of the full master enterprise TCM ontology, the system building process must: i) not smudge the TCM ontology of the target system; and ii) introduce mistakes when adding in either the physician's key-in (standard) or hand-written (non-standard) diagnosis/prescription information. This led to Nong's primordial argument that D/P system building should start with a meta-interface (MI) specification. Ideally the icons in the MI specification should be selected from the master icon library, and every icon should be a semantic path encoded in the basis/master ontology of the target system, which should be automatically generated for the given MI specification in one-to-one correspondence. The investigation results in this aspect contributed to two publications: p6 and p10.

The description of this PhD research, which is represented by the 14 tasks as shown in Figure pa.2, is organized into ***one Preamble*** and ***11 chapters*** in this thesis as follows:

a) **Preamble** – This is the extended abstract for the thesis, telling the reader what is to come next. The definitions of some useful terms are listed to help readers understand this thesis more clearly.
b) **Chapter 1 Research Scope and Methodology** – This defines the aim and objectives and describes some relevant issues. The research methodology to guide the research to success is also presented and explained.
c) **Chapter 2 Review of Related Work** – Those related work pieces, which are eminently relevant to the formulation of objectives in this thesis, are critically reviewed.
d) **Chapter 3 The Realm of Metadata Usage** – The TCM ontology should be annotated clearly by a metadata system. To absorb the rich user experience already in existence, the intention is to follow the W3C (World Wide Web Consortium) recommendations.
e) **Chapter 4 Ontology, Enterprise Ontology, Semantic Transitivity and Knowledge Discovery by Data Mining** – This explains the essence of successful knowledge discovery from ontological constructs and discusses some relevant and significant issues.
f) **Chapter 5 Essential Software Engineering Support** – This explains why appropriate software engineering support is essential for successful discoveries of herbal ingredients from ontological constructs. The new

ontology-based technique that supports automatic generation of prototypes for tests is also introduced. These new techniques were called different names at different times in the course of the PhD research, including: MI (*meta-interface*) paradigm, CTSS (*customized telemedicine software system*) approach, WTS (*web-based telemedicine system*) approach, and EOD-ISD (*enterprise ontology driven information system development*). These names represent the course of evolution of this technique from the primordial MI concept to the mature EOD-ISD approach. They are considered synonyms in this thesis.

g) **Chapter 6 Living Ontology, Semantic Aliasing and Relevance Index** – A living ontology evolves continuously to keep itself abreast of events over time. It is a requirement in the research to enliven the local ontology core of the prototypes, which are basically generated from the given static Nong's TCM onto-core. The "living ontology mechanism (LOM)" extends the knowledge in the local onto-core but does not change the skeletal information that is the same as the given onto-core. The argument is that any new information to be included in any ontology must go through an "official" process of consensus certification. If the LOM is inhibited, the prototype operates with the "original" knowledge, as if nothing had happened. A **novel** automatic semantic aliasing mechanism is proposed to serve two purposes: i) to discover new prescriptions from the enlivened/original TCM onto-core; and ii) to absorb and standardize new knowledge from open sources (e.g. open web or handwritten information) for the same illnesses. The relevance index indicates/measures the degree of similarity between the

"host" illness and the newly "discovered one". The work in this chapter is the consequence of addressing the issues in Chapter 4 successfully.

h) **Chapter 7 Knowledge Classification for Herbal Discovery** – It argues that the neural network (NN) is the better alternative for TCM herbal classification and discovery than the algorithmic approach because the former is free of axiomatic ambiguity, more flexible, and less programming-error-prone.

i) **Chapter 8 Discovery of Individual Herbal Ingredients** – This argues that the neural network (NN) approach is also suitable for discovering individual herbal ingredients. The discovery procedures make use of the philosophy of relevance index (RI).

j) **Chapter 9 Walkthrough of Selected Experimental Results** – It critically walks through, analyzes and discusses the relevant verification results that were obtained in the course of the *WD²UHI* research.

k) **Chapter 10 Review of the Proposed Solutions, Achievements and Contributions** – It focuses on: i) originality of the proposed solutions; ii) overall achievements; and iii) contribution by this thesis to the body of telemedicine knowledge.

l) **Chapter 11 Conclusion and Suggested Future Work** – This emphasizes: i) the motivation of the research; ii) the aim and objectives; iii) the overall achievements, iv) the original contribution to the body of knowledge, namely, "*Web-based Data Mining and Discovery of Useful Herbal Ingredients (WD²UHI)*"; and v) the suggested direction for the future work.

Chapter 1 Research Scope and Methodology

The research *scope* is abstracted by the project title: *"Web-based Data Mining and Discovery of Useful Herbal Ingredients (WD^2UHI)"*. This abstraction represents three basic goals: i) data-mine the web (and other open sources) for new knowledge; ii) find a way to incorporate the mined information into extant knowledge base to from a new updated knowledge embodiment – extant/old knowledge base "enlivened" and equipped for on-line evolution; iii) discover new knowledge from both the old knowledge base as well as the new embodiment. This research is a TCS (Teaching Company Scheme) joint venture between the Hong Kong Polytechnic University and the Nong's Company Ltd., which is a subsidiary of the PuraPharm Group in the Hong Kong Special Administrative Region (HKSAR). It combines two knowledge domains: Information Technology (IT) and Traditional Chinese Medicine (TCM).

TCM practice in the Hong Kong SAR is protected by law - the Chinese Medicine Ordinance. This legal protection inspires deep TCM research in the HKSAR of China. The PuraPharm group is a frontrunner in the TCM industry. Firstly, it is the sole supplier of herbs to all the hospitals under the control of the Hong Kong Hospital Authority (HA), which was formerly a governmental division and is now a semi-governmental organization. Secondly, it is a reputable pharmaceutical (mainly herbal) manufacturer. Thirdly, it supplies TCM drugs to more than 80 hospitals and clinics in mainland China. Fourthly, it has deployed the first TCM pervasive mobile clinic in the HKSAR

successfully. Since their deployment the mobile clinics treat hundreds of patients daily. In fact, the PuraPharm Group is operating with the web-based mobile business (e-business) model [Thomas03] (Figure pa.1). In this model, the company or enterprise spontaneously visualizes the supply and demand forces and balances them as the proactive response that helps maximize the profits. Since PuraPharm is a drug manufacturing enterprise, it needs to discover new herbs or herbal products to remain ahead of the competition. In this light PuraPharm and the Hong Kong PolyU have paired together to explore the pristine area of *web-based data mining and discovery of useful herbal ingredients*, and this has become my PhD pursuit under the TCS (Teaching Company Scheme). While I am conducting the research I am also working with, and gaining experience directly, from the mobile clinics (MC). The Chief Supervisor of my PhD research, Dr. Allan Wong, was the architect of the extant PuraPharm's TCM mobile-clinic telemedicine system, which is deployed successfully in the HKSAR and has been treating hundreds patients daily for the last three years. This MC system was first deployed by Yan Oi Tong (YOT) in its TCM operation and later by others. YOT is one of the biggest sub-vented charity organizations in Hong Kong. It is operating various hospitals, clinics and geriatric homes.

1.1 Aim and Objectives

The aim is to propose a conceptual framework for useful *web-based data mining and discovery of useful TCM (Traditional Chinese Medicine) herbal ingredients*. This area of research is pristine, and there is little published experience.

The ***two*** possible main objectives to be achieved in this research include:

a) ***First objective –*** Contribute to the development of a ***trusted*** WD^2UHI platform: This platform should support useful and meaningful web-based data mining and discovery of useful TCM herbal ingredients. The platform is trusted because it satisfies ***two very basic requirements***: i) it provides reliable wireline/wireless communication to support correct, pervasive, and responsive client/server interactions; and ii) it is supported by a "standard", unambiguous, and semantically-transitive knowledge base to support local and global interoperability to a varying degree. Due to the colossal amount of research work required, it was decided that the ***focus*** of this thesis should be ***mainly*** on the contribution to the ***second basic requirement above***, after serious and extensive preliminary exploration was conducted for the first requirement. It is envisioned that the WD^2UHI framework and thus the prototypes for the verification experiments would be characterized by the following:

i) ***3-layer architecture*** **–** The bottom layer is the TCM ontology; the middle layer has ***two main items***, namely, the ***semantic net*** which is the machine understandable form of the ontology and the ***parser*** to process it; the top layer, which is a query system that abstracts the semantic net, for human understanding and manipulation.

ii) ***An unambiguous and axiomatic TCM ontology core (onto-core)*** **–** This is a local/enterprise view (or subset) carved out of the global TCM ontology, which is made up of all the available TCM classics, treatises, case histories, and new scientific findings, by the process of consensus certification.

iii) ***On-line evolution*** **–** The platform should absorb new knowledge from the open sources continuously and in a real-time manner. Yet, the absorbed knowledge is only a temporary extension of the enterprise TCM onto-core of the running system (or prototype) because permanent knowledge absorption must be vetted by the consensus certification process instituted by the system owner or enterprise.

iv) ***Ontology-based software engineering (SE)*** **–** A new paradigm in this direction is needed to be proposed so that correct WD^2UHI based prototypes can be quickly, correctly and automatically generated in a single step from the named TCM onto-core given as the basis. The advantage of this new SE paradigm is that the semantic transitivity among the three system layers (i.e. bottom ontology, middle semantic net and

upper query system) in the prototype can be guaranteed. This cross-layer semantic transitivity confirms if the ontology and therefore the discovered material are trustworthy. The new SE paradigm produces semantically correct *customized telemedicine software systems* (CTSS) from the given TCM onto-core from the specification provided by the client. Even though the contents of specifications vary with different clients, the terminology/lexicon/vocabulary within these specifications should be "standard" in respect to the given TCM onto-core. For the sake of convenience, this specification is called meta-interface (MI) specification or simply MI as in the primordial MI SE paradigm, which was first proposed by the Nong's without all the necessary details for implementations (i.e. a "shell" paradigm). To generalize, the ontology-based SE paradigm generates automatically the target *web-based telemedicine system* (WTS) in one step from the given TCM onto-core with the given MI specification. Therefore, the local operational ontology in the target WTS or CTSS (synonyms), in reality, is customized from the same TCM onto-core given. It is a variant because different MI specifications produce functionally different variants. To make the primordial MI paradigm proposed by Nong's work for my PhD research, new elements are defined, verified and added as follows: *cross-layer semantic transitivity, automatic software generator, master library of*

icons, metadata impact on ontology representation, living ontology (i.e. ontology evolution on-line), standard keyed-in ontological information versus non-standard hand-written information, and support for meaningful discoveries. These elements will be explained later in details. The name of MI is changing as necessary in the course of the research, and it has the following synonyms (please refer to the definitions of useful terms): WTS, CTS, and EOD-ISD.

b) ***Second objective*** – Propose novel methods to discover herbal ingredients correctly and meaningfully: In this research herbal discoveries are divided into two levels: prescriptions (high-level) and individual herbal ingredients (low level). A TCM prescription is usually made up of at least four types of herbs by their capacities/roles: i) the *principal*, to treat the illness head-on; ii) the *courtier*, to enhance the principle's curative power; iii) the *assistant*, to aid the courtier by pacifying its possible ill effects; and iv) the *messenger*, to bring the curative effect to the "cause or point" of the illness spot-on. The relationship between an herbal ingredient and a prescription, however, is not transitive. From another angle, any discovery can be Type 1 (if it is outside the current ontological context) or Type 2 (if it is within the current ontological context).

Achieving the above two research objectives means addressing many relevant issues including the following satisfactorily:

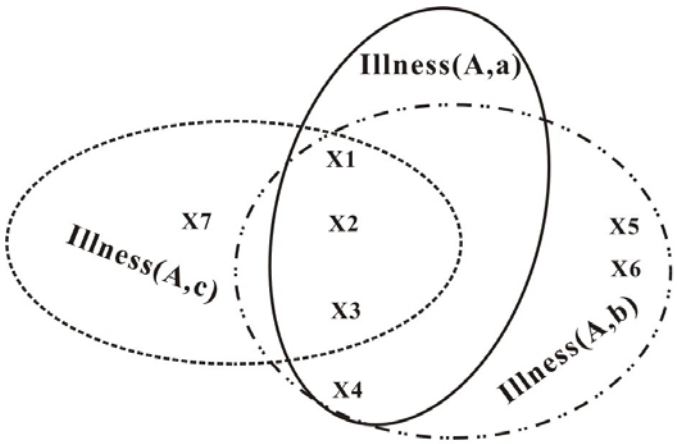
1) What does useful herbal discovery in TCM mean?
2) What kind of structure should be there to support accurate and consistent TCM knowledge representation in evolution?
3) What is the relationship among distributed computing over the web, herbal knowledge repertoire, discovery, and information retrieval?
4) What has been done/achieved out there in the computing knowledge domain in light of medical/herbal knowledge representation, telemedicine, discovery techniques, successful examples, information retrieval, and machine learning (to absorb new medical/herbal knowledge)?
5) If the open web was the universal knowledge base, how could newly emerging and relevant knowledge be discovered (i.e. identified, retrieved, standardized, and absorbed) in a value-added manner?
6) Can the Nong's TCM mobile-clinic telemedicine system help shed light on how such a system can self-learn and transform informal personal experience of epidemiological origins into formal knowledge to a varying degree?

The subsequent preliminary investigation of the above relevant issues had revealed that their complexity together is much more than a PhD thesis. Therefore, I had narrow down the scope of work, with consensus from my thesis supervisor and the research endorser (i.e. PuraPharm) by discontinuing the deeper investigation into the communication issues such as the impact of Internet traffic on the stability of the *WD*^{2}*UHI* platform. More precisely, the conceptual *WD*^{2}*UHI* framework proposed in my thesis should now address ***two***

main issues only: i) develop a trusted WD^2UHI framework <u>*without*</u> *deeper scrutiny of trusted communication issue*; and ii) propose methods to discover herbal ingredients correctly and meaningfully.

Herbal discoveries should base on formal principles of the two mains:

a) ***TCM domain***: This is the formal diagnostic principle enshrined in the TCM classics – The "同病異治, 異病同治" [WHO07]" in Chinese terminology or SIMILARITY/SAME (i.e. "同") principle in English [JWong09a]. If the three different sets of prescriptions for Illness (A, a) (i.e. illness A for region a), Illness (A, b) (i.e. the same illness name for region b) and Illness (A, c) (i.e. same illness name for region c) are PAa, PAb, and PAc respectively, by the SIMILARITY/ SAME principle the total/common set of usable prescriptions for treating the three illnesses should be $P_{all} = PAa \cup PAb \cup PAc$. The $\cup$ operation associates the three different sets of prescriptions into a single pool (i.e. common set P_{all}) by their common attributes/factors. In fact, the three illnesses are defined according to very different attributes other than the common set, due to geographical and epidemiological differences. This formation of a common set by similarity (in terms of common attributes) is called *automatic semantic aliasing*. This is illustrated in Figure 1.1.1, in which the three illnesses have common attributes (e.g. symptoms) X1, X2, and X3.



Legend: a, b and c are 3 different geographical regions

Figure 1.1.1 Automatic semantic aliasing

b) ***Logical domain***: Here, we deal with sets of entities and their axiomatic logical relationship. In this light, the three illnesses in Figure 1.1.1 are three subsets of entities created from the entity population by random experiments, which arbitrarily include some entities into their subsets. Then, the subset Illness (A, a) contains the entities X1, X2, X3, and X4, and similarly the subset Illness (A, b) is defined by the entities X1, X2, X3, X4, X5, and X6. Using Illness (A, a) as the ***referential point/host***, then Illness (A, b) is *very or strongly similar but not logically the same*; the similarity is 4/4 or 1 (as there are 4 entities from the view of Illness (A,a) – the numerator, and Illness (A, b) contains all 4 entities – the denominator). From Illness (A, b)'s angle, Illness (A, a) is only 4/6 or 66.7% similar, for the different number of constituent attributes or entities in the two sets. The sets S1 and S2 are synonyms (exactly or 100% same) only if $S1 = S2$ logically holds (i.e. logically equivalent). From the angle of probability (P) theory, the $P(S1 \cup S2) = P(S1) + P(S2) - P(S1 \cap S2)$ expression indicates that S1

and S2 are only *aliases* with the resemblance probability equal to $P(S1 \cap S2)$ conceptually, where $\cup$ and $\cap$ are union and intersection operators respectively.

1.2 First Recap

This PhD research project is entitled: "**Web-based Data Mining and Discovery of Useful Herbal Ingredients (** WD^2UHI **)**". It combines both the IT and TCM domains and being a joint venture between the Hong Kong Polytechnic University and the Nong's Company Ltd. – a subsidiary of the PuraPharm Group in Hong Kong SAR. This project is a TCS (Teaching Company Scheme) based, and for this reason I will be able to carry out all the verification experiments in the Nong's TCM telemedicine clinical environment. After narrowing down the research scope carefully, the effort is mainly on *knowledge-based hierarchy, knowledge absorption* and *herbal discovery*. The knowledge discoveries can be Type 1 (outside the current ontological context) and Type 2 (within the current ontological context). It is "high-level" if the discovery is a prescription and "low-level" if it is an herbal ingredient or logical relationship among herbal ingredients.

1.3 Research Methodology

Research success depends on the appropriateness of the methodology adopted [Ketchen04]. A research process may have many phases, for example, the seven-step model proposed by Kumar [Kumar96]. The steps are: a) problem

formulation, b) design conceptualization, c) instrument construction for data collection, d) selection of sample types for experiments and testing, e) research proposal writing, f) collection of data for the selected type(s), and g) analysis of the sampled data. The research methodology must suit the problem domain for good results. In fact, many strategies and methods are in existence to cover different problem domains and research types. In the computing research area, three basic types of research may be identified, namely [Yu87]:

1) *Exploratory research*: This tackles a little known problem for which the research details cannot be formulated very well at the beginning. The result of this kind of research usually pushes the boundaries of the knowledge frontiers and leads to discovery of new knowledge. *My PhD thesis belongs to this category*.
2) *Testing-out research*: This finds the limits of previous generalizations.
3) *Problem-solving research*: This usually starts with a specific real-world problem of well-defined characteristics and then brings all the available intellectual resources together for a reasonable solution.

Even when an appropriate research methodology has been identified, suitable basic strategies may have to be defined to timely smooth out the research activities in the process:

1. *Top Down*: The objectives are defined and realized step by step from conceptual to prototype building. The typical examples include: 1) Waterfall model for traditional software engineering based on

algorithmic programming that does not encourage user intervention, and 2) *fast prototyping* that encourages repetitive user input until the system is finally accepted. In both cases, there could be a lot of backtracking to do.

2. *Bottom Up*: The coordination framework, as a connective, is first proposed so that what is available (commodities and/or intellectual resources) can effectively be interconnected into a single system. This research strategy, which has a sense of integration, advances the complexity of the target system with time.

The Top Down approach is suitable for testing-out research, and the Bottom Up approach is natural for the problem-solving type.

The nature this PhD research is exploratory and the topic involves the **novel** notion of ***semantic TCM***, which has little published experience in the literature. After my analysis plus enlightenment from other TCM experts I had decided to start my research in a top-down fashion in the following phases: literature search, problem statement, proposed solutions, and data collection. It is, however, difficult to apply the Top Down approach in a strict sense in my case because the research process will inevitably involve repetitive backtracking and cross-referencing to gain insight at different stages before proceeding further. It is not easy to find an extant research methodology that can support this kind of repetitive, backtracking activity. After extensive reading and careful consideration, I decided to propose my own version of the IEP methodology [Wu02], namely, "*investigate & experiment & proceed with*

possible backtracking, cross referencing and looping (IEP)". I call my new research methodology the ***new IEP*** (N-IEP).

The main difference between the N-IEP and a strict top-down approach is that in the N-IEP, explorations and investigations start from the root through the branches to a leaf and then back again. The traversals up and down the branches and leaves represent a heuristic process, and they may repeat many times before enough material, insight and data can be crystallized sufficiently to start the next stage of the research. The N-IEP details are concisely condensed in the roadmap in Figure 1.3.1, which guides the thesis through. For example, a N-IEP traversal may be the following path: *Dynamic buffer tuning → Test it with the extant Nong's web-based telemedicine system for some insight → Use the test data to improve the extant Nong's TCM onto-core's visualization mechanism→ Construction of the new TCM onto-core and test it with the visualization support to gain some insight→ Apply techniques from the semantic web proposals in this construction→ Apply text mining to build the new skeletal onto-core→ Verify the logical correctness of this new onto-core with formal methods → Build the first prototype for gaining more insight.* This path, however, is only one of the many possible "*operation*" paths in the course of the whole PhD research. Traversing the N-IEP roadmap back and forth may be necessary, but it should be guided by the data collection and analysis result at the time, involving cross-reference, data refinement and/or comparison. In Figure 1.3.1 those items that should be investigated first are in "*solid-line boxes*", and those "*dotted-line boxes*" would usually be investigated later. The research activities are basically organized as a ***fast prototyping process***, which

feeds the current useful experience to the next stage incrementally to re-orient the direction when necessary. In fact, fast prototyping is natural in my research because the trustworthiness my prototypes for experiments should be verified in the clinical environments, and the verifications usually involve TCM physicians who may use the prototypes to actually help treat patients.

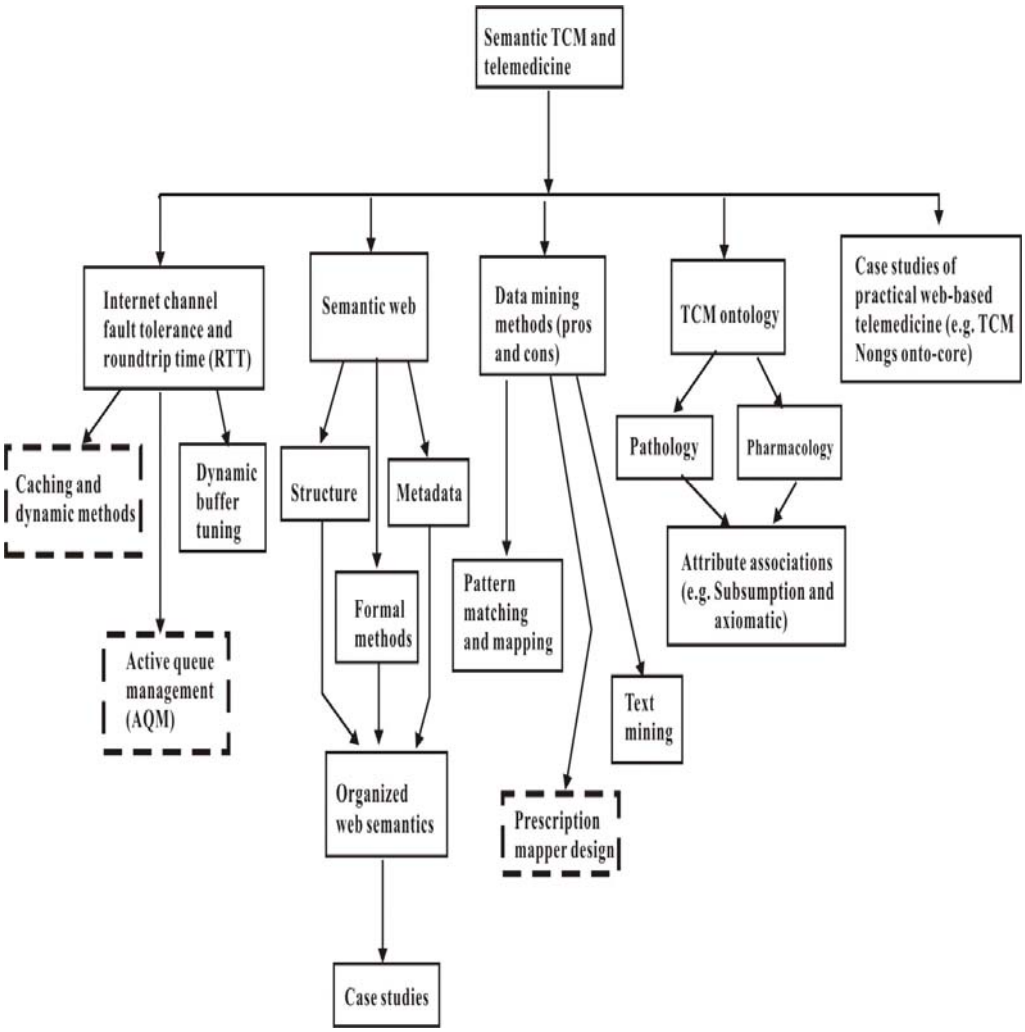


Figure 1.3.1 The proposed N-IEP strategy/roadmap for project management

Traversals of the roadmap in Figure 1.3.1 may involve the following activities that are intertwined overtime in different ways with backtracking:

1. Understanding the following as thoroughly as possible to gain insight:

    i) Internet channel fault tolerance and roundtrip time (RTT) [Lin06a, Lin06b, Paxson99, Molnár99, Chatranon04, Wu06a, Wu06b].

    ii) Semantic web [Rifaieh06].

    iii) Data mining methods [Bloehdorn05, Witten00, Yu06, Lu00, Agrawal96, Wu02, Fayyad96]

    iv) TCM ontology [JWong09a].

    v) Strengths and weaknesses of the extant/old Nong's TCM onto-core.

2. Identifying carefully the general requirements in D/P (diagnosis/prescription) telemedicine over the mobile Internet.
3. Evaluating the different text-mining strategies and scrutinizing different ways whereby a web-based text-mining technique can be constrained for focal TCM applications.
4. Proposing how an extant TCM onto-core can be enlivened for on-line evolution, with data mining (e.g. text-mining) support.
5. Define the scope and levels of herbal discoveries.
6. Propose methods to guarantee the trustworthiness of a discovery in light of: i) semantic transitivity, time-honoured TCM principle(s), and

axiomatic/logical correctness from the IT angle.

7. Propose methods whereby trustworthy prototypes for verifications can be quickly built and deployed.

1.4 Second Recap

Although the SE paradigm for developing the *WD²UHI* based prototypes is proposed, the different research activities have to be timely coordinated to make the research a success. Through information gathered from previous experience and my own careful analysis, I have proposed the N-IEP research methodology.

1.5 Conclusion and Connective Statement

This research, "***Web-based Data Mining and Discovery of Useful Herbal Ingredients (*** *WD²UHI* ***)***" combines both the Traditional Chinese Medicine (TCM) and IT (Information Technology) domains. The aim is to discover useful TCM herbal ingredients and prescriptions from the given TCM ontology core. Yet, the discoveries should apply to both the static onto-core and the open knowledge sources such as the web. It is therefore, a project requirement to make the given static onto-core a living entity that can evolve and keep itself abreast of contemporary changes over time. Since this is a TCS (Teaching Company Scheme) based PhD research, the endorsing company Nong's Company Ltd. permits my verification exercise to be conducted in its successful TCM telemedicine clinical environment. The prototypes are

customized from the Nong's master/enterprise TCM onto-core. To implement these prototypes quickly for the experiments the research has made use of and perfected the primordial "shell" MI SE paradigm, which was first proposed by Nong's with no implementation details. That is, the usable version of the MI "shell" concept is the **novel** SE approach contributed by this thesis for automatic system generation from the given ontology core. With this, *WD*^{2}*UHI* based prototypes are generated quickly in a single step – only the iconic MI specification is needed. Since this research involves intensive and pristine activities that cannot adhere directly to any published previous experience, I proposed the N-IEP research methodology to coordinate all necessary research steps and activities. The herbal discoveries are based on two principles that I proposed and condensed: i) the SAME principle in the TCM domain; and ii) the axiomatic logical relationship in the IT domain.

The next chapter is to discuss the review of literature and related work that have helped the formulations of the proposed solutions.

1.6 Key References

[Agrawal96] R. Agrawal and J. Shafer, Parallel Mining of Association Rules, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996, 962-969

[Bloehdorn05] S. Bloehdorn, P. Cimiano, A. Hotho and S. Staab, An Ontology-based Framework for Text Mining, LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, May 2005, 87-112

[Chatranon04] G. Chatranon, M.A. Labrador and S. Banerjee, A Survey of TCP-friendly Router-Based AQM Schemes, Computer Communications, Vol. 27, No. 15, September 2004, 1424-1440

[Fayyad96] U.M. Fayyad, S.G. Djorgovski and N. Weir, Automating the Analysis and Cataloging of Sky Surveys, in Advances in Knowledge Discovery and Data Mining, eds. Y.M. Fayyad, AAAI/MIT Press, 1996

[JWong09a] J.H.K. Wong, A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), 2009

[Ketchen04] D. Ketchen, C. Snow and V. Hoover, Research on Competitive Dynamics: Recent Accomplishments and Future Challenges, Journal of Management, Vol. 30, No. 6, December 2004, 779-804

[Kumar96] R. Kumar, Research methodology, A Step-by-step Guide for Beginners, Melbourne: Longman Australia, 1996

[Lin06a] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, Application of Soft Computing Techniques to Adaptive User Buffer Overflow Control on the Internet, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 36, No. 3, May 2006, 397-410

[Lin06b] W.W.K. Lin, A.K.Y. Wong and R.S.L. Wu, Applying Fuzzy Logic and Genetic Algorithms to Enhance the Efficacy of the PID Controller in Buffer Overflow Elimination for Better Channel Response Timeliness over the Internet, Concurrency and Computation: Practice & Experience, Vol. 18, No. 7, June 2006, 725-747

[Lu00] H. Lu, L. Feng and J. Han, Beyond Intra-transaction Association Analysis: Mining Multi-dimensional Inter-Transaction Association Rule, ACM Transactions on Information Systems, Vol. 18, No. 4, October 2000, 423-454

[Molnar99] S. Molnar, T.D. Dang, and A. Vid acs, Heavy Tailedness, Long-range Dependence and Self-similarity in Data Traffic, Proceedings of the 7th International Conference on Telecommunication Systems Modeling and Analysis, Nashville, Tennessee, USA, March 1999

[Paxson99] V. Paxson. Bro: A System for Detecting Network Intruders in Real-Time, Computer Networks, Vol. 31, No. 23-24, 1999, 2435-2463

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[Thomas03] S.F. Thomas and M.L. Gillenson, Mobile Commerce: What It Is and What It Could Be, Communications ACM, Vol. 46, No. 12, December 2003, 33-34

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

[Witten00] I.H. Witten, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Academic Press, 2000

[Wu02] R.S.L. Wu, A.K.Y. Wong and T.S. Dillon, Comparing Four Novel Scalable Split/Aggregate Algorithms (Mobile Agent Based) for Distributes Mining of Multimedia Association Rules over the Internet, Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, Vol. 2, 24-27 June 2002, 760-766.

[Wu06a] R.S.L. Wu, A.K.Y. Wong and T.S. Dillon, A Novel Dynamic Cache Size Tuning Model with Relative Object Popularity for Fast Web Information Retrieval, Journal of Supercomputing, 2006

[Wu06b] R.S.L. Wu, W.W.K. Lin and A.K.Y. Wong, Harnessing Wireless Traffic is an Effective Way to Improve Mobile

Internet Performance, Proceedings of the 1st Australian Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless), Sydney, Australia, March 2006

[Yu87] P.S. Yu, C.M. Krishna and Y.H. Lee, An Adaptive Optimization Model with Applications to Testing, Computer Performance and Reliability, 1987, 503-515

[Yu06] S. Yu, H. Qiang and S. Jing, A Framework of XML-Based Geospatial Metadata System, Proceedings of the APWeb Workshops, 2006, 775-778

Chapter 2 Review of Related Work

The literature review focused on the following major areas: i) trusted communication; ii) reliable software engineering; iii) concept of ontology; iv) ontology-based software engineering; v) D/P (diagnosis/prescription) medical systems vi) text mining; and vii) neural networks. The aim was to let the review help define the scope of my PhD research so that it could be finished within the expected time constraints. The conclusion from the literature review was that my PhD research should exclude the area of trusted communication, which is vast and time-consuming, and concentrate mainly on addressing the issue of herbal discoveries, namely Type 1 and Type 2. This decision was a collective one made by my PhD supervisor, myself and PuraPharm that funds this PhD research under the TCS (Teaching Company Scheme) agreement with the Hong Kong Polytechnic University.

2.1 Trusted Communication

Trusted client/server communication is a fundamental requirement for the success of any time-critical applications [Lin06a, Braden98, JWong07], including telemedicine [Lin08]. According to [RFC2828], system trustworthiness should be defined conceptually by the set of parameters chosen for the application domain. As far as the Internet is concerned, one of the parameters for gauging trustworthiness is system ***response***, which is greatly affected by the traffic pattern. The Internet has two kinds of traffic: i) the connection-oriented TCP (Transmission Control Protocol) traffic, and ii) the

send-and-forget UDP (User Datagram Protocol) traffic. Both the TCP and UDP are built on top of the IP (Internet Protocol), which interconnects remote networks over the globe together. Since its dawn, the TCP, which is for in-order, reliable client/server interactions, has been supported by various built-in "*flow control*" for congestion avoidance [Chatranon04], including: i) timeout and retransmission; ii) "*slow start*" of the AIMD (*additive increase & multiplicative decrease*) nature; and iii) throttling of runaway senders by the receiving router. Since the TCP traffic responds to these built-in mechanisms it is ***responsive***.

In contrast, UDP traffic does not respond to these mechanisms, therefore it is ***unresponsive***. Since the UDP users such as providers of video-on-demand are opportunistic, they would increase their sending rate whenever they sense that there is extra Internet bandwidth. As a result, unfair bandwidth usage would surface, for the *unresponsive* traffic has usurped the network bandwidth from *responsive* traffic. For this reason the modern AQM (*active queue management*) mechanisms (e.g. the BLACK) will arbitrate the network bandwidth between TCP and UDP users [Chatranon04]. It is not so long ago that Internet users realized the importance of effective AQM, and since then more effective methods to support AQM have appeared continually [Braden98, Ren02].

Despite the existence of usable congestion avoidance supports, it has been difficult to use the Internet for serious mission and/or time-critical applications such as remote *diagnosis and prescription* (D/P) operations in telemedicine [JWong08b]. The reasons are as follows:

a) ***Uncontrollable system response time***: It is difficult to monitor and control the service roundtrip time (RTT) in any TCP based client/server interaction. The reason is that it is impossible to monitor all the dynamic faults and errors of a TCP channel because the IP underneath works by hop-by-hop routing. If we encapsulate all these dynamic errors by the error probability δ , the average number of trial (ANT) to get a transmission success can be defined by the expression $ANT = \sum_{j=1}^{k\to\infty} jP_j = \sum_{j=1}^{k} j[\delta^{j-1}(1-\delta)] \approx \frac{1}{(1-\delta)}$; reducing δ would lower ANT. In fact, all the extant AQM mechanisms contribute to reduce ANT. The issue of how to reduce δ is an important one but so far there are not many usable real-time solutions in the field. The more well-known examples in the literature include: i) dynamic buffer tuning [Lin06a]; ii) dynamic cache tuning [Wu06a, Wong03, Podlipnig03]; and iii) real-time IEPM (Internet end-to-end performance measurement) [Cottrell99, Wong01]. In fact, Paxson observed and argued that any system running on the Internet would fail if it was designed with a preconceived mathematical model in mind [Paxson95]. Since then, many researchers have been correlating this kind of failures with changes in the mean service RTT values. This correlation is the backbone to the IEPM school of thought [Cottrell99].

b) ***Discrepancy between traffic modeling and reality***: The Internet traffic can be stationary of independent increments or "*non-stationary*" at other

times [Wong08]. Stationary traffic can be of *short-range dependence* (SRD) and *long-range dependence* (LRD). LRD traffic can be very complex; for example, it can be self-similar [Crovella97], heavy-tailed or both in a mixed manner. Yet, any stationary traffic pattern can be statistically modeled [Wong08] to a varying degree of accuracy, but the empirical approach can mask out computational errors if they are too fine to be detected (masking means inadvertently neglecting). As a result, even it was equipped with a good traffic filter [Wu06b] the receiver might have received packets from "*a system which had failed – failure undetected*". This leads to the dilemma of "*using incoherent or corrupted data*" that would lead to unforeseen consequences. The results from some more recent research [Wong08] confirm that: i) every data point in an aggregate, which was sampled for statistical traffic modeling, should be examined if system failure had actually occurred (i.e. *if fractal breakdown had occurred*); ii) this examination can be achieved by using the *Holder exponent*; and iii) if a fractal breakdown had occurred then the channel communication and the message receiver are totally untrustworthy. This leads to the novel argument [Lin09] that a system should invoke its fail-fix mechanism if fractal breakdown has been detected. This invocation is absolutely necessary for time-critical applications, because information accuracy cannot be compromised in, for example, the area of telemedicine.

The two points above together make it difficult, if not impossible, in the current state of the art in the area of trusted communication to aid clinical

telemedicine. This is particularly true when the ontology that supports all the clinical practices is distributed – a distributed knowledge base. This provided the impetus for me to narrow down the scope of my PhD research by excluding deeper exploration of the trusted communication issue in my subsequent work. Otherwise it would have been difficult for me to finish my PhD study within the time and resource constraints.

2.2 Reliable Software Engineering

Software engineering (SE) has a long history and has gone through several eras of paradigm changes. Despite these changes the failure rate of software development projects has remained the same as it was three decades ago – 70% [Coplien04, Ausi00, Standish04, Cheah07]. From the 21st century and on, we cannot simply rely on traditional techniques to improve the software development success rate, as pointed out by [Boehm08]. The reason is that we need to overcome many emerging formidable and ever-evolving challenges that include rapid, uncertain technological changes/emergence; cultural diversity leading to ambiguous understanding of the target system; and hardware and software heterogeneity that prevent interoperability. The paper [Boehm08] sums the formidable challenges nicely, and one of the guidelines is to avoid THWADI ("***that's how we've always done it***"). This guideline is especially suitable for developing the remotely deployable ubiquitous web-based telemedicine systems, which is a new phenomenon of the 21st century [Karr99, Bardram07].

In reality, the THWADI guideline is unavoidable, for computing requirements had evolved rapidly through different eras, following the Moore's Law [Lewis96]: i) Amdahl's era (early 1960s) – synchronizing sequential processes correctly was the focus; ii) Gustafson-Barsis era (mid-1980s) – parallel computing (i.e. High Performance Computing (HPC)) to yield speedup; iii) megacomputing era (mid-1990s) – distributed systems running on the Internet as the platform; and iv) pervasive era (early 2000) – concerns of how mobility of hardware and software entities with support of location-aware capability can be effectively achieved.

Despite the rapid SE evolution driven by various contemporary forces, we find that: i) the traditional Waterfall model can still provide the minimum SE procedural guideline; ii) optimal placement of distributed program tasks correctly and optimally is still a focal issue; and ii) sound synchronization and data/control coherence among distributed tasks and data modules are still essential for obtaining meaningful results. Yet, a new philosophy is needed for configuration control so that the dominant forces of the current era can be balanced for producing qualitative software systems. One of the new factors for modern configuration control framework is the ***enterprise ontology*** (EO) concept [Uschold07]. The argument that the EO approach is essential for producing reliable software can be divided into two fronts:

a) ***Project management***: The EO serves as the standard enterprise SE vocabulary or lexicon so that different collaborating teams of software engineers (i.e. in multisite software development) would have a uniform, unambiguous understanding of a single concept [Cheah07].

b) ***Code generation automation***: If the specification is a collection of semantic paths encoded in the given ontology already, the corresponding target system can be correctly generated in one step. Every function in this target system should work correctly in the context of the given ontology [JWong08b, Wongthongtham09]. This kind of semantic correctness is a basic requirement of software reliability and trustworthiness.

In fact, software applications can be divided into two categories: control-oriented and data-oriented. Control-oriented software directs all the execution flow independent of data availability. On the contrary, data-oriented software works by data availability. The algorithmic languages such as C++ and Java are control-oriented, and the declarative language LISP is typically data-oriented. For data-oriented distributed computing that supports load balancing naturally, the Linda programming language (a form of coordination languages), which was once very popular [Yeung98] is a good example. No matter if the SE goal is for control-oriented or data-oriented applications, the ultimate goal is to ensure that the target system indeed works correctly as expected. To achieve this goal the traditional Waterfall procedure (Figure 2.2.1) can still serve as a

good guideline. The Waterfall model consists of four main repetitive phases for quality control as follows:

a) *Requirement Specification & Analysis*: The natural language specification must be written in an unambiguous way so that the system functionality can be analyzed and consolidated correctly. This representation can be achieved by formal approaches such as the Petri net so that the logical correctness of the specification can be verified quickly at least by simulations.
b) *Design Specification*: The synchronization among collaborating entities (e.g. inter-process and intra-process) should be explicit and unambiguous. The control flow should be verifiable and the mapping between technology and software mechanism should be seamless.
c) *Implementation*: The transformation of the design details into the final artefacts must be meticulous.
d) *Testing and debugging*: The verification and validation processes of the implementations must be supported by appropriate tools in both the alpha (developer) and beta (user) testing phases. This is not easy for the distributed environment because traditional tools are not applicable.

It is not hard to find that the above Waterfall phases are full or flaws in the sense that human errors can be easily and inadvertently introduced in the process, due to the intimate and continuous human involvement. The most difficult part in developing a mission-critical software system is, in fact, to combine the exact understanding of the concept by both the domain expert, who

is the knowledge master, and the software engineer, who is the technology master. Seemingly, the most effective technique to achieve this and produce a reliable piece of software is the ontology-based SE paradigm [Cheah07, JWong08b, Wongthongtham09]. This paradigm has two parts: i) application of appropriate multisite SE development management techniques; and ii) generation of correct software artefacts that exactly match the corresponding functional specifications. The ontology-based SE paradigm actually does not conflict with the Waterfall philosophy. Instead, it absorbs the first two Waterfall phases into the "*semantic specification process*", which is similar in fashion to the MI (meta-interface) approach originally proposed by Nong's [JWong08b]. The Waterfall implementation phase is now made automatic because the final system is generated from the given semantic specification in one step.

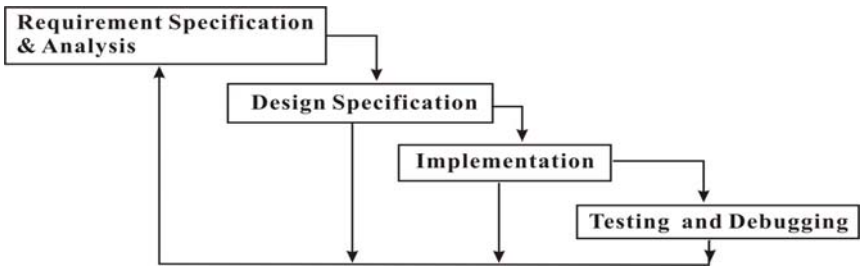


Figure 2.2.1 The generic Waterfall software engineering paradigm

2.3 Concept of Ontology

In philosophy, the ontology of a knowledge boundary is a conceptual schema of what exists within [Witmer04]. Therefore the schema is a nomenclature (i.e. a system of names) of concepts and the associations among

them. These concepts and associations are formally and axiomatically defined to constrain their interpretations (i.e. they form the unique vocabulary). Yet, these ontology-based interpretations can be syllogistic because they are valid logically but not necessarily true/factual. In applied science factual conclusion is usually more important than syllogism, and to this effect, only facts are included. If new facts can be added to the extant domain ontology to keep its knowledge abreast of time, we can say that this ontology has the ability to evolve. In this light, any ontology capable of evolution is a living ontology [JWong08c].

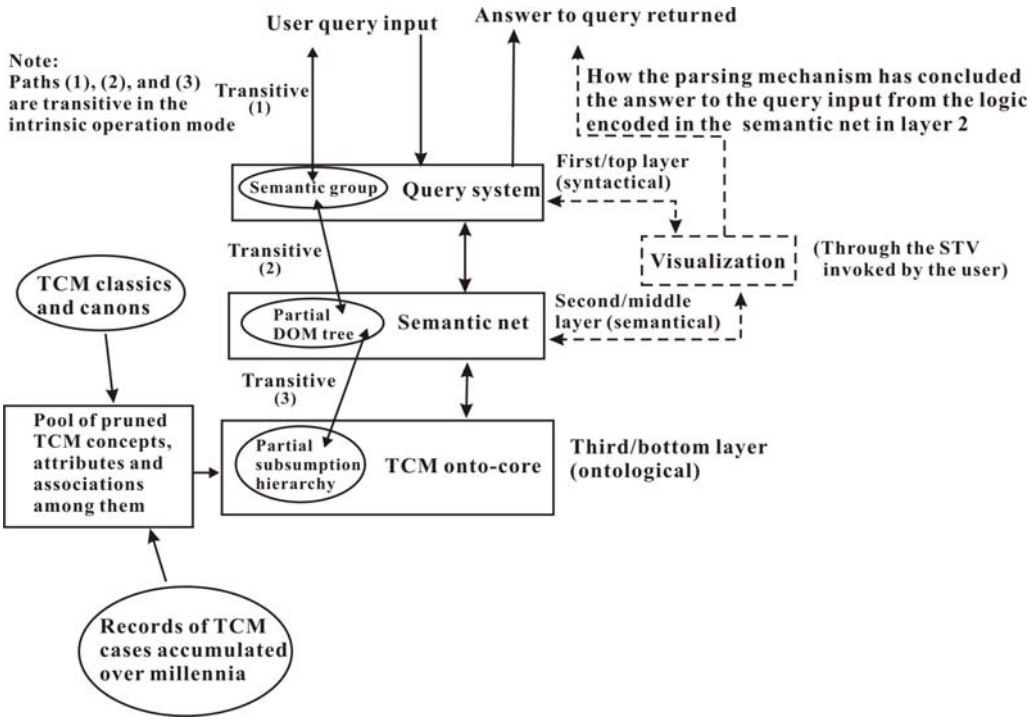


Figure 2.3.1 A 3-layer ontological architecture with cross-layer semantic transitivity

Under the concept of "what exists" within a knowledge boundary, any construct or abstraction (e.g. a program) is ontology, however detailed the

contents within the boundary are. The ontology as such is basically an explicit specification of conceptualization [Gruber93a], which should have three layers of abstractions [Guarino95]:

a) ***Bottom layer*** – This is the representation of the desired knowledge scope, which is the standard of a community, domain or application because it has been agreed by a sufficient number of experts in a process of consensus certification.
b) ***Middle layer*** – This is the "*semantic network*" (or simply semantic net), which correctly and fully represents the knowledge in the ontology underneath. This is the explicit subsumption hierarchy to aid machine processing – logical inference.
c) ***Top layer*** – This is the query system that abstracts the ontology for human understanding and manipulation. Every query should be a semantic path within the explcit subsumption hierarchy.

Any system that has the three layers would work unambiguously only if cross-layer semantic transitivity exists. This means that for any term picked up from any of the three layers, the corresponding representations in the other two should surface in a consistent fashion [Ng08]. Figure 2.3.1 shows the three layers of any typical working ontology-based systems.

The aim of any ontological knowledge base is to support correct, effective communication. Therefore, the contents are annotated by metadata

(e.g. XML, RDF, and OWL [W3Ca, W3Cb]) at the appropriate level to aid unambiguous understanding. The following are some examples:

a) ***Knowledge transfer***: Different levels of knowledge transfer require different metadata system for facilitation. For example, for primary school students simple taxonomies are sufficient but for university students formal representation is important. Therefore, transferring the same piece of communal knowledge means different representations for different levels of recipient groups.

b) ***Interoperability***: This is a different requirement because a piece of knowledge must be understood consistently as a single concept – no multi-representations [Rifaieh06]. This is especially important for applied sciences (e.g. computing, engineering and medicine), where working systems are always synthesized, verified and validated as the final artefacts.

c) ***Consultation*****:** This is for interactive learning, reference, and bridging knowledge gaps in a global sense. For example, in conventional or allopathic medicine different countries may have different definitions for an observed phenomenon P. One country may construe P as defined by two attributes (e.g. $P_1 = \{a_1, a_2\}$), whereas $P_2 = \{a_1, a_2, a_3\}$ is the definition in another country. Then, the degree of similarity between P_1 and P_2 can be defined by the probability term $P_r(P_1 \cap P_2)$ in the following expression: $P_r(P_1 \cup P_2) = P_r(P_1) + P_r(P_2) - P_r(P_1 \cap P_2)$. In order to resolve the similarity and differences of various definitions, as well as language peculiarities in a global scale, meta-thesauri can serve

as an effective means as shown by the ULMS (***Unified Medical Language System*** [UMLS]), developed by the US National Library of Medicine.

The WD^2UHI platform deals with clinical requirements and therefore it always adheres to facts (from classics and case histories). It relies on axioms to contrain interpretations for precision to the benefit of patients and the TCM industry. These axioms can be divided into two sides: i) computing axioms such as $P_r(P_1 \cup P_2) = P_r(P_1) + P_r(P_2) - P_r(P_1 \cap P_2)$; and ii) TCM medical axioms such as the SAME principle (refer the the defintions of terms in the preamble section).

2.4 Ontology-Based Software Engineering

According to the statistics of software development, less than 50% of the projects in the Western world were completed successfully [Standish04]. Even though different software engineering (SE) techniques (e.g. the traditional Waterfall model) have been proposed over the last few decades [Osterweil08, Boehm08, Cheah07], the software project success chance is still 30% today as it was a few decades ago. One of the reason is the SE process is overall expanded to reap the potential advantage of international talents; that is, multisite SE across the continents.

In the beginning, some major corporations had actually reaped some benefits from multisite software development arrangements. There has been a

trend to move software development projects by these corporations to countries where the employee wages are relatively lower. Multisite environments involve multiple teams residing across cities, regions, or even countries speaking different languages to work together in a networked, distributed fashion to develop software [Wongthongtham09]. From multisite software development projects, five major characteristics can be identified [Wongthongtham06a]:

i) Different concepts and terminologies are used by different teams in software engineering and project management
ii) Different levels in understanding the problem domain they are dealing with.
iii) Different training resulting in different levels of knowledge and skills among the supposedly collaborating teams.
iv) Any issues raised by different teams cannot be solved immediately or in time.
v) No sense of ownership about the project and therefore "it is always somebody else's fault"; in this way time is wasted on arguments rather than resolving the project problems.

In fact, the problems of multisite software development increase in both size and depth because collaborating team members and/or project leaders may not be in the same locations to resolve them. In addition, it could be communication problems among team members of different sites as well. This is aggravated by incongruent interpretation of details in documentation, design, or diagrams. Ontology can help resolve the above problems because it provides

distinctive representations of the same concept to be share an ambiguously project teams even they do not speak a common language [Davenport98].

Ontology, in the field of computer science, is an explicit specification of a conceptualization [Gruber93a]. In it, the definitions associate the names of concepts in the universe of discourse (e.g. class, relations and functions) with a description of what the concepts mean. Then, formal axioms are defined to constrain the interpretations and guarantee the well-formed use of these terms [Beuster02]. The domain knowledge of software engineering can be regarded as ontology a well as a well-founded model of reality. In this way, terms in the ontology can be used to analyze the meaning of conceptual models or constructs [Wand99], to check if they accurately reflect the reality.

In order to share the SE domain knowledge publicly and correctly, there must be a common agreement of how information should be communicated correctly. Therefore, ontological commitment is described as the agreement about concepts and relationships among them [Gruber93b]. Once the ontological commitment has successfully been established, in light of semantics of the concepts and relationships, software engineers can to share knowledge unambiguously. The SE ontology is similar to OWL ontological constructs in other domains. It consists of instances, properties, and class [Wongthongtham09]. The instances include specific project data, and binary relations among software engineering concepts.

Intelligent software agents in conjunction with ontology constructs to support multisite software development may be useful. These intelligent agents have three distinctive ability characteristics [Wongthongtham04]:

i) They reason and manage issues that arise in the entire SE course.
ii) They classify attributes, roles, and concepts with the support of ontological constructs (e.g. SE ontology, project management ontology, domain ontology (also known as business ontology) and "issues and solutions" ontology).
iii) They remotely communicate with software developers, in light of classifying questions and issues and providing ontology-based answers autonomously.

2.5 D/P Medical Systems

The TCM (Traditional Chinese Medicine) is protected by law and therefore its practice is common in the Hong Kong SAR. For this reason there is a drive in the Hong Kong SAR to develop web-based computer-aided D/P (diagnosis/prescription) systems. Through the graphical user interface (GUI) of such a system, the physician can diagnose and prescribe more accurately. This is true because the system database contains both TCM knowledge and patients cases accumulated over time from different physicians, hospitals and institutions. The first successful pervasive D/P example in the Hong Kong SAR as well as the rest of China, after my exhaustive literature search, is the Nong's D/P system. This system has been deployed by the user YOT (Yan Oi

Tong), which is one of largest charitable organizations in the SAR that operates clinics and hospitals, among other things. Since its deployment three years ago in the mobile clinics of YOT, the D/P system has been treating hundreds of patients daily. The detailed descriptions of the Nong's D/P system can be found in different documents and publications (e.g. [PTeC06, PTeC07]).

The YOT D/P system is also called the "***Chinese Medicine Vehicle Information System***" (CMVIS) [PTeC06]. It was developed by the traditional Waterfall software engineering approach. In this approach the system development goes through several stages: i) user specification; ii) design; iii) implementation; iv) verification and tests; and v) maintenance and migration. The programming is basically algorithmic and is therefore error-prone [JWong08b]. The CMVIS is made up of three major software modules as follows:

a) ***Graphical User Interface (GUI)*** – This allows the physician to conduct computer-aided "***diagnosis & prescription***" operations. If it is necessary, help can be solicited from remote peer physicians or from the central YOT management.
b) ***Database*** – The database is made up of TCM classical knowledge for clinical practice as well as useful cases that have been accumulated over the past millennia (e.g. [Li16, Circa722BC]. This database was created by many TCM domain experts, who were invited by Nong's, in a collectively manner with consensus for clinical practice. In effect, it is a consensus certification (CC) process within the PuraPharm enterprise, in

which Nong's is a subsidiary. Medical experts from the user side, namely, YOT, also participated in the CC process. This database was a prelude to the subsequent Nong's TCM onto-core construction.

c) ***Inventory*** –This fulfils three requirements: i) YOT herbal inventory, which is virtually linked with the Purapharm Group supply chain and thus manufacturing process; ii) toxicity control and management statistics to be included in the reports by YOT to the Hong Kong SAR Government in light of drug and disease control; and iii) drug and herb quality control by YOT and thus the PuraPharm Group.

After a few years of the Nong's or YOT D/P operation, several pros and cons became apparent [PTeC07, JWong08a, JWong09a, JWong09b]:

a) ***Customization difficulty***: The YOT success has spurred many requests to have specific customized versions of the D/P system by customers from Mainland China and different countries across the globe (e.g. U.K. and Australia). Yet, the original algorithmic programming approach makes the customization error-prone and difficult. The difficulty is more acute in the understanding of medical terms especially. The customization process normally involves the PuraPharm software experts and teams from the clients outside the SAR (i.e. multisite development), and misunderstanding due to multi-representations of terms make the final product often disfunctional. In fact, this is exactly what the UMLS [UMLS] has faced and tries to resolve the multi-representation problems by using thesauri and meta-thesauri.

b) ***Evolution problems***: The database of the D/P system, though "*consensus-certified*" and full of time-honoured classical TCM knowledge and medical records, is static and not equipped to evolve automatically with time. That is, the clients, who own customized D/P variants, cannot keep the system abreast of scientific TCM advances.

c) ***Semantic transitivity murkiness***: It is difficult to assess if semantic transitivity exists in the D/P database. Semantic transitivity means if $a \to b$, then $b \to a$ is also true, where $\to$ for 'implying". This makes any discovery of new knowledge difficult.

In order to resolve the three main problems above, Nong's has proposed the following:

a) ***Meta-Interface paradigm***: The argument is that customization of any D/P variants can be easily, quickly, and accurately achieved by using an iconic specification, provided that the functionality of every icon is a semantic path encoded in the supporting enterprise TCM ontology core (onto-core) for clinical practice. The client simply selects and/or creates icons and puts them into a single MI specification, which is the sole input to the "*system generator*". The MI paradigm was simply a shell (lacking in details for implementation) when I started this PhD research.

b) ***Enterprise TCM onto-core***: This onto-core's creation is a massive consensus certification process, but Nong's succeeded by converting the old database in the extant D/P system into a true TCM ontology construct. This has provided the necessary basis for meaningful

development of the MI paradigm afterward. In fact, it is part of my PhD research pursuit to enrich, implement, and verify the MI paradigm by using the newly created Nong's enterprise TCM onto-core as the basis with permission. This pursuit is necessary for quickening my research, for the MI paradigm enables me to construct prototypes for conducting experiments quickly, correctly, and in a trustworthy manner. The criterion for achieving the necessary trustworthiness is that TCM physicians should be able to use my prototypes to treat patients in a computer-aided manner in the real clinical environments.

c) ***Semantic transitivity visualization***: Nong's proposed to verify the semantic transitivity by visualization but without providing any details for implementation. This has inspired my proposal of the novel STV (semantic transitivity visualizer) for checking the cross-layer semantic verification in the enterprise Nong's TCM onto-core anytime and anywhere.

Yet, Nong's and PuraPharm experts have never addressed the following issues, which instead have become focal investigations in my PhD scope:

a) ***TCM onto-core evolution***: The issue is how to help the onto-core evolve in a continuous and real-time manner. Meanwhile, the evolution should not violate the principle of consensus certification so that global trust of the onto-core is not breached.

b) ***Herbal discoveries***: It was a PuraPharm wish but was never seriously considered until my PhD pursuit, which is TCS (Teaching Company

Scheme) based. The difficulty facing the company experts is how to define the scope for meaningful herbal discoveries. Besides, any discovery is not trustworthy unless it adheres to the classical TCM clinical and/or pharmacological principles. Therefore, the first step in my exploration of this issue is to define usable principle(s) with help from TCM domain experts. Once the principle(s) has been established, the next step is to address the question of what techniques can help achieve the discovery goals. In addition, these goals may have to be classified as well.

2.6 Text mining

Text mining or *knowledge discovery from text* (KDT) was mentioned for the first time by Feldman [Feldman95]. The subject deals with machine-based analysis of text. Text mining makes use of the techniques from information retrieval, information extraction, natural language processing (NLP), and connects them with algorithms or methods of knowledge discovery, data mining, machine learning and statistics. It has also been defined in different perspective in the following areas:

a) *Information extraction* – It deals with extraction of facts from texts.
b) *Text data mining* – It deals with extracting useful data patterns and statistics from texts [Nahm02, Gaizauskas03].
c) *Knowledge discovery* – It extract semantic data or knowledge according to the *knowledge discovery process model* [CISP99]. It is made up of a

series of partial steps that involves different techniques suitable at the time, for pattern matching and classification operations [Kodratoff99, Hearst99, Hidalgo02].

In fact, we can summarize that information extraction and text data mining are two different forms of knowledge discovery procedures.

Current research in the area of text mining concentrates on tackling problems of text representation, classification, clustering, information extraction or the search for and modeling of hidden patterns [Hotho05]. It has been revealed that both the selection of characteristics and the influence of domain knowledge play an important role in the process. Thus, an adaptation of the known data mining algorithms to suit particular text data is usually necessary. The following paragraphs introduce different areas that are actually part of the text mining context:

a) ***Information retrieval***: It is the science of searching for: i) documents; ii) information within documents; iii) metadata about documents; iv) data in relational databases; and v) World Wide Web (WWW) data [Singhal01]. To achieve this goal, statistical measures are used for automatic processing and comparing text data. Information retrieval, in the broader sense, deals with the entire range of information processing, from data retrieval to knowledge retrieval. The retrieving process begins when a user has entered a query into the system. Queries are formal statements of information needs (e.g. search strings in web search

engines). Actually information retrieval does not require a query to uniquely identify a single object in the collection. Instead, several objects that match the query with varying relevancy may be retrieved.

b) ***Natural language processing (NLP)***: It is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages [Manning99]. An NLP system converts information from computer databases into readable human language. There could be a more formal representation, such as parse trees or first-order logic structures, to enable help computer programs in the manipulation process. The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system. The evaluation determines if the system answers the goals of its designers, or meets the needs of its users (to a certain extent).

c) ***Information extraction***: It is a type of information retrieval, and the goal is to automatically extract structured information, which is usually categorized and contextually/semantically well-defined data. The extraction is usually made from unstructured machine-readable documents [Sarawagi08]. The task of information extraction naturally decomposes into a series of processing steps, typically including: tokenization, sentence segmentation, part-of speech assignment, and the identification of named entities (i.e. name of person, location, and name of organization). At a higher level, phrases and sentences may have to be parsed, semantically interpreted, and integrated. Finally the required

pieces of information are collated and then entered into the database.

2.7 Neural Network

The result of my literature survey [Ren02, Karray02, Ghosh03, Connor94, Mitra94, Lin06a, Zhao03] indicates that the neural networks (NN) approach is one of the popular and mature soft computing techniques. It has a wide range of application. The other soft computing techniques include genetic algorithms, fuzzy logic, and Bayesian approaches. The popularity of the NN approach associates with the following reasons:

a) ***Reusability***: A NN construct is usually generic. Once verified the same construct can be trained to assume to any assigned intelligent roles [Zhao03].
b) ***Simplicity***: A NN construct is usually simple even when the neurons are fully connected. Therefore, it is easy to program and less error-prone.
c) ***Data-orientation***: The logical points inside a NN will converge to the logical operation as required by the set of training data. For example, the same logical point a in Figure 2.7.1 may converge to an AND operation in one training session but an OR operation in another. This contrasts the *algorithmic programming paradigm* in which point a must be defined logically and clearly for implementing the supporting software module. That is, for the AND and OR operations two separate software modules may be required.
d) ***Versatility***: Every NN construct is a building block, and it may be

combined with its clones or other constructs to form a larger, more complex NN configuration.

e) ***Adaptability***: This can be divided into two views: i) the activation function (e.g. Sigmoid) of a neuron can be replaced as required; and ii) the input parameters to a neuron can be weighted and normalized according to the need.

f) ***Optimization***: Feed-forward neural networks can be effectively optimized for the particular operation or even a designated period of operation. One of the formal adherences for effective NN optimization is the Hessian matrix (refer to section 6.6 in Chapter 6). If the Eigen values of this matrix is positive definite, then $f(x)$ definitely have a minimum point of a set of minima. This is the conceptual basis for pruning (the minimum or minima), but the actual application is subject to the domain of application and the activation function adopted at the time.

g) ***Commodity***: There are many NN constructs in the form of freeware in the public domain and have rich user experience. We can select and try out different NN freeware until the suitable one is found.

h) ***Accuracy***: The accuracy of the NN (backpropagation) inference is not directly proportional to the number of neurons in the hidden layer. As long as the number of the hidden neurons is twice that of the input neurons, the result is usually accurate [Hagan96, Gallant92].

The inspiration for my investigation to possibly adopt the NN approach as the knowledge classification technique to support Type 2 herbal discoveries

came from the Nong's Company Limited. When diagnosis/prescription (D/P) system for mobile-clinic was being developed by Nong's, the Waterfall software engineering process was very laborious, difficult and error-prone. The difficulty particularly came from the process in customizing D/P versions for different clients. In this process programming errors were introduced inevitably, inadvertently and repeatedly. The enlightenment gained from the customization process is that software faults and errors can be greatly reduced if the data-oriented programming paradigm is adopted. This particularly true if the NN can converge automatically to suit the peculiarity of the input data in the locale of operation [Mitra94]. As a result, software errors are intrinsically eliminated by training the NN without further programming effort required. This becomes the rationale for adopting the NN approach as the knowledge classification techniques in this thesis.

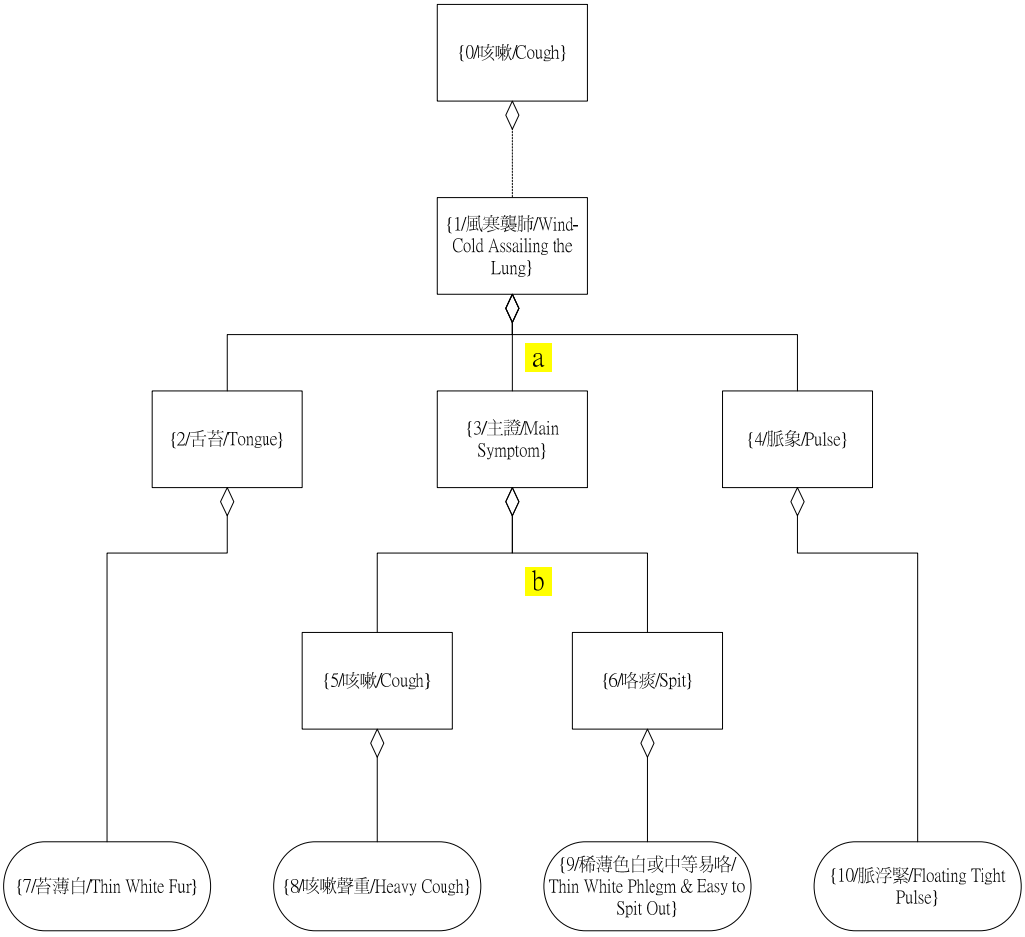


Figure 2.7.1 UML organization of raw clinical data

2.8 Conclusion and Connective Statement

This chapter reviews all the relevant literature so that the scope and the roadmap for the research can be clearly defined. The important literature areas that have been scrutinized include: trusted communication, reliable software engineering, concept of ontology, ontology-based software engineering, D/P medical systems, text mining, and neural network. In fact, the conclusions drawn from the literature review has helped narrow down the scope of the research. For example, deeper pursuit in the area of trusted communication was ruled out because of the time constraints imposed on the PhD study. The

literature review is undoubtedly an important step that contributes to the success of this thesis.

Since the success of the *WD²UHI* project depends on the correctness of the given core ontology, it is necessary in this research to critically scrutinize and analyze the metadata system of the Nong's master TCM onto-core, which is XML-annotated. The metadata system scrutiny and analysis have contributed to the contents of the next chapter – ***Chapter 3 The Realm of Metadata Usage***.

2.9 Key References

[Ausi00] Australian Bureau of Statistics, Finance, Australia 2000 Special Article – Information Technology and Telecommunications in Australia, 2000, http://www.abs.gov.au/Ausstats/abs@.nsf/0/9053E0EB512D0DDC4CA256F2A0007346F?Open

[Bardram07] J.E. Bardram and H.B. Christensen, Pervasive Computing Support for Hospitals: An Overview of the Activity-Based Computing Project, IEEE Pervasive Computing, Vol. 6, No. 1, January 2007, 44-51

[Beuster02] G. Beuster, Ontologies Talk Given at Czech Academy of Sciences, 2002, http://www.uni-koblenz.de/~gb/papers/2002_intro_talk_ontology_bang/agent_ontologies.pdf

[Boehm08] B. Boehm, Making a Difference in the Software Century, IEEE Computer Society, Vol. 41, No. 3, March 2008, 32-38

[Braden98] B. Braden, Recommendations on Queue Management and Congestion Avoidance in the Internet, RFC2309, April 1998

[Chatranon04] G. Chatranon, M.A. Labrador and S. Banerjee, A Survey of TCP-friendly Router-Based AQM Schemes, Computer Communications, Vol. 27, No. 15, September 2004, 1424-1440

[Cheah07] C. Cheah, Ontological Methodologies - From Open Standards Software Development to Open Standards Organizational Project Governance, Computer Science and Network Security, Vol. 7, No. 3, March 2007

[Circa722BC] Yellow Emperor's Canon of Internal Medicine (Huang Di Nei Jing), China, Circa 722 B.C.

[CISP99] Cross Industry Standard Process for Data Mining, 1999, http://www.crisp-dm.org/

[Connor94] J.T. Connor, D. Martin and L.E. Atlas, Recurrent Neural Networks and Robust Times Series Prediction, IEEE Transactions on Neural Networks, Vol.5, No. 2, March 1994, 240–253

[Coplien04] J. Coplien, Organizational Patterns: Beyond Technology to People, Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, 2004

[Cottrell99] L. Cottrell, M. Zekauskas, H. Uijterwaal and T. McGregor, Comparison of Some Internet Active End-to-End Performance Measurement Projects, 1999, http://www.slac.stanford.edu/comp/net/wan-mon/iepm-cf.html

[Crovella97] M.E. Crovella and A. Bestrvros, Self-similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Transaction son Networks, Vol. 5, No. 6, December 1997, 835–846

[Davenport98] T.H. Davenport and L. Prusak, Working Knowledge: How Organizations Manage What They Know, Harvard Business School Press, 1998

[Feldman95] R. Feldman and I. Dagan, KDT - Knowledge Discovery in Texts, Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD), Montreal, Quebec, Canada, 20-21 August 1995, 112–117

[Gaizauskas03] R. Gaizauskas, An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications, 2003, http://www.itri.bton.ac.uk/projects/euromap/

[Gallant92] A.R. Gallant and H. White, On Learning the Derivatives of an Unknown Mapping and Its Derivatives Using Multiplayer Feedforward Networks, Neural Networks, Vol. 5, 1992

[Ghosh03] A. Ghosh and S. Tsutsui, Advances in Evolutionary Computing: Theory and Applications, Springer, 2003

[Gruber93a] T.R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, No. 2, 1993, 199-220

[Gruber93b] T.R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Proceedings of the International

Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation, Padova, Italy, 17 March 1993

[Guarino95] N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, 1995, 25-32

[Hagan96] M. Hagan, Neural Network Design, PWS Publishing Company, 1996

[Hearst99] M. Hearst, Untangling Text Data Mining, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), 15 October 1999

[Hidalgo02] J.M.G. Hidalgo, Tutorial on Text Mining and Internet Content Filtering, Tutorial Notes Online, 2002, http://ecmlpkdd.cs.helsinki.fi/pdf/hidalgo.pdf

[Hotho05] A. Hotho, A. Nurnberger and G. Paab, A Brief Survey of Text Mining, GLDV-Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, 2005, 19-62

[JWong07] J.H.K. Wong, Advanced and Research Topics in Parallel and Distributed Computing, Technical Report COMP6813, Department of Computing, April 2007

[JWong08a] J.H.K. Wong, T.S. Dillon, A.K.Y. Wong and W.W.K. Lin, Text Mining for Real-time Ontology Evolution, Data Mining for Business Applications, Springer, 2008, ISBN: 978-0-387-79419-8, 143-150

[JWong08b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon**,** An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008

[JWong08c] J.H.K. Wong, W.W.K. Lin and A.K.Y. Wong, Real-Time Enterprise Ontology Evolution to Aid Effective Clinical Telemedicine with Text Mining and Automatic Semantic Aliasing Support, Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, 11-13 November 2008

[JWong09a] J.H.K. Wong, A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), 2009

[JWong09b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong, Enterprise-ontology-driven TCM (Traditional Chinese Medicine) Telemedicine System Generation, Proceedings of the 4th International Food – New Horizons in Chinese Medicine and Health Food Symposium, Hong Kong, 29-30 October 2009

[Karr99] J.F. Kaar, International Legal Issues Confronting Telehealth Care, Telemedicine Journal, March 1999

[Karray02] F. Karray, F. Gueaieb and A. Al-Sharham, The Hierarchy Expert Tuning of PID Controllers Using Tools of Soft Computer, IEEE Transactions on System, Man and Cybernetics, Vol. 32, No. 1, 2002, 77–90

[Kodratoff99] Y. Kodratoff, Knowledge Discovery in Texts: A Definition and Applications, Lecture Notes in Computer Science, 1999, 1609 – 1629

[Lewis96] T. Lewis, The Next 10000 Years: Part 1, IEEE Computer Society, Vol. 29, No. 4, 1996, 64-70

[Li16] S.Z. Li, Canon on Materia Medica (Ben Cao Gang Mu), 16th Century, China

[Lin06a] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, Application of Soft Computing Techniques to Adaptive User Buffer Overflow Control on the Internet, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 36, No. 3, May 2006, 397-410

[Lin08] W.W.K. Lin, J.H.K. Wong and A.K.Y. Wong, Applying Dynamic Buffer Tuning to Help Pervasive Medical Consultation Succeed, Proc. of the 1st International Workshop on Pervasive Digital Healthcare (PerCare), Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications, Hong Kong, 17–21 March 2008, 675-679

[Lin09] W.W.K. Lin, J.H.K. Wong and A.K.Y. Wong, A Novel Real-Time Traffic Sensing (RTS) Model to Improve the

Performance of Web-based Industrial Ecosystems, IEEE Transactions on Industrial Electronics (TIE), 2009

[Manning99] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 978-02-6213-360-9

[Mitra94] S. Mitra and S.K. Pal, Self-Organizing Neural Network as a Fuzzy Classifier, IEEE Transactions on Systems, Man and Cybernetics, Vol. 24, No. 3, 1994, 385-399

[Nahm02] U. Nahm and R. Mooney, Text Mining with Information Extraction, Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002

[Ng08] S.C.S. Ng and A.K.Y. Wong, RCR – A Novel Model for Effective Computer-Aided TCM (Traditional Chinese Medicine) Learning over the Web, Proceedings of the International Conference on Information Technology in Education (CITE), Wuhan, China, July 2008

[Osterweil08] L.J. Osterweil, C. Ghezzi, J. Kramer and A.L. Wolf, Determining the Impact of Software Engineering Research on Practice, IEEE Computer Society, March 2008, 39-49

[Paxson95] V. Paxson and S. Floyd, Wide-area Traffic: The Failure of Poisson Modelling, IEEE/ACM Transactions on Networking, Vol. 3, No. 3, 1995, 226 – 244

[Podlipnig03] S. Podlipnig and L. Bozormenyi, A Survey of Web Cache Replacement Strategy, ACM Computing Surveys, Vol. 5, No. 54, December 2003, 374 - 398

[PTeC06] Service Description: YOT Chinese Medicine Vehicle Information System Project, PolyU Technology & Consultancy Company Limited, 2006

[PTeC07] Service Description: A Feasibility Study on the Effective Generalization of the Present PP-N's Diagnostic/Prescription (D/P) System into Mobile-Business Framework, PolyU Technology & Consultancy Company Limited, 2007

[Ren02] F. Ren, Y. Ren and X. Shan, Design of a Fuzzy Controller for Active Queue Management, Computer Communications, Vol. 25, 2002, 874–883

[RFC2828] Internet Security Glossary, 2000, ftp.isi.edu/in-notes/rfc2828.txt

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[Sarawagi08] S. Sarawagi, Information Extraction, FnT Databases, Vol. 1, No. 3, 2008

[Singhal01] A. Singhal, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, 2001, 35–43, http://singhal.info/ieee2001.pdf

[Standish04] The 3rd Quarter Research Report: Chaos Demographics, The Standish Group International, 2004, http://standishgroup.com/sample_research/ darts_sample.php

[UMLS] http://umls.nlm.nih.gov/

[Uschold07] M. Uschold, M. King, S. Moralee and Y. Zorgios, The Enterprise Entology, Artificial Intelligence Applications Institute, University of Edinburg, UK, 2007, http://citesee.ist.psu.edu/cache/papers/cs/11430/ftp:zSzzSzftp.aiai.ed..ac.ukzSzpubzSzdocumentszSz1998zSz98-ker-ent-ontology.pdf/uschold95enterprise.pdf

[W3Ca] W3C, Ontology Definition MetaModel, 2005, http://www.omg.org/docs/ad/05-08-01.pdf#search='Ontology%20Definition%20Metamodel

[W3Cb] W3C, Web Service Architecture (Working Paper), http://www.w3.org/TR/ws-arch/

[Wand99] Y. Wand, V.C. Storey, and R. Weber, An Ontological Analysis of the Relationship Construct in Conceptual Modeling, ACM Transactions on Database Systems, Vol. 24, No. 4, 1999, 495-528

[Witmer04] G. Witmer, Dictionary of Philosophy of Mind-Ontology, May 2004, http://www.artsci.wustl.edu/~philos/MindDict/ontology.html

[Wong01] A.K.Y. Wong, T.S. Dillon, W.W.K. Lin and T.W. Ip, M^2RT: A Tool Developed for Predicting the Mean Message Response Time for Internet Channels, Journal of Computer Networks, Vol. 36, 2001, 557-577

[Wong03] A.K.Y. Wong, M.T.W. Ip and R.S.L. Wu, A Novel Dynamic Cache Size Adjustment Approach for Better Data Retrieval

Performance over the Internet, Computer Communications, Vol. 26, 2003, 1709-1720

[Wong08] A.K.Y. Wong, T.S. Dillon, and W.W.K. Lin, Harnessing the Service Roundtrip Time over the Internet to Support Time-Critical Applications – Concept, Techniques and Cases (invited and contracted by Nova Science Publishers, Incorporated, New York, Feburary 2008

[Wongthongtham04] P. Wongthongtham, E. Chang and T.S. Dillon, Ontology-based Multi-agent System to Multi-site Software Development, Proceedings of the Workshop on Quantitative Techniques for Software Agile Process, Newport Beach, California, USA, November 2004

[Wongthongtham06a] P. Wongthongtham, E. Chang, T.S. Dillon and I. Sommerville, Ontology-based Multi-site Software Development Methodology and Tools, Journal of Systems Architecture, Vol. 52, No. 11, 2006, 640-653

[Wongthongtham09] P. Wongthongtham, E. Chang, T.S. Dillon and I. Sommerville, Development of a Software Engineering Ontology for Multisite Software Development, IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 8, August 2009, 1205-1217

[Wu06a] R.S.L. Wu, A.K.Y. Wong and T.S. Dillon, A Novel Dynamic Cache Size Tuning Model with Relative Object Popularity for Fast Web Information Retrieval, Journal of Supercomputing, 2006

[Wu06b] R.S.L. Wu, W.W.K. Lin and A.K.Y. Wong, Harnessing Wireless Traffic is an Effective Way to Improve Mobile Internet Performance, Proceedings of the 1st Australian Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless), Sydney, Australia, March 2006

[Yeung98] D.S. Yeung and A.K.Y. Wong, The OORHS: A Conceptual Framework that Provides Easy and Reversible Distributed Programming, International Journal of Computer Systems, Science and Engineering, Vol. 13, No. 15, 1998, 289 - 301

[Zhao03] W. Zhao, R. Chellappa, P.J. Phillips and A. Rosenfeld, Face Recognition: A Literature Survey, ACM Computing Surveys, Vol. 35, No. 4, 2003, 339–458

Chapter 3 The Realm of Metadata Usage

This chapter presents the results of our investigation of the metadata usage in defining the TCM ontology. This is absolutely relevant to this research because of the following goals:

a) The defined ontology should be accurate and easily changeable (addition, deletion, and update modifications).
b) The language/tool use should be a kind of de facto standard to reduce the chance of obsolescence.
c) It should support semantic transitivity as well as easy logical extensions.

3.1 Introduction

Metadata is the "data about data", which usually comes from a specific domain. Yet, the territory of metadata systems is very wide because there are many such systems, including the one supported by the W3C (World Web Consortium), which has proposed XML, RDF, and OWL. A metadata can be a simple content item or a construct of multiple content items. It facilitates understanding and management of data items as well as the associations among them. For example, it is often used to support locating and retrieving information. Metadata systems have existed for a long time in various forms; the table of contents (TOC) for a book is the one of the most common metadata to describe the book's organization (e.g. authors, chapters, and publication date).

In a library, the TOC (Table of Contents) metadata helps librarians organize the shelves to enable easy location of any book.

More sophisticated definitions of metadata keep appearing out of need. For example, metadata may be regarded as structured, encoded data that describes the characteristics of information-bearing entities. The aim is to aid identification, discovery, assessment, and management of the described entities [ALA99]. For other applications, metadata is a set of optional structured descriptions that are publicly available to explicitly assist in locating objects [Bultermann04]. Therefore, it is reasonable to generalize that metadata as a means to describe the structure, management, and usage of information in a domain.

Metadata applications cover a wide spectrum, and a few examples of their use include:

a) ***Speeding up and enriching the searching of web objects:*** This is typified by Web browsers, P2P applications and file management software. Usually metadata improves the speed of file searching.
b) ***Linking files:*** As a result, documents can be converted into an electronic format that eases the storage in the document repository such as Documentum; this facilitates file retrieval process.
c) ***Bridging the semantic gaps, because the relationships between data items can be axiomatically specified:*** In this way it helps complex information retrieval operations. For example, if the search engine

acquired the knowledge that “Aristotle” was a “Greek philosopher”, users may provide a search query on “Greek philosopher” with a link to a Web page about “Aristotle”, even if the exact term “Greek philosopher” never occurs in that page. This approach is generally called “Knowledge Representation”, which is usually of special interest to the Semantic Web and artificial intelligence.

3.1.1 Problems of Formability, Ambiguity and Implicit Semantics

This PhD thesis makes use of the experience of the successful original Nong’s TCM (Traditional Chinese Medicine) diagnosis/prescription (D/P) telemedicine system for mobile-clinic (MC) deployments and applications. The original Nong’s D/P system, however, was not based on the ontology concept even though the creation of the master knowledge base for the successful D/P system was also a form of consensus-certification. In this form, a group of TCM experts pruned and agreed to consolidate the final contents for the master knowledge base, which then became the basis for any subsequent D/P variants to be customized for individual clients. The customization process was based on the Waterfall model, which provided the basis for the creation of the first and original D/P system. The Waterfall model involves too much human intervention and is thus prone to errors. For this reason, Nong’s had proposed the **novel** MI (meta-interface) paradigm (*but only a “shell” concept without implementation details*), and transformed the original master TCM D/P knowledge base into the enterprise TCM ontology core (onto-core). The metadata system for annotations in this onto-core is the XML due to its

flexibility. The problem was that the MI paradigm proposed by Nong's is very high-level and did not contain sufficient details for implementation. The MI paradigm is, however, superbly novel because the aim is to generate a workable D/P system variant for the given iconic specification. The local TCM onto-core of the customized D/P variant is a subset of the Nong's master TCM onto-core. In this way all the cognate customized D/P system variants should be interoperable to a varying degree.

The main advantage of employing semantically based customization such as the MI paradigm is that ontology usually has a subsumption hierarchy of explicit semantic paths. If a software module basically executes a collection of explicit semantic paths that are selectively "extracted" from the hierarchy, then it has explicit semantics because every execution has the corresponding semantic path in a consistent fashion. In contrast, an ordinary organization of raw clinical data may have very implicit semantics as shown in Figure 3.1.1.1. In this example, there are only ten clinical entities uniquely identified. In the set {0/咳嗽/Cough} the unique symbols, "0", "咳嗽" and "Cough" have the same connotation – "0" as the identifier. Solid-line arcs (e.g. between "3" and "4") indicate strong associations between entities. The arcs are bi-directional because their traversals should be logically transitive. The entities can be placed anywhere within the database but their retrieval depends on the "***retrieval algorithm (RA)***" implemented as part of the D/P system software.

The RA is working by predicate logic; for example: (i) if "8" is true then "5" is true; (ii) if "9" is true then "6" is true; and (iii) if "5" and "6" are true

then “3” is true – the logic for point “b” is then a logical “AND” function. For the same Figure 3.1.1.1, if another RA interprets point “b” as a logical OR (i.e. if “5” or “6” are true then “3” is true), then this system is not physically compatible with the previous one. In our research, if one has to scrutinize the RA code in order to find out the exact meaning of “b”, then the RA code has implicit semantics – diagrammatically (i.e. Figure 3.1.1.1) the two systems look similar superficially. In the same light, the implementation for the logical point “a” may vary from one system to another – these systems are incompatible variants. Yet, once the predicates for the different logical points are axiomatically defined, a variant is formal.

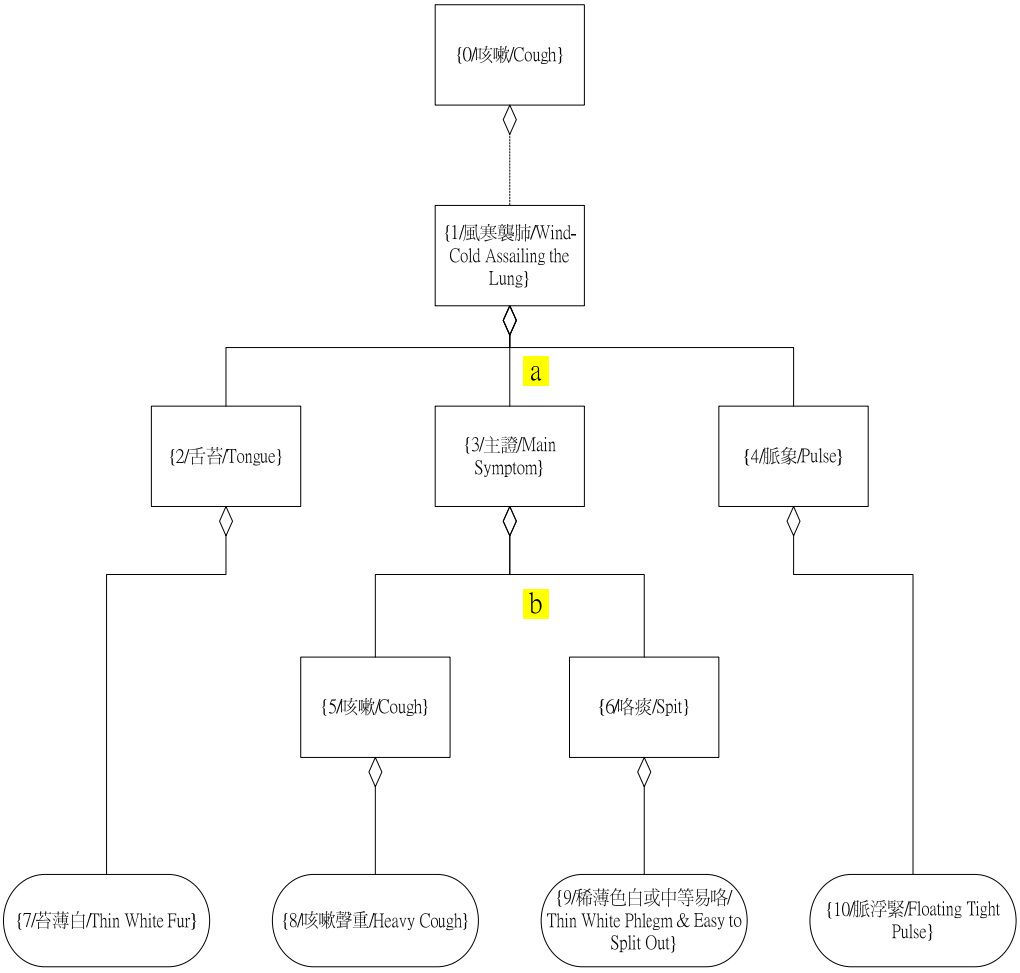


Figure 3.1.1.1 Organization of raw clinical data (e.g. in the original Nong's D/P system) ("0" and "5" are logically different as "0" is an illness name while "5" is a symptom.) (excerpt of Figure 2.7.1)

Figure 3.1.1.2 is the formal Petri net representation of the hierarchy in Figure 3.1.1.1, where logical point "a" and "b" have the OR and AND implementations respectively. In fact, this Petri net can be used to simulate/check the logical correctness of Figure 3.1.1.1. This Petri net (PN) is bipartite and is made up of three sets of symbols, arcs (A), transitions (T), and places (P), as follows:

a) PN = [P, T, A]

b) P = [{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}]

c) T = [T1, T2, T3, T4, T5, T6, T7, T8, T9]

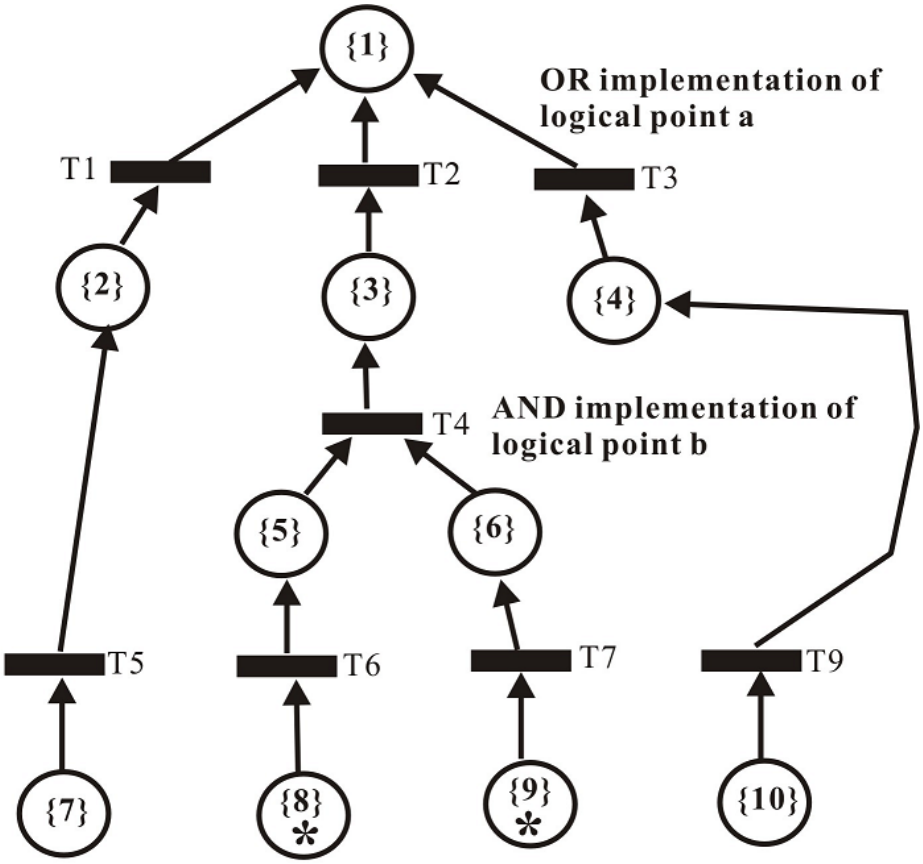


Figure 3.1.1.2 Formal Petri net representation of Figure 3.1.1.1

In Figure 3.1.1.2, T1, T2 and T3 represent a logical OR operation and T4 is the logical AND operation. Places {8} and {9} are initialized with two tokens. When a token is in a place then the place assumes the "true" state. T4 fires as long as both places {5} and {6} have token(s) but in a unpredictable fashion. The transition firing is atomic, and this means that at anytime only one transition can fire even it takes zero time to complete. In Figure 3.1.1.2, with the tokens in places {8} and {9}, the successive firing will generate the final

token in place {1}, which represents the logical conclusion. In fact, the traversals from places {8} and {9}, through places {5}, {6} and then {3}, for the token to reach place {1}, makes the "parsing" process. The traversals can be depicted clearly by the reachability graph in Figure 3.1.1.3. In this graph the state vector changes with time when the transitions fire in an indeterminate fashion. Every vector (e.g. M0) indicates the current state at the time, with a "1" to indicate the presence of a token – a "logically true" state. The leftmost bit position in the vector is for the place {1} and the rightmost for the place {10}.

If we reverse the logical operations for the point "a" (i.e. to become logical AND) and "b" (i.e. to become logical OR), then the corresponding Petri net and its reachability graph will be very different. Therefore, logically speaking Figure 3.1.1.2 differs from the Petri net with the reversed (i.e. "a" becomes "AND" and "b" becomes "OR") variant. The variants are, however, *both formal and axiomatic, <u>but not</u> logically compatible*.

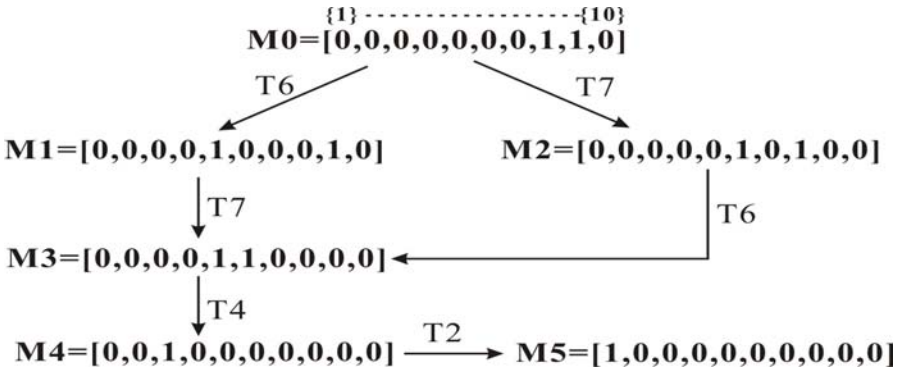


Figure 3.1.1.3 Reachability graph for the Petri net in Figure 3.1.1.2

3.1.2 W3C and XML

W3C the strong promoter: Tim Berners-Lee in 1989 proposed the World Wide Web (WWW) concept, which was then supported by the World Wide Web Consortium (W3C) team since 1992. The W3C mission is to help standardize the semantic web development. Since then, the W3C has been proposing various protocols and guidelines. The aim is to aid long-term web growth and make the web live up to its full potential [Fensel03]. Since 1999 W3C has proposed more than 110 recommendations, which provide the basis for open-forums and discussions. The basic requirement for the web to succeed is to have compatible web technologies provided by different vendors. This requirement, from the W3C point of view, is "*web interoperability*." With the support from powerful web languages and protocols, this requirement would avoid web market fragmentation.

XML (Extensible Markup Language) metadata: It is a generic specification or metadata system for creating extensible, customized markup languages. It lets users mark up and define data constructs in their own ways. In this way it facilitates sharing of pre-defined structured data across different information systems, particularly over the Internet [Bray04]. XML was developed by the XML Working Group (known as the SGML (Standard Generalized Markup Language) Editorial Review Board originally) that was formed under the World Wide Web Consortium (W3C) in 1996. The design goals of XML include [Fensel03]:

i) XML shall be straightforwardly usable over the Internet.

ii) XML shall support a wide variety of applications.

iii) XML shall be compatible with SGML.

iv) It shall be easy to write programs that process XML documents.

v) The number of optional features in XML is to be kept to the absolute minimum, ideally zero.

vi) XML documents should be human-legible and reasonably clear.

vii) The XML design should be prepared quickly.

viii) The design of XML shall be formal and concise.

ix) XML documents shall be easy to create.

x) Terseness in XML markup is of minimal importance.

XML allows users to structure data with regards to their content rather than their presentation [Yergeua04]. They are behavioral in nature and provide the tool to ***map tacit knowledge into explicit knowledge***. Therefore, large amounts of data could be analyzed, diagnosed, seen, and understood by both humans and machines. In this light, the XML is useful for ontological construction because, according to Gruber [Gruber93a], ontology is the explicit specification of a conceptualization, which should be human-understandable and machine process-able. XML, which limits and controls the vocabularies and namespaces, provides the ***first level*** of metadata. The following is a simple XML annotation example: <price currency="HKD">100</price>. In this example, however, the labels do not bear any meaning for machines but humans. To reduce such limitations, W3C came up with a better solution – XML schema. The XML schema provides structured descriptions for XML documents, which

typically are expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntax constraints imposed by XML itself. As an example, it can specify that the content of a “price” label should be a rational number. An XML schema provides a view of the document type at a relatively high level of abstraction. The basic XML is poor in data type specification.

The schema does not help specify the meaning of tags, despite the feature to build hierarchies of element types. This hierarchy contains no conceptual knowledge, but only functions as a syntactic shortcut to allow reuse of complex definitions. A naïve way of relating the ontology with XML documents is to match the labels in a XML document syntactically with the names of concepts/properties in the ontology that are associated with it. But, what role the data in the XML document fulfils is not clear at all. Referring to the previous example <price currency=“HKD”>100</price>, it is not yet clear whether “price” is the data type of “100”, or the value of price is “100”. For reliable use on the Semantic Web, it is necessary to interpret data and its type unambiguously. This inspired the emergence of the RDF data model to support unambiguous data interpretation.

3.2 Semantic Web

The *semantic web concept* is an evolving extension of the World Wide Web in which the semantics of information and services on the Web are defined. Via some types of interface intelligence, it is possible for the Web contents to

be understood and to satisfy the information retrieval by people and/or machine [Berners01, W3Cc]. It was the vision of Tim Berners-Lee's to use the Web as a universal medium/platform for data, information and knowledge exchanges [Herman08]. Before this is achievable we need to organize the web and desired semantics (i.e. web semantics) [Taniar06]. One way to support this organization is the ontological approach. From the viewpoint of philosophy [Corazzon04], ontology is a knowledge base that contains consensus-certified items and has an evolvable boundary. Usually the concepts, attributes, and the associations among them within the ontology are descriptive, and the formality of the knowledge stems from the process of consensus certification, which constrains interpretation. Yet, it is difficult to force the same constraints for all knowledge domains. One approach to resolve such domain-related "multi-representations" for the same items is formulating domain "views", which act like filters to accentuate the desired features and mask out the irrelevant ones.

The concept of problem domains and their applications of specific salient features arouse the concept of portable ontology specification. Gruber proposed that any computer-application ontology should have the high-level characteristics as follows [Gruber93a]:

a) It is an explicit specification of a conceptualization that caters to: i) representation of the consensus-certified domain knowledge (e.g. human disease [Wongthongtham06b] and allopathic medicine [UMLS]; ii) the ontology should have a logical representation or machine-understandable form (i.e. the semantic net that is alternatively known as

the DOM (document object model)), and iii) the corresponding human-understandable form (i.e. the syntactical layer for implicit or explicit query formations). Therefore, a system with ontology support should have three layers: ontology at the bottom, a semantic net (and the corresponding parsing mechanism or parser) in the middle, and a query system at the top. The parser provides the logical conclusion for the query according to its actual parameters specified by the user by working at the DOM tree by inference.

b) The knowledge in the form of descriptive philosophy must be translated correctly into the application knowledge/data.

The second characteristic in the Gruber's proposal is not easy to achieve as it involves the issues of appropriate and flexible data structures as well as the issue of supporting interoperability. Today, the computer is basically the web or Internet even within a commercial enterprise. Sharing the same ontological knowledge/data based over the web may involve machines, communication technologies, and software from various vendors [Hamilton02].

To confine the meaning of Gruber's second characteristics in the field of applications, Guarino proposed that the ontological contents should be organized into a subsumption hierarchy [Guarino00] and the interpretation of the associations among sub-ontologies should be axiomatically constrained. This proposal had alleviated the ontology phobia of many potential applicants [Taniar06]. Since then, building enterprise ontology to house the necessary knowledge of an enterprise or corporation has started to flourish [Clark05,

Coplien04, Dunn05]. There was success in adopting the enterprise concept for managing software development, which involves different teams across the globe (i.e. *multi-site software development*). If the ontology is the enterprise vocabulary with constrained interpretation that all the team involved must use as a base, then there is no chance for multi-representation of the same terms; modules can be integrated perfectly [JWong08b].

Meanwhile different groups have been actively researching how to support the translation process (Gruber's 2nd characteristic) and it has been generally concluded that the use of metadata (e.g. XML, RDF, and OWL) is viable [Lopez99, IBM03]. Other researchers have been busy proposing ontology building and managing tools [Denny04], but unfortunately these tools are exclusively inoperable. One conclusion, however, is seemingly unanimous – we can use semantic web technology to integrate enterprises – and the W3C has contributed tremendously in this aspect. Indirectly the accumulated XML, RDF, and OWL metadata experience in the field has benefited this research direction enormously.

3.3 Ontology and Semantic Web

The aim of classical ontology is to weave/intertwine philosophical knowledge that has been agreed by consensus certification – a knowledge engineering approach. Normally, the ontology reflects knowledge up to a particular time t (i.e. $Onto_t$). For example, the philosophical foundations, which were laid down by philosophers such as Aristotle (Greek philosopher) and Lao

Zi (Chinese philosopher) have since been expanded as time progresses. It is, however, difficult to realize all the philosophical principles in full. Instead, implementations have to be adapted with respect to the environmental forces, needs, and temporal factors. In the area of computer science, ontology is considered as a viable approach for organizing usable knowledge for scientific and engineering applications. Perhaps, Gruber [Gruber93a] and Guarino [Guarino00] together have laid down a good theoretical basis of how ontology can be used to disambiguate a "community knowledge base", which should be created by domain experts in a consensus certification process. In fact, the ultimate aim of using ontology is to disambiguate information interchange by constraining terminology with respect to the standard community vocabulary/lexicon. If the "community knowledge base" is for usage within an enterprise, it would be the enterprise knowledge base or enterprise ontology. The constraining process disallows certain meanings of the same word (i.e. multi-representations). To put this into perspective, the meanings of terms are filtered so that only those relevant meanings are allowed to pass through; these filters are the domain/application views. A successful example is enterprise ontology for "distributed software engineering", which may involve different groups in different regions/countries across the globe (e.g. Verizon). The enterprise ontology supported by predefined views filters out regional multi-representations for the same term, making software modules developed anywhere, anytime "integrate-able" with minimal errors. A key ontology feature is that, through formal, real-world semantics and consensual terminologies, it interweaves human and machine understanding [Fensel03]. In

this way, the ontology facilitates sharing and reuse of knowledge for both humans and machines.

The web contains worldly but rather unorganized knowledge in an intertwined fashion, and the meaning and importance of a piece of web information depends on subjective interpretations of the potential users. To benefit a community, subjective interpretations should depend on collective agreements – the prelude to consensus certification. The first logical step to reap the potential usefulness of web sources is to organize them into a standard (i.e. community/domain) repertoire. The organization involves: view definitions of terms and attributes, concepts and their associations, classification or categorization of these view definitions (sub-ontologies), and define the subsumption hierarchy for these sub-ontologies. The information/knowledge organization finally agreed by consensus becomes the "community, domain or enterprise" ontology, which is then made into the machine-usable form. In this light, the ontology approach makes the web semantics potentially usable; from the user's point of view, the web is now semantic (i.e. meanings can be easily extracted and used – semantic web). If one could neutralize the subsumption hierarchy in the web ontology into a Petri net, then every network path (in the Petri net) would have a meaning (semantic). Then, tracing such a path for a set of parameters is a logical process or inference. The traced path is the logical conclusion for the given set of parameters; its connotation can be predefined in the specific context via a consensus certification process.

The UMLS [UMLS] is a real-life example of how a specific allopathic medical ontology is organized into three layers: i) the allopathic context, from a specific angle, into the bottom-layer ontology; ii) the middle-layer semantic net to logically represent the bottom ontology for machine processing; it is the machine-process-able form and the processor is the logical parser; and iii) the top-layer syntactical representation of the semantic net so that human users can formulate the corresponding queries. In fact, all three layers connote the same knowledge in different forms for different purposes. The knowledge, however, must be represented formally and in a retrievable form for practical use. The representation or retrieval form can be a meta-data system (e.g. XML). For production systems the meta-system may be converted into technology-oriented databases (e.g. Microsoft (MS) relational database – SQL server) so that tools from different vendors are readily applicable. In this way the chance of premature obsolescence is prevented because the databases are tied in with the related proprietary artefacts of data interoperability. This also applies to the metadata systems; for example, XML and its successors (RDF and OWL) have less chance of obsolescence because they are supported by the influential W3C.

As the semantic web keeps developing [Fensel03], its knowledge contents are becoming richer and more useful. As a result it should be managed to the benefit of mankind on a global scale. Tim Berners-Lee actually envisioned the semantic web as the natural evolution from the current web. Its evolution involves additions of machine-readable information and automated services. Berners-Lee thought that explicit representations of the semantics

underlying the web data, programs, pages, and other web resources would support qualitative new services [Fensel03].

3.3.1 RDF

RDF (Resource Description Framework) was proposed by W3C as a more formal format for making assertions and leveraging the XML format to represent and transport information. It is a language specifically designed for representing resources and information in the World Wide Web. It is a metadata system extended from the XML, and it clearly defines the details of web resources; for example, title, author, web page modification date, copyright, licensing information of a web document, and/or the availability schedule for sharing resources.

By providing a common framework/infrastructure/standard the RDF facilitates encoding, exchanging and reusing structured data in the metadata system. Information exchanged via the RDF would have intact meaning. On top of the RDF platform there are many commercial support tools by various vendors. But, the RDF standard makes the data handled by different tools interoperable. The obvious RDF advantage for the application designers is the availability of common RDF parsers and processing tools; simply select and use them. The tools usually recognize things with web identifiers (known as the Uniform Resource Identifiers, or URIs) that describe resources in terms of their simple properties and values.

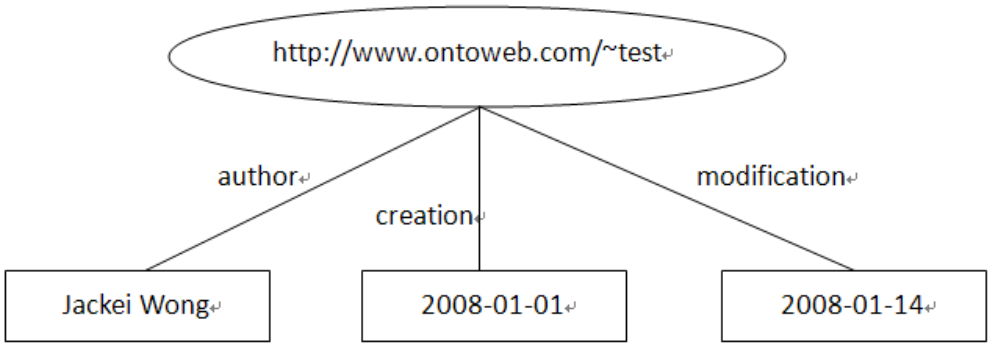


Figure 3.3.1.1 URL hierarchy

Figure 3.3.1.1 shows how a URL represents a hierarchy of information:

i) Resource: http://www.ontoweb.com/~test

ii) Property: The element <author>

iii) Value: The string "Jackei Wong"

A statement may also be described in terms of (Subject, Predicate, Object). Referring to the same example:

i) Subject: http://www.ontoweb.com/~test

ii) Property: The element <author>

iii) Value: The string "Jackei Wong"

RDF is a generally syntax-independent model for representing resources and their corresponding descriptions, as shown in Figure 3.3.1.2. (The model can be expressed in XML, and the specification uses XML as its syntax for encoding metadata). RDF provides an enhanced representation over XML in defining the relationship such as the concept of class and subclass, and also the

triple (Resource, Property, Value) or (Subject, Predicate, Object). RDF is extensible, which means that descriptions can be enriched with additional descriptive information, as shown in Figure 3.3.1.3.

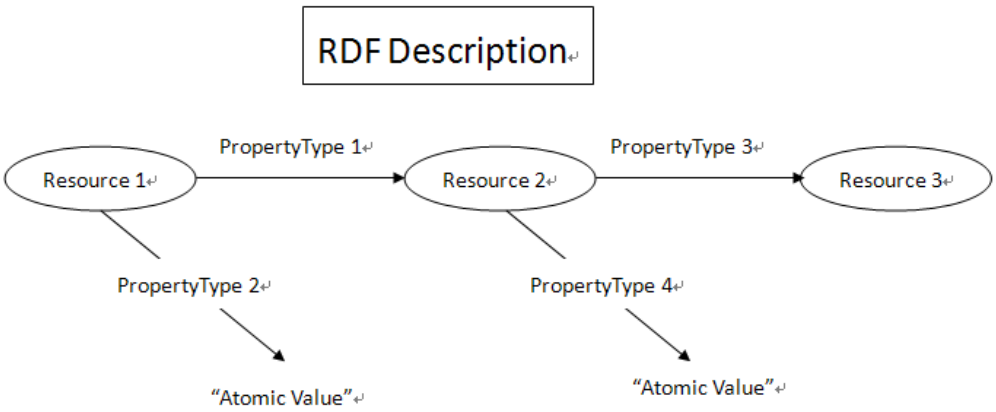


Figure 3.3.1.2 RDF description of resources

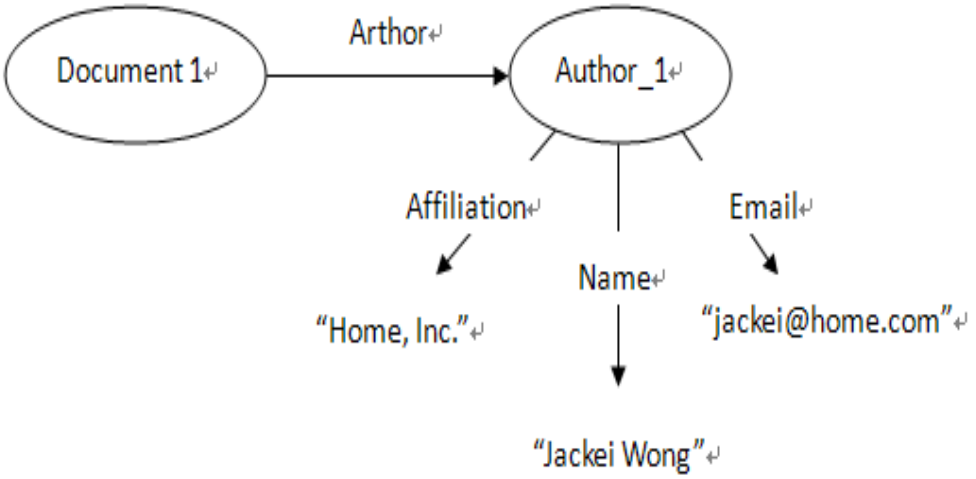


Figure 3.3.1.3 Extensible RDF representation

Although RDF is generally syntax-independent, it provides an XML syntax called serialization syntax. The following is an example:

```
<?xml version”1.0”?>
<RDF>
        <Description about=http://www.ontoweb.com/~test>
                <author>Jackei Wong</author>
                <created>2008-01-01</created>
                <modified>2008-01-14</modified>
        </Description>
</RDF>
```

RDF metadata can be inserted or nested into XML code or vice versa. Yet, for information operability RDF metadata formats have to be pre-defined in order to facilitate correct information exchange over the web anytime, anywhere and with anyone that operates with the same “format core”. The Dublin Core is an example in which a set of properties for describing documents are predefined. The first set of Dublin Core property was defined at the Metadata Workshop in Dublin, Ohio 1995. It is currently maintained by the Dublin Core Metadata Initiative. Dublin Core is a standard for cross-domain information resource description. It provides a simple and standardized set of conventions for describing things online in ways that make them easier to find. It is widely used to describe materials such as video, sound, image, text, as well as those composite media like web pages. Implementations with support from the Dublin Core typically work with both XML and RDF. Simple Dublin Core consists of the following 15 metadata elements [W3Schools] (Table 3.3.1.1).

Elements	Definition
Contributor	An entity responsible for making contributions to the content of the resource
Coverage	The extent or scope of the content of the resource
Creator	An entity primarily responsible for making the content of the resource
Format	The physical of digital manifestation of the resource
Date	A date of an event in the lifecycle of the resource
Description	An account of the content of the resource
Identifier	An unambiguous reference to the resource within a given context
Language	A language of the intellectual content of the resource
Publisher	An entity responsible for making the resource available
Relation	A reference to a related resource
Rights	Information about rights held in and over the resource
Source	A reference to a resource from which the present resource is derived
Subject	A topic of the content of the resource
Title	A name given to the resource
Type	The nature of genre of the content of the resource

Table 3.3.1.1 Original 15 elements in Simple Dublin Core

Conceptually the elements in Simple Dublin Core could be construed as predefined "data types". In contrast the XML has only one data type "string".

After the original 15 elements specification, refinement of the Dublin Core Metadata Element Set (DCMES) continued. Additional terms were identified by working groups in the Dublin Core Metadata Initiative (DCMI) and judged by the DCMI Usage Board to be in conformance with principles of good practice for the qualification of Dublin Core metadata elements.

RDF element refinements narrow its meaning, and a refined element shares the meaning of the unqualified element but with a more restricted scope. In additional to element refinements, Qualified Dublin Core includes a set of recommended encoding schemes, designed to aid the interpretation of an

element value. The schemes include controlled terms and formal notations or parsing rules. A value expressed using an encoding scheme may thus be a token selected from a controlled vocabulary, or a string formatted in accordance with a formal notation. For example, “2008-03-10” can be used as the standard expression of a date. DCMI also maintains a small, general vocabulary recommended for use within the element Type. This vocabulary currently consists of 12 terms. In effect, the refinement is a standardization process.

An example of RDF and XML namespace is shown below. An XML namespace is used to unambiguously identify the schema for the Dublin Core terms by pointing to the definitive Dublin Core resource that defines the corresponding semantics. In this example, RDF is nested inside XML – the two are interoperable.

```
<?xml version=”1.0”?>
<rdf:RDF
        xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
        xmlns:dc=http://purl.org/dc/elements/1.1>
                <rdf:Description rdf:about=http://uri-of-Document-1>
                        <dc:creator>Jackei Wong</dc:creator>
                </rdf:Description>
</rdf:RDF>
```

An example using description element is shown below:

```
<?xml version=”1.0”?>
<rdf:RDF
        xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
        xmlns:cd=http://www.recshop.fake/cd>
                <rdf:Description
                        rdf:about=http://www.recshop.fake/cd/Empire_Burlesque>
                        <cd:artist>Bob Dylan</cd:artist>
                        <cd:country>USA</cd:country>
                        <cd:company>Columbia</cd:company>
                </rdf:Description>
</rdf:RDF>
```

RDF is a formalism in light of metadata annotation; a way to work with XML harmoniously. But, it does not provide the special meanings to some terms such as subClassOf of Type. The definitions of these terms rely on the RDF Schema (RDFS), which allows users to define the terms and their associations or relationships. The definitions provide “extra meanings” to particular RDF predicates and resources. The “extra meaning” or semantics specifies how a term should be interpreted. In RDFS, terms include Class, Property, type, subClassOf, range, domain etc. Some examples are as follow:

i) <Person, type, Class>

ii) The type of “Person” is a “Class”

iii) <Professor, subClassOf, Person>

iv) “Professor” is a subclass of the class “Person”

v) <Allan, type, Professor>

vi) The type of “Allan” is a “Professor”

vii) <hasColleague, domain, Person>

viii) The domain of “hasColleague” is “Person”

The RDF, however, has some problems as follows:

i) No distinction between classes and instances (individuals)

ii) No localized range and domain constraints

iii) No existence/cardinality constraints

iv) No transitive, inverse or symmetrical properties

v) Too weak to describe resources in sufficient detail

vi) Difficult to provide reasoning support

Meanwhile, powerful web ontology language is expected to have some basic qualities:

i) It extends existing Web standards (E.g. XML, RDF, RDFS)

ii) It is easy to use and understand

iii) It is formal

iv) It has adequate expressive power

v) It can provide automated reasoning support

The few web ontology languages developed after RDF include:

i) OIL – European researchers

ii) DAML-ONT – US researchers in DARPA DAML programme

iii) DAML+OIL – Joint EU/US Committee on Agent Markup Languages and extended RDF

iv) OWL – W3C – Web-ontology working Group and based on DAML+OIL

3.3.2 OWL

From the literature, we see that annotations of web semantics have evolved from XML, through RDF to OWL (Web Ontology Language). These three metadata systems are interoperable and nested in one another. OWL basically compensates the RDF shortcomings. It is a family of knowledge representation languages for authoring ontology constructs, and it is endorsed by the World Wide Web Consortium (W3C). It is a family of languages based on two (largely, but not entirely, compatible) systems: OWL DL and OWL Lite. Again these two systems are based on Description Logics [Horrocks04], which have attractive and well-understood computational properties. The OWL Full uses a novel semantic model, intended to have compatibility with RDF Schema. OWL ontology constructs are usually serialized using RDF/XML syntax.

The W3C-endorsed OWL specification includes the definition of three variants of OWL, with different levels of expressiveness:

i) OWL Lite – This is a subset of OWL DL and supports classification hierarchy and simple constraints, for example it only permits cardinality values of 0 or 1

ii) OWL DL – This supports maximum expressiveness, computational completeness, decidability; it includes all language constructs; its functions correspond with Description Logics (a fragment of FOL (First Order Logic))

iii) OWL Full – This supports maximum expressiveness, syntactic freedom of RDF, no computational guarantees; it is the union of OWL syntax and RDF; its reasoning software cannot support complete logical reasoning

Every one of the following sublanguages is a syntactic extension of their simpler predecessors. But, it should be noted that relationship may not be transitive; for example, the following set of relations hold but their inverses do not:

i) Every legal OWL Lite ontology is a legal OWL DL ontology

ii) Every legal OWL DL ontology is a legal OWL Full ontology

iii) Every valid OWL Lite conclusion is a valid OWL DL conclusion

iv) Every valid OWL DL conclusion is a valid OWL Full conclusion

In the sequel, a few main OWL characteristics will be shown with examples:

OWL – Namespace

<rdf:RDF

xmlns=http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#

xmlns:vin=http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#

xmlns:food=http://www.w3.org/TR/2004/REC-owl-guide-20040210/food#

xmlns:owl=http://www.w3.org/2002/07/owl#

xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#

xmlns:rdfs=http://www.w3.org/2000/01/rdf-schema#

xmlns:xsd=http://www.w3.org/2001/XMLSchema#>

OWL – Headers

<owl:Ontology rdf:about="">

- Ontology name

<rdfs:comment>An example OWL ontology</rdfs:comment>

<owl:priorVersion rdf:resource="http://www.w3.org/TR/2003/PR-owl-guide-

20031215/wine"/>

- Ontology versioning

<owl:imports rdf:resource="http://www.w3.org/TR/2004/REC-owl-guide-20040210/food"/>

- Import other ontologies

<rdfs:label>Wine Ontology</rdfs:label>

- Natural language label

OWL – Classes

<owl:Class rdf:ID="PotableLiquid"/>

<owl:Class rdf:ID="http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#Region" />

<owl:Class rdf:ID="Wine">

<rdfs:subClassOf rdf:resource="&food;PotableLiquid"/>

<rdfs:label xml:lang="en">wine</rdfs:label>

<rdfs:label xml:lang="fr">vin</rdfs:label>

```
    ...
</owl:Class>
```

OWL – Individuals

```
<Region rdf:ID="CentralCoastRegion" />

<owl:Thing rdf:ID="CentralCoastRegion" />
<owl:Thing rdf:about="#CentralCoastRegion">
    <rdf:type rdf:resource="#Region"/>
</owl:Thing>
```

- "CentralCoastRegion" is a ***member*** of "Region"

OWL - Properties

```
<owl:Class rdf:ID="WineDescriptor" />
<owl:Class rdf:ID="WineColor">
    <rdfs:subClassOf rdf:resource="#WineDescriptor" />
    ...
</owl:Class>

<owl:ObjectPropertyrdf:ID="hasWineDescriptor">
    <rdfs:domain rdf:resource="#Wine" />
    <rdfs:range rdf:resource="#WineDescriptor" />
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="hasColor">
    <rdfs:subPropertyOf rdf:resource="#hasWineDescriptor" />
    <rdfs:range rdf:resource="#WineColor" />
    ...
```

```
</owl:ObjectProperty>

<owl:Class rdf:ID="Vintage">
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#vintageOf"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">
            1</owl:minCardinality>
      </owl:Restriction>
   </rdfs:subClassOf>
</owl:Class>

<owl:ObjectProperty rdf:ID="vintageOf">
   <rdfs:domain rdf:resource="#Vintage" />
   <rdfs:range rdf:resource="#Wine" />
</owl:ObjectProperty>
```

OWL – Datatype Properties

```
<owl:Class rdf:ID="VintageYear" />
<owl:DatatypeProperty rdf:ID="yearValue">
   <rdfs:domain rdf:resource="#VintageYear" />
   <rdfs:range rdf:resource="&xsd;positiveInteger"/>
</owl:DatatypeProperty>

<VintageYear rdf:ID="Year1998">
   <yearValue rdf:datatype="&xsd;positiveInteger">1998
      </yearValue>
</VintageYear>
```

OWL – Transitive Property

- P(x,y) and P(y,z) implies P(x,z)

```
<owl:ObjectProperty rdf:ID="locatedIn">
   <rdf:type rdf:resource="&owl;TransitiveProperty" />
   <rdfs:domain rdf:resource="&owl;Thing" />
   <rdfs:range rdf:resource="#Region" />
</owl:ObjectProperty>

<Region rdf:ID="SantaCruzMountainsRegion">
   <locatedIn rdf:resource="#CaliforniaRegion" />
</Region>

<Region rdf:ID="CaliforniaRegion">
   <locatedIn rdf:resource="#USRegion" />
</Region>
```

OWL – Inverse Property

- $P_1(x,y) \rightarrow P_2(y,x)$

```
<owl:ObjectProperty rdf:ID="IsTaughtBy">
   <rdf:type rdf:resource="&owl;InverseProperty" />
   <rdfs:domain rdf:resource="#Teacher" />
   <rdfs:range rdf:resource="#Teacher" />
</owl:ObjectProperty>
```

OWL – Symmetric Property

- P(x,y) iff P(y,x)

```
<owl:ObjectProperty rdf:ID="adjacentRegion">
   <rdf:type rdf:resource="&owl;SymmetricProperty" />
   <rdfs:domain rdf:resource="#Region" />
   <rdfs:range rdf:resource="#Region" />
</owl:ObjectProperty>

<Region rdf:ID="MendocinoRegion">
   <locatedIn rdf:resource="#CaliforniaRegion" />
   <adjacentRegion rdf:resource="#SonomaRegion" />
</Region>
```

OWL – Functional Property

- P(x,y) and P(x,z) implies y=z

```
<owl:Class rdf:ID="VintageYear" />
<owl:ObjectProperty rdf:ID="hasVintageYear">
   <rdf:type rdf:resource="&owl;FunctionalProperty" />
   <rdfs:domain rdf:resource="#Vintage" />
   <rdfs:range rdf:resource="#VintageYear" />
</owl:ObjectProperty>
```

OWL – Property Restriction

```
<owl:Class rdf:ID="Wine">
   <rdfs:subClassOf rdf:resource="&food;PotableLiquid" />
   ...
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#hasMaker" />
         <owl:allValuesFrom rdf:resource="#Winery" />
```

```
      </owl:Restriction>
   </rdfs:subClassOf>
   ...
</owl:Class>
```

- allValuesFrom: For all wines, if they have makers, all the makers are wineries
- someValuesFrom: For all wines, they have at least one maker that is a winery

```
<owl:Class rdf:ID="Vintage">
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#hasVintageYear"/>
         <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1
            </owl:cardinality>
      </owl:Restriction>
   </rdfs:subClassOf>
</owl:Class>
```

- Every Vintage has exactly one VintageYear

OWL – Ontology Mapping

- To show a particular class or property in one ontology is ***equivalent*** to a class or property in a second ontology

```
<owl:Class rdf:ID="Wine">
   <owl:equivalentClass rdf:resource="&vin;Wine"/>
</owl:Class>

<owl:Class rdf:ID="TexasThings">
   <owl:equivalentClass>
```

```
      <owl:Restriction>
        <owl:onProperty rdf:resource="#locatedIn" />
        <owl:someValuesFrom rdf:resource="#TexasRegion" />
      </owl:Restriction>
    </owl:equivalentClass>
</owl:Class>

<Wine rdf:ID="MikesFavoriteWine">
    <owl:sameAs rdf:resource="#StGenevieveTexasWhite" />
</Wine>
```

OWL – Complex Classes

- OWL supports the basic set operations, namely ***union***, ***intersection*** and ***complement***

```
<owl:Class rdf:ID="WhiteWine">
    <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#Wine" />
      <owl:Restriction>
        <owl:onProperty rdf:resource="#hasColor" />
        <owl:hasValue rdf:resource="#White" />
      </owl:Restriction>
    </owl:intersectionOf>
</owl:Class>

<owl:Class rdf:ID="Fruit">
    <owl:unionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#SweetFruit" />
      <owl:Class rdf:about="#NonSweetFruit" />
    </owl:unionOf>
</owl:Class>
```

From the examples above, it is proven that OWL is more powerful than RDF(S).

3.4 Concise Comparison of XML/RDF/OWL

	Data Types		Types of Properties		Property Element		Classes	
	Primitive data-type	Numeric min, max	Transitive	Inverse	Import element	Individual element	Negation/ Disjoint classes	Inheritance
RDF(S)	No	No	No	No	No	Yes	No	Yes
DAML+ OIL	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
OWL	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 3.4.1 Comparing RDF, DAML+OIL and OWL

Table 3.4.1 compares RDF, DAML+OIL and OWL. The focus of this chapter, however, is on XML, RDF and OWL and thus, DAML+OIL will not be covered here. XML is the 1st-level metadata that contains no meaning for machines but does have meaning to humans. RDF is a metadata system extended from the XML, and it can clearly define the details of web resources (e.g. title, author, web page modification date, copyright, licensing information of a web document, and/or the availability schedule for sharing resources). We can compare RDF and OWL qualitatively in three areas:

i) *Data types* – RDF(S) does not contain primitive data-type, nor the numeric minimum or maximum functions, but OWL does. *Types of*

relationship – No transitive and inverse relationships can be found when using RDF(S) but OWL can have them.

ii) *Property elements* – Only OWL can import elements into it while RDF(S) cannot. Both of them can set the property of elements individually.

iii) *Classes* – OWL contains the classes of negation and disjoint while RDF(S) does not. Both of them have the property of class inheritance.

3.5 Transformation of Markup Languages

A markup language (ML) uses a set of annotations to describe the structure of a text, and layout its format. ML of different forms has been used for decades; for example, computer typesetting and word-processing systems use markup languages.

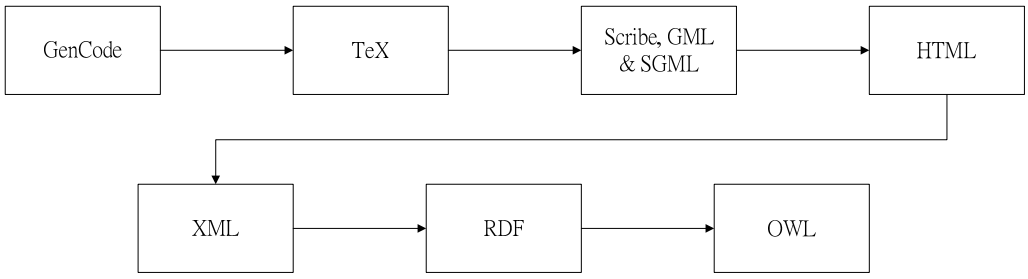


Figure 3.5.1 Transformation of markup languages

The diagram in Figure 3.5.1 shows the transformation of markup languages. The idea of "markup languages" was first presented by publishing executive William W. Tunnicliffe at a conference in 1967; it was called "generic coding" at the time. He then led the development of a standard called "GenCode" for the publication industry. In later years, Donald Knuth created

another major publishing standard – TeX, and kept continuously refining it in 1970s and 1980s. The focus of TeX is on the detailed layout of text and the font description in order to typeset mathematical books in professional quality. TeX requires considerable skill from the users, so that it is mainly used in academic area. It is the *de facto* standard in many scientific disciplines. LaTeX, a TeX macro package, provides a descriptive markup system on top of Tex and is widely used in academic and technical publications (e.g. Springer Verlag).

Scribe, developed by Brian Reid in 1980, is the first language to make a clean and clear distinction between presentation and structure. It was revolutionary at that time because it introduced the idea of styles to be separated from the marked up document, as well as using grammar to control the usage of descriptive elements. It also influenced the development of Generalized Markup Language (GML), and later Standard Generalized Markup Language (SGML), and is a direct ancestor of HTML and LaTeX.

In the early 1980s, SGML was created for the idea that the markup operation should be focused on the structural aspects of a document, with the visual presentation of that structure left to the interpreter. SGML was developed by a committee chaired by Goldfarb. It incorporated ideas from many different sources including Tunnicliffe's project, GenCode. SGML specified a syntax which includes the markup in documents, and it separately describes what and where the tags are allowed (Document Type Definition (DTD) or schema). This syntax allows authors to create and use any markup they want, or select tags

that make the most sense in their own languages. So, SGML is recognized as a proper meta-language from which other markup languages are derived.

Tim Berners-Lee learned SGML and used its syntax to create HyperText Markup Language (HTML), which is similar to other SGML-based tag languages, but much simpler. However, some computer scientists disputed that HTML was hard to use because it restricts the tag placement. These scientists argued that easy-to-use markup languages should be hierarchical instead of being just a **"***language of container***"**. The hierarchical argument led to the emergence of other languages such as XML and its interoperable partners including RDF and OWL. The contemporary markup languages follow the "what you see is what you get" style.

3.6 First Recap

In philosophy, ontology is a problem of essence (nature) and existence. In computer science, the word "ontology", borrowed from philosophy, represents a set of precisely defined terms (vocabularies), which are accepted by a domain or community. Gruber argued that ontology should be an explicit specification of a conceptualization [Gruber93a], which represents concepts and the relationships among them. Guarino later proposed that the ontological contents should be organized into a subsumption hierarchy [Guarino00] for clarity.

An investigation on the metadata languages, XML, RDF and OWL had been done. XML is the first level of metadata that contains no meaning for machines but has meaning for humans. RDF is more powerful than XML as it can represent relationships among the data entities. OWL is the most powerful as it covers useful functions for representing ontology cases. Different groups have been actively researching how to support the translation process (Gruber's 2nd characteristic) and it has been generally concluded that the use of metadata (e.g. XML, RDF, and OWL) is viable [Lopez99, IBM03]. A conclusion is that we can use the semantic web technology to integrate enterprises. In this respect, the W3C has contributed tremendously in the past two decades.

3.7 Metadata System Manipulation

As it was indicated in the first recap, my preliminary research had found that metadata systems are quite suitable for expressing and accentuating TCM knowledge in terms of entity relationships. This had inspired the design of some experiments to deepen the investigation.

3.7.1 Study of JENA

Jena is a Java framework for building Semantic Web applications [JENA]. It provides a programmatic environment for RDF, RDFS, OWL and SPARQL, and includes a rule-based inference engine. It is open source software from HP Labs Semantic Web Programme. The Jena framework includes:

i) A RDF API

ii) Reading and writing RDF in RDF/XML, N3 and N-Triples

iii) An OWL API

iv) In-memory and persistent storage

v) SPARQL query engine

3.7.2 JENA and RDF

The “RDFNode” interface provides a common base for all elements that can be parts of the RDF triples. The “Literal” interface represents literals such as “Blue Car” that can be used as the <Object> in {Predicate, Subject, Object} triples. The “Literal” interface provides accessor methods to convert literals to various Java types such as String, integer and double. Objects implementing the “Property” interface can be the <Predicate> in {Predicate, Subject, Object} triples. The “Statement” interface represents a {Predicate, Subject, Object} triples. It can also be used as the <Object> in a triple since RDF allows statements to be nested. Objects Implementing the “Container”, “Alt”, “Bag” or “Seq” interface can be the <Object> in {Predicate, Subject, Object} triples.

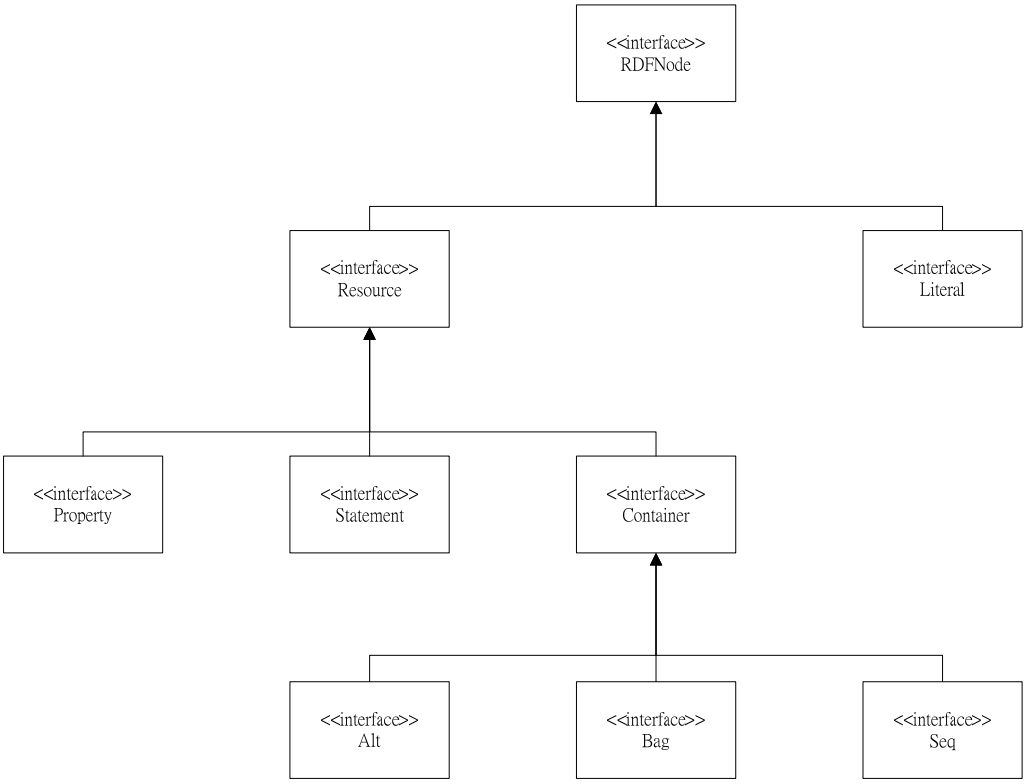


Figure 3.7.2.1 Interfaces for accessing/manipulating RDF statements in Jena

3.8 Verification of the RDF and OWL Suitability

Many experiments were conducted to verify the suitability of RDF and OWL for annotating a clinical TCM onto-core with the Nong's enterprise TCM onto-core as the basis. The suitability conclusion is drawn on the fact that the metadata system can indeed yield the necessary *cross-layer semantic transitivity* in a D/P system of three layers: the query system in the top layer to abstract the semantic net in the middle layer for human understanding and manipulation; the semantic net in the middle layer to fully represent the subsumption hierarchy of the onto-core in the bottom layer for machine understanding and processing; the ontology in the bottom layer to store the necessary knowledge for clinical operation. With cross-layer semantic

transitivity support, for an entity picked from any layer the corresponding representations in the other two layers would surface consistently. All the experimental results indicate unanimously that the, RDF and OWL are both suitable for annotating the TCM ontology constructs, except that OWL is more powerful than RDF because it supports both transitive logical interpretation and tracing. Some selected experimental results are presented in the "**8.2.1 RDF and OWL Verifications**" section for demonstration.

3.9 Second Recap

The W3C (World Wide Web Consortium) first proposed the XML metadata system, which requires the user to declare the relationship among entities and concepts. This kind of declaration provides the following: i) *disadvantage* – if the recipient is not given the meanings of the declarations, then it is difficult to achieve interoperability; and ii) *advantage* – the disadvantage above becomes the advantage as far as data security is concerned. The RDF is more powerful than the XML in the sense that logical relationships is intrinsic in the metadata but only in a *"forward only fashion" – not transitive*. The OWL overcomes the RDF's intransitive shortcoming. In fact, in the three metadata systems, XML, RDF and OWL, annotations can be nested in a single document. Some experiments were conducted with the aim of verifying the annotation power of the RDF and OWL with the help of the STV (Semantic Transitivity Visualizer), which is another **novel** contribution of this thesis. The results from the different experiments indicate that both the RDF and OWL are suitable for annotating the Nong's enterprise TCM onto-core. OWL is, however,

potentially more powerful than RDF because it indicates the relationships among different elements so that user does not need to declare the relationships before data preparation. In this light, OWL can support more complex query systems because the logical relationships among entities can be more clearly and richly declared. The XML, however, has the advantage of simplicity over the RDF and OWL. The security embedded in the XML annotations is somewhat intrinsic because the recipient of the annotations may not understand the embedded logic unless prior explanation was received.

It is shown in my experiments that XML, RDF and OWL can be nested at will in a single document. This finding is useful because the three metadata systems may be nested for optimal document constructions (according to the concensus certification of the domain experts).

3.10 Conclusion and Connective Statement

Many experiments were conducted to verify the suitability of RDF and OWL for annotating TCM ontology. The conclusion is that they are indeed suitable but OWL is more powerful than RDF because it supports transitive logical interpretation. In fact, XML, RDF and OWL can be nested at will, and this is useful for constructing knowledge base to suit the degree of optimality in mind. The next logical step is to lucidly explain the following in Chapter 4: ontology, enterprise ontology, semantic transitivity, and knowledge discovery by data mining (from open sources) in the context of this thesis.

3.11 Key References

[ALA99] American Library Association, Task Force on Metadata Summary Report, June 1999

[Berners01] T. Berners-Lee, H. James and L. Ora, The Semantic Web, Scientific American Magazine, 17 May 2001

[Bray04] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler and F. Yergeua, Extensible Markup Language (XML) 1.0 (4th Edition) – Origin and Goals. World Wide Web Consortium, 2004

[Bultermann04] D.C.A. Bultermann, Is It Time for a Moratorium on Metadata? IEEE MultiMedia, October-December 2004

[Clark05] D. Clark, Position Paper for Rules Concerning Project Management Ontologies, 2005

[Coplien04] J. Coplien, Organizational Patterns: Beyond Technology to People, Proc. of the 6th International Conference on Enterprise Information Systems (ICES2004), Porto, Portugal, 2004

[Coppin04] B. Coppin, Artificial Intelligence Illuminated, Jones & Bartlett Publishers, 2004

[Corazzon04] R. Corazzon, Descriptive and Formal Ontology – A Resource Guide to Contemporary Research, 2004, http://www.formalontology.it/

[Denny04] M. Denny, Ontology Tools Survey, 2004, http://ww.xml.com/lpt/a/2004/07/14/onto.html

[Dunn05] C. Dunn and J. Hollander, The REA Enterprise Ontology: Value System and Value Chain Modelling, Enterprise Information Systems: A Pattern-based Approach, McGraw Hill, 2005

[Fensel03] D. Fensel, Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, 2nd Edition, Springer, Berlin/Heidelberg, 2003

[Gruber93a] T.R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, No. 2, 1993, 199-220

[Guarino00] N. Guarino and C. Welty, Towards a Methodology for Ontology Based Model Engineering, 2000, http://citseer/ist/psu/edu/b12206.htm

[Hamilton02] J.A. Hamilton, J. Rosen, and P.A. Summers, Developing Interoperability Metrics, Auburn University, 2002, http://www.eng.auburn.edu/users/hamiliton/security/spawar/6_Deveoping_Interoperability_Metrics/pdf

[Herman08] I. Herman, Semantic Web Activity Statement, W3C, 7 March 2008

[Horrocks04] I. Horrocks and P.F. Patel-Schneider, Reducing OWL Entailment to Description Logic Satisfiability, Web Semantics, Vol. 1, No. 4, 2004

[IBM03] IBM and Sandpiper Software Incorporated, Ontology Definition Metamodel – Third Revised Submission to

OMG/RFP ad, 4 March 2003, http://www.omg.org/docs/ad/05-8-01.pdf

[JENA] http://jena.sourceforge.net/

[JWong08b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008

[Lopez99] F. Lopez, Overview of Methodologies for Building Ontologies, 1999, http://www.ontology.org/maim/presentations/madrid/analysis.pdf

[Taniar06] D. Taniar and J.W. Rahayu, Web Semantics & Ontology, Idea Group Publishing, 2006

[UMLS] http://umls.nlm.nih.gov/

[W3Cc] W3C Web site, http://www.w3.org/

[W3Schools] W3 Schools, http://www.w3schools.com/

[Wongthongtham06b] P. Wongthongtham, Ontology and Multi-agent-based Systems for Human Disease Studies, PhD Thesis, Curtin University, 2006

[Yergeua04] F. Yergeua, T. Bra, J. Paoli, S. Sperberg-McQueen and E. Maler, Extensible Markup Language (XML) 1.0 (3rd ed.), W3C Recommendation, 2004

Chapter 4 Ontology, Enterprise Ontology, Semantic Transitivity and Knowledge Discovery by Text Mining

4.1 Introduction

Knowledge appears (or is annotated) in different forms throughout human history, including paintings, murals, texts, chanting and engraving. Usually similar knowledge items may have different meanings or semantics in different communities. The interpretation of the same knowledge item may vary among different communities because they use different rules and grammar. The method of interpretation for a small community may be guarded by a few selected individuals such as tribal shamans. The overwhelming amount of knowledge in big communities cannot rely on a few individuals for interpretations that benefit society. Instead more systematic knowledge organization, interpretation methods, and verification approaches are needed so that useful knowledge is disseminated and passed on, and the public is effectively educated.

It is normal for a knowledge embodiment to consist of both facts (true) and philosophically valid items. In philosophy, this means one reasons syllogistically and accepts something as valid based on the axioms laid down. Valid arguments may not be the truth except that they are logically correct according to the axioms. These kinds of arguments have been rife in different periods of history with axioms agreed by philosophers at the time. That is, the argument is consensus-certified by a sufficient number of extant experts of the

particular community. For example, the great ancient philosophers like Plato and Aristotle in the West, and Lao Zi [LaoZi] in China, dealt with syllogism rather than pure truth. Sometimes, truth (TR) exists only at a defined period of time *t*; it is function of time mathematically: $TR(t) = f(t)$. For example, the so called North Pole today was once the South Pole. Therefore, it is truer to say that earth has magnetic polarity than the North Pole has always been geographically where it is today.

If a knowledge domain addresses a school of thought (e.g. "Naturalism" by Lao Zi), then the domain of concepts, entities and their associations form the ontology. At the start, the school of thought was based on a few logical axioms, and as time passes more axioms may be added to enrich the domain. Intrinsically reasoning with any ontology is syllogistic inference [Bassler98]. Therefore, any ontology is living and incrementally richer provided that it has followers, who every now and then agree on new axioms. Special characteristics of ontology can be summarized as follows:

1) It is axiomatic, syllogistic and living conceptualization.
2) Its knowledge items have explicit semantics.
3) Reasoning or inference is based on predicates (i.e. If…Then….Else…) and bounded by the pre-defined vocabulary or lexicon.

4.2 Ontology and Computing

The advent of the Internet has spawned many new, different opportunities in respect to research, business, communication and living comfort. The overwhelming amount of information accumulated in the web is a potential treasure trove because the web data has significant semantics waiting to be discovered. The prelude to such discoveries is to organize the data for the purposes of interpretation, transfer, sharing, and interoperability. This all falls into the domain of "web semantics" and tremendous international research effort has been spent in achieving the "semantic web". One of the well-known consortiums that try to provide guidance to achieve semantic web is the W3C (World Wide Web Consortium), which emphasizes annotating web data effectively with metadata systems (e.g. XML, RDF, and OWL) [Yu06, BRMT, Dymond].

Argumentatively, computing systems involve three components/strata as follows:

1) *Data to be acquired, manipulated, and processed*: Data items are classified into types and their values are assumed by pre-defined variables (e.g. Inter: X – X is a variable of the integer type), which should be operated by the suitable type of operators (e.g. multiplication).
2) *Logic to process data*: The data items are processed or transformed in sequence according to the pre-defined logical flow to yield the expected outcome. The logical flow is normally specified or designed explicitly

by the human operator, indicating what data are to be transformed by which operator and at what point in time. The design is expressed or implemented as a program, which is then transformed into the machine-processable software. Therefore, the program is for human understanding and the software or executable code is for machine operation (i.e. machine-processable).

3) *Information retrieval*: When a piece of software is in action it needs to fetch/store data from/in a designated place, and this happens continuously throughout the software's service lifetime. The place to facilitate information retrieval is literally the database.

Since the dawn of computer science, it has been a difficult task to achieve perfectly congruent computing strata despite continued research effort in the field (e.g. [Osterweil08]). This is indicated by different statistics, including the following types [Ausi00, Standish04, Coplien04]:

1) *High cost*: In Australia and USA, it is common for 50% or more of enterprise expenditures to be spent on software development and maintenance [Ausi00].
2) *Prone to failure*: Less than 50% of software development projects in the Western world were completed successfully [Standish04].
3) *Same trend*: The trend of roughly 70% software project failures will continue in the future, remaining the same as it was three decades ago [Coplien04].

Even though different software engineering techniques have been proposed over the last few decades [Osterweil08, Boehm08, Cheah07] software project success has remained at 30% today, as it was a few decades ago.

The Waterfall model (later augmented by fast prototyping) is perhaps the earliest software engineering approach, but it has the following shortcomings:

1) *Imperfect programming logic*: It is hard to produce an error-free program, which in an unobtrusive manner, can be impregnated with deadlocks and different type of bugs (e.g. time, logic and even Byzantine [Boyer90]). For this reason different techniques were proposed to express the programming logic in a way that logical correctness can be checked and verified immediately (e.g. Petri Nets [PetriNets]).
2) *Imperfect program logic transformation*: It is almost impossible to transform the logic in a large program into its executable form in a transitive fashion. That is why reverse software engineering has stagnated – one-to-many reverse versions.
3) *Implicit data semantics*: Traditionally the database is mainly concerned with where and how can data be stored and retrieved, and sometimes with data coherence emphasis. The semantics and associations among the stored data are implicit – it is subject to the interpretation of the beholder's context. Meanings and associations of the data items depend

on the program logic that processes it, and this programming approach is innately control-oriented.

In contrast to control-oriented programming, data-oriented programming triggers the execution of particular modules when the data items being waited for are ready. The typical data-oriented programming language is the LISP [LISP] and Linda [Yeung98]. Yet, the semantics of the data remain implicit.

4.2.1 Programming with Explicit Data Semantics

The main advantage of programming supported by the given ontology is explicit data semantics. This approach is supported by the 3-layer architecture: i) the top syntactic layer; ii) the middle semantic-net layer; and iii) the bottom ontology. Combining Gruber's and Guarino's [Gruber93a, Guarino95] argument together, the ontology should be perfectly and completely represented by its semantic net or DOM tree, the machine-processable form. The top syntactical layer abstracts the semantic net for human understanding and manipulation. If the 3-layer architecture of the software system works correctly, then semantic transitivity among them is always present. That is, if an entity is picked from any layer, its corresponding representations in the other two layers should always surface consistently. The ontological subsumption hierarchy is a collection of organized explicit semantic paths (i.e. explicit data semantics). Therefore, any program is simply a subset of this hierarchy, consisting of distinctive explicit semantic paths. If an automatic software generation

mechanism exists, then the correct software shall be generated from the subset (i.e. the specification) in one step. The only human intervention in the process is to provide the subset of subsumption hierarchy in the form of a specification. This is the original argument in the MI paradigm proposed by Nong's and then augmented and perfected in this research. The new/improved MI paradigm or EOD-ISD, as it is called in this thesis, differs from the error-prone Waterfall approach.

Since its dawn, software engineering has retained only a 30% success rate despite the fact that various techniques had been proposed to improve the situation. In general the techniques and tools fail to keep abreast of contemporary advances and the changing requirements. For example, traditional software debugging tools are not applicable for the parallel distributed software/processing or HPC (high performance computing) domain. Although program/software visualization looks promising for debugging distributed processing [Wong00, Katifori07], it is still a long way to maturity. Traditionally software development adheres to the Waterfall model, which starts with the user/functional specification. This specification is then converted into the corresponding design specification for implementation, and then testing. During implementation users may participate to ensure the functionality correctness (fast prototyping). To reduce inadvertent malicious changes the entity relationship (ER) diagrams may be used to clarify control and data flows. In this way, no variables or modular functions would be changed without first referencing the effect on others. Yet, the control flows and semantics in the program modules remain implicit, and this reduces the correctness of reusing

the modules. Debugging of these modules is also difficult because of the implicit semantics. Implicit semantics is one of the obstacles to successful reverse software engineering because the same executable code may produce several possible program versions of different logical representations.

Data-oriented programming has a wide connotation. In general it refers to invoking the software function only when the needed data is ready. This is well represented by the LISP declarative approach. In distributed computing data-oriented programming (DOP), however, focuses on designing the appropriate algorithm for the type of data intended to be processed with speedup [Wong00b]. The DOP approach in the distributed computing area is therefore characterized by the following:

a) A large database is partitioned into smaller manageable modules or blocks so that different programs or clones of the same program would process different blocks in parallel, in a total parallel for interleaved fashion.
b) The coherence of parallel processing endeavours or events are co-ordinated as a time sequence. Exchanges of data blocks and/or results among concurrent modules follow the pre-defined "network topologies" to cut the inter-process communication overhead. The efficiency of a distributed/parallel program can be gauged by the computation-to-communication (CTC) ratio. A high CTC ratio means more CPU time spent on doing useful work as verified by some previous research publications [Wong00b].

Yet, the DOP approach is still plagued by the problem of implicit program semantics, which depends on the programmer's interpretation of the logic flow for the actual software implementation.

4.3 Enterprise Ontology

Different problems due to implicit software semantics, which means that the interpretation of the semantics rests with the beholder only, have dented the software reliability of system developed by the Waterfall model. This led to the argument and proposal of the explicit-semantics (ES) software engineering [Rifaieh06], which is so far mainly for developing systems with ontology support. The term ontology has its origin in philosophy and it encompasses all the syllogistically correct arguments that are based on the axioms laid down. Within the ambit of the ontology of interest, for example, TCM (Traditional Chinese Medicine), there are facts (e.g. case histories) as well as theoretical arguments (e.g. treatises) waiting to be validated. Usually the ontology, which represents the ***total knowledge*** of a domain or area of expertise, is consensus-certified by domain experts over time. In this thesis, total knowledge is synonymous to ***global knowledge***. Total/global knowledge is the standard vocabulary of knowledge (or lexicon) of a "***community***", and its evolution has be sanctioned by repeated consensus certifications in a manual, laborious and communal manner.

In some ontology-related applications only the ontological factual parts are allowed, for example, clinical TCM practice. The factual part forms the

master or ***enterprise*** ontology of an establishment. Compared to the global/total/community ontology the enterprise ontology is the ***local*** master of the enterprise/establishment. In this light, other variants derived partially from the enterprise ontology are the target ontology constructs of the ***local-of-the-local*** nature. This thesis differentiates three levels of ontology constructs: i) global; ii) local/enterprise; and iii) target or local-of-the-local. For example, the pervasive Nong's mobile-clinics (MC) TCM telemedicine, known as the diagnosis/prescription (D/P) system, is supported by the Nong's master or enterprise TCM ontology core (TCM onto-core). This onto-core was created by extracting relevant information from the available TCM classics, treatises, and case histories, which, if put together exhaustively, would form the global ontology of the TCM domain. Since the extracted contents of the Nong's proprietary TCM ontology were consensus-certified by a sufficient number of TCM experts, it is the Nong's local enterprise TCM ontology – a corporation standard for clinical telemedicine practice. Yet, many of Nong's clients want their own customized versions of the Nong's D/P system. In order to cope with the demands as such, Nong's started to propose the "shell' meta-interface (MI) paradigm, which still lacked the details for implementation at the start of this thesis research. The aim of MI is to customize the local D/P systems automatically from the clients' specifications. On top the MI paradigm proposal, Nong's had transformed its original knowledge base for D/P operation into the corresponding enterprise TCM onto-core. ***Since the original MI paradigm proposal did not contain sufficient details for implementation, one of the objectives in this thesis is to make the MI paradigm work. As a result, prototypes in this thesis research can be quickly, correctly constructed to***

verify the proposed solutions in the clinical environment. In the MI paradigm, every target D/P system, including my prototypes, (at the local-of-the-local level) would have a subset of the Nong's master or enterprise TCM onto-core. The improved workable MI paradigm is then called the EOD-ISD approach, and it is an original contribution in light of ontology-based software engineering by this thesis. Therefore, the MI paradigm and/or EOD-ISD differentiate three levels of ontological constructs:

a) *Global ontology*: This refers to the complete TCM knowledge domain consisting of formal classics, syllogistic arguments, treatises, and case histories.
b) *Local enterprise ontology*: This is created by extracting the wanted portion of the global ontology of the domain. The Nong's master onto-core is an example.
c) *Local system ontology*: This is a subset of the enterprise ontology which is customized according to the given specification, for example, the Nong's MI specification (i.e. local-of-the-local level).

If the enterprise ontology is for supporting a computer system (e.g. the Nong's D/P system to help physician administer TCM medicine in a computer-aided manner), then it is an explicit conceptualization of concepts, entities, and associations among them [Gruber93a]. For practical purposes this conceptualization has a subsumption hierarchy of sub-ontologies [Guarino95]. Another typical example is the Unified Medical Language System (UMLS) [UMLS], which was developed by the US National Library of Medicine to

resolve differences in clinical terminology due to regional and/or national disparities. The UMLS also fits the concepts by Gruber and Guarino together as shown by its 3-layer architecture:

a) *Top layer/level* - the modularized query system (modules are semantic groups) for human understanding and manipulation.
b) *Middle layer* – the semantic net to logically represent the semantics in the bottom ontological layer; this is the machine processable form of the bottom layer.
c) *Bottom layer* – the integrated ontology [Rifaieh06] which is a subsumption hierarchy of the medical semantics organized by consensus certification into sub-ontological groups. Disambiguate terminology differences due to natural languages in a meta-thesaurus is present.

4.3.1 Similarity to the UMLS

The UMLS is not a clinical system but a consultation setup that people can interact to sort out terminology problems. Yet, the 3-layer architecture of the UMLS has provided some useful information on which to base the improvements for the shell MI paradigm:

a) If the bottom ontological layer is the "required" knowledge by consensus certification, the system created for a semantic group (top layer) is supported by the relevant portion of the ontology, isolated as the "local/target" system ontology.

b) If the semantic group can be correctly represented and verified, the MI mechanism should automatically generate the corresponding physical *customized* system variant, which is error-free and immediately ready for use. The automation cuts development cost and ensures customer satisfaction by its short development cycle. From the enterprise viewpoint, creation of a semantic group to meet the client's specific requirement is commercial customization. Extant techniques for managing web semantics may be adoptable [Taniar06].

c) If the "total" ontological knowledge could be drawn as a network or DOM (document object model) tree, the isolated "local/target" system ontology should have its customized semantic net for the local parser to work with. The parser finds the answer for the query $Q\{p_1, p_2\}$ at the syntactic semantic group level by inference. It traces out the unique operation or semantic path in the DOM tree in a stochastic manner for the input parameter set. Machine processing in the ontological context is parsing.

Figure 4.3.1.1 depicts the flow of the improved MI paradigm, which is alternatively known as the *enterprise ontology driven information system development* (EOD-ISD) approach of the following phases: i) the MI specification of selected icons is first created by the client; ii) the EOD-ISD generator extracts the portion of the master/enterprise TCM onto-core of Nong's to form the D/P TCM onto-core to be customized for the local system; iii) the EOD-ISD generator constructs the semantic net (or DOM tree) and inserts the standard ready-made Nong's parser; and iv) the EOD-ISD

mechanism constructs the GUI, which is the syntactical layer of the semantic net. The semantic net is the machine-processing form of the local TCM onto-core and the syntactical layer is the query system, which abstracts the semantic net, for human understanding and manipulation.

The main argument for the EOD-ISD approach is that the top syntactical layer should be a *graphical user interface* (GUI), which has the ***same appearance*** as the MI or iconic specification. Then, all the symptoms keyed-in via the GUI by the physician (e.g. s_1, s_2, s_3) are captured as actual parameters for the query (e.g. $Q(s_1, s_2, s_3)$). These actual parameters are implicitly (i.e. user-transparently) put together by the GUI system as a query for the parser to decipher. The parsing mechanism draws the logical conclusion from the DOM tree (i.e. the corresponding illness for the $Q(s_1, s_2, s_3)$ query). Precisely, the customized local system's ontological layer defines the ambit of the D/P operation. This layer is the vocabulary and the operation standard of the entire customized system. If a MI-based local D/P system has been correctly customized by the EOD-ISD approach, the three layers should be intrinsically and semantically transitive [Ng08]. With semantic transitivity, given an item from any layer the corresponding ones in the other layers always surface consistently.

From the above discussion, the UMLS and the MI approach are similar in the sense that the parser draws the logical conclusion for the input query. This conclusion is drawn from the semantic net, which is the machine-processable form of the ontology underneath.

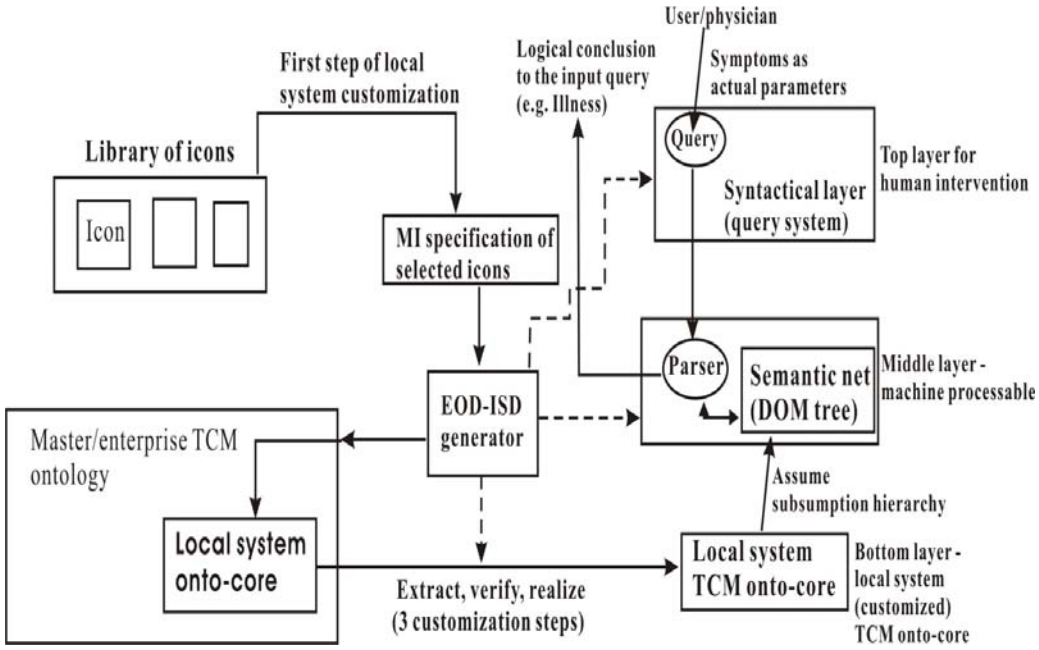


Figure 4.3.1.1 Local system customization flow with enterprise ontology support

Although Nong's had proposed the primordial MI paradigm "shell" first, there was insufficient detail for its implementation. It is an objective in this thesis to add new elements to make the primordial MI paradigm work for my research, in the sense that it can customize accurate prototypes quickly for tests in the real clinical environment – based on the Nong's enterprise TCM onto-core for clinical practice. Figure 4.3.1.1 is the practical model improved from the primordial MI paradigm in this research. It has new elements as follows: semantic transitivity; EOD-ISD automatic software system generator; master library of icons; W3C metadata for effective ontology annotations; living ontology (i.e. capable of real-time evolution; standard keyed-in ontological information; non-standard hand-written information; and support for useful discoveries. The improved MI version is the novel EOD-ISD

*(enterprise ontology driven information system development) approach, which is a contribution by this thesis. This approach also embeds another novel, important constituent conceptual element, namely the OCOE&CID (**On**-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery) technique.*

Some relevant experimental results that are selected to demonstrate local system customization by the EOD-ISD approach are presented in section 9.2.2 of Chapter 9.

4.4 Semantic Transitivity

For correct operation, it is important that the ***semantic transitivity*** exists among the three layers. Having semantic transitivity means that for any entity picked from any layer, its corresponding representations in the other two layers should always surface consistently. The Semantic Transitivity Visualizer (STV), which is another novel contribution by this thesis, enables the user to visualize and check anytime and anywhere the cross-layer semantic transitivity in the ontology. One of the experimental STV visualization results is presented in section 9.2.3 of Chapter 9.

4.5 Knowledge Discovery by Text Mining

Knowledge discoveries in the context of this thesis can be divided into two categories as follows:

a) *Living ontology support*: An enterprise ontology core such as the Nong's master/enterprise TCM onto-core is usually stagnated with classical information that was consensus-certified when the ontology was first created. In order to enliven the enterprise ontology so that it can evolve to keep abreast of contemporary scientific findings, we need two elements: i) text mining that searches the open sources (e.g. the open web and the repertoire of case histories accumulated by individual physicians) for new information; and ii) a way to append this newly acquired information to the extant ontological contents. The appendage should be temporary and real-time because permanent incorporation of new contents into the extant ontology requires another round of consensus certification.

b) *Discoveries of new prescriptions and herbal ingredients*: This is independent of but, if necessary, also includes the "living ontology". This aim is to discover new prescriptions and/or new herbal ingredients with respect to the named referential illness. The details for this category are presented in Chapter 5. In this thesis prescription discovery is more "high-level" than the herbal ingredient discovery (i.e. low-level).

The two principles to support knowledge discoveries in general are: i) automatic semantic aliasing that pairs with the technique of relevance indexing [JWong08c]; and ii) logical axiomatic relationship [JWong08a]. Some of the experimental results for demonstrating these two principles are presented in Chapter 9. The first principle above actually is the realization of the clinical SAME principle, which in Chinese terminology is known as "同病異治, 異病同治".

Text mining is used mainly to plough through the open sources to find new useful information. The text-mining tool used in the research is the WEKA, which is defined by the following WEKA parameters:

a) *Term frequency* (or simply tf): This weighs how important the i^{th} term t_i is, by its occurrence frequency in the j^{th} document (or d_j) in the corpus of K documents. If the frequency of t_i is tf_i and $tf_{l,j}$ is the frequency of any other term t_l in d_j, then the relative importance of t_i in d_j is $tf_{i,j} = \frac{tf_i}{\sum_{i=1}^{L} tf_{l,j}}$, for $l = 1,2,..i,..L$; $\sum_{i=1}^{K} tf_{l,j}$ includes tf_i.

b) *Inverse document frequency* (idf): It measures the importance of the t_i term in the corpus of size K as $idf_{i,k} = \frac{K}{\{d : t_i \in d\}}$, where d is the set of documents that contain t_i.

4.5.1 Making the WEKA Choice

The WEKA choice is based on the results of a careful survey. In fact, text mining is a branch of the data mining technique that can be used to discover knowledge patterns in textual sources [Chen96, Fayyad96, Holzman03, Bloehdorn05, Yu06]. In practice various inference techniques can be exploited for effective text mining [Walt06, Pedrycz98, Agrawal94], including case-based reasoning, artificial neural network (ANN), statistical approach, fuzzy logic, and algorithmic approaches. Table 4.5.1.1 summarizes our survey of text mining techniques and tools in the field. Our in-house experience, however, indicates that WEKA is by far the most effective text mining approach [Yu06]. For this reason it was adopted and adapted to support the OCOE operation. WEKA gauges how important a term/word is in a document/corpus (a bundle of documents) in a statistical manner.

Tools	Strengths	Weaknesses
Clementine	● Visual interface ● Algorithm breadth	● Scalability
Darwin	● Efficient client-server ● Intuitive interface options	● No unsupervised algorithm ● Limited visualization
Data Cruncher	● Ease of use	● Single Algorithm
Enterprise Miner	● Depth of algorithms ● Visual interface	● Harder to use ● New product issues
Gain Smarts	● Data transformations ● Built on SAS ● Algorithm option depth	● No supervised algorithm ● No automation
Intelligent Miner	● Algorithm breadth ● Graphical tree/cluster output	● Few algorithms ● No model export
Mine Set	● Data visualization	● Few algorithms ● No model export
Model 1	● Ease of use ● Automated model discovery	● Really a vertical tool
Model Quest	● Breadth of algorithms	● Some non-intuitive interface options
PRW	● Extensive algorithms ● Automated model selection	● Limited visualization
CART	● Depth of tree options	● Difficult file I/O

		• Limited visualization
Scenario	• Ease of use	• Narrow analysis path
Neuro Shell	• Multiple neural network architectures	• Unorthodox interface • Only neural networks
OLPARS	• Multiple statistical algorithms • Class-based visualization	• Date interface • Difficult file I/O
See5	• Depth of tree options	• Limited visualization • Few data options
S-Plus	• Depth of algorithms • Visualization • Programmable or extendable	• Limited inductive methods • Steep learning curve
Wiz Why	• Ease of use • Ease of model understanding	• Limited Visualization
WEKA	• Ease of use • Ease of understanding • Depth of algorithms • Visualization • Programmable or extendable	• (None visible)

Table 4.5.1.1 Strengths and weaknesses of the different text mining tools/techniques

4.6 Recap

The original Nong's mobile-clinic TCM telemedicine D/P system, which has been treating hundreds of patients daily in the Hong Kong SAR since its deployment three years ago, was developed by applying the traditional Waterfall model for algorithmic software engineering. This makes it hard for Nong's to customize D/P system variants correctly for individual clients. For this reason, Nong's proposed the shell MI paradigm and created the enterprise TCM onto-core by transforming the original knowledge base. The aim is to make ontology-based software engineering, which support automatic system generation, possible. It was unfortunate that the original MI paradigm did not contain sufficient details for implementation. This research has improved and transformed the MI paradigm into the novel *enterprise ontology driven information system development* (EOD-ISD) approach. In this novel approach several goals have been achieved: i) a D/P system can be generated in a single

step from the MI specification given using the named onto-core as the basis; ii) the customized onto-core is enlivened so that it evolves with time and otherwise it is stagnated with old classical TCM information within the ambit of the master enterprise TCM onto-core (as a closed onto-core); iii) the automatic semantic aliasing mechanism helps achieve the goals of standardizing new TCM knowledge and discovering new prescriptions for the referential illness; iv) the semantic transitivity among the three layers in the target D/P system can be verified anytime and anywhere the novel STV (Semantic Transitivity Visualizer); and v) the WEKA text miner ploughs through the open sources (e.g. open web) to find new knowledge and append it to the new "open" onto-core. WEKA was chosen only after a thorough survey of different text miner in the field. With the EOD-ISD, accurate prototypes can be customized for the experiments in the research.

4.7 Conclusion and Connective Statement

This research has transformed the original high-level MI paradigm proposed by Nong's into the novel EOD-ISD version. With the EOD-ISD accurate prototypes can be precisely customized in a single step from the given MI or iconic specification. In fact, the EOD-ISD approach is a framework that supports the following: i) customization of the corresponding prototype accurately from the named TCM onto-core in a single step from the given MI/iconic specification [JWong09c, JWong08b]; ii) opening up of the otherwise closed customized TCM onto-core via text mining; iii) continuous support for real-time onto-core evolution; and iv) enabling of useful and

meaningful herbal discoveries. Without the EOD-ISD support it would be difficult to verify the proposed methods for herbal discoveries in the clinical environment in a trustworthy manner. It is trustworthy because the enlivened ontology has explicit semantics and the semantic transitivity of the three layers in the prototypes can be checked and verified anytime, anywhere. Therefore, the next logical step is to explain the EOD-ISD approach, which is the wholesome essential software engineering support, in detail.

4.8 Key References

[Agrawal94] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, 12-15 September 1994

[Ausi00] Australian Bureau of Statistics, Finance, Australia 2000 Special Article – Information Technology and Telecommunications in Australia, 2000, http://www.abs.gov.au/Ausstats/abs@.nsf/0/9053E0EB512D0DDC4CA256F2A0007346F?Open

[Bassler98] O.B. Bassler, Leibniz on Intension, Extension, and the Representation of Syllogistic Inference, Synthesis, Springer 1998, 117-139

[Bloehdorn05] S. Bloehdorn, P. Cimiano, A. Hotho and S. Staab, An Ontology-based Framework for Text Mining, LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, May 2005, 87-112

[Boehm08] B. Boehm, Making a Difference in the Software Century, IEEE Computer Society, Vol. 41, No. 3, March 2008, 32-38

[Boyer90] R.S. Boyer and J.S. Moore, A Theorem Prover for a Computational Logic, Lecture Notes in Computer Science, Springer, 1990, 1-15

[BRMT] http://ilrt.org/discovery/2000/08/bized-meta/index.html

[Cheah07] C. Cheah, Ontological Methodologies - From Open Standards Software Development to Open Standards Organizational Project Governance, Computer Science and Network Security, Vol. 7, No. 3, March 2007

[Chen96] M.S. Chen, J.S. Park, and P.S. Yu, Data Mining for Path Traversal Patterns in a Web Environment, Proceedings of the 16th International Conference on Distributed Computing Systems, Hong Kong, 27-30 May 1996, 385-392

[Coplien04] J. Coplien, Organizational Patterns: Beyond Technology to People, Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, 2004

[Dymond] http://www.dymondassoc.com/metadata/mos.htm

[Fayyad96] U.M. Fayyad, S.G. Djorgovski and N. Weir, Automating the Analysis and Cataloging of Sky Surveys, in Advances in Knowledge Discovery and Data Mining, eds. Y.M. Fayyad, AAAI/MIT Press, 1996

[Gruber93a] T.R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, No. 2, 1993, 199-220

[Guarino95] N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, 1995, 25-32

[Holzman03] L.E. Holzman, T.A. Fisher, L.M. Galitsky, A. Kontostathis and W.M. Pottenger, A Software Infrastructure for Research in Textual Data Mining, The International Journal on Artificial Intelligence Tools, Vol. 14, No. 4, 2004, 829-849

[JWong08a] J.H.K. Wong, T.S. Dillon, A.K.Y. Wong and W.W.K. Lin, Text Mining for Real-time Ontology Evolution, Data Mining for Business Applications, Springer, 2008, ISBN: 978-0-387-79419-8, 143-150

[JWong08b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon**,** An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008

[JWong08c] J.H.K. Wong, W.W.K. Lin and A.K.Y. Wong, Real-Time Enterprise Ontology Evolution to Aid Effective Clinical Telemedicine with Text Mining and Automatic Semantic Aliasing Support, Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, 11-13 November 2008

[Katifori07] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis and E. Giannopoulou, Ontology Visualization Methods – A Survey, ACM Surveys, Vol. 39, No. 4, October 2007

[LaoZi] http://www.iep.utm.edu/l/laozi.htm

[LISP] http://en.wikipedia.org/wiki/Lisp

[Ng08] S.C.S. Ng and A.K.Y. Wong, RCR – A Novel Model for Effective Computer-Aided TCM (Traditional Chinese Medicine) Learning over the Web, Proceedings of the International Conference on Information Technology in Education (CITE), Wuhan, China, July 2008

[Osterweil08] L.J. Osterweil, C. Ghezzi, J. Kramer and A.L. Wolf, Determining the Impact of Software Engineering Research on Practice, IEEE Computer Society, March 2008, 39-49

[Pedrycz98] W. Pedrycz, Fuzzy Set Technology in Knowledge Discovery, Fuzzy Sets and Systems, Vol. 98, No. 3, 1998, 279-290

[PetriNets] http://www.petrinets.info/

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[Standish04] The 3rd Quarter Research Report: Chaos Demographics, The Standish Group International, 2004, http://standishgroup.com/sample_research/ darts_sample.php

[Taniar06] D. Taniar and J.W. Rahayu, Web Semantics & Ontology, Idea Group Publishing, 2006

[UMLS] http://umls.nlm.nih.gov/

[Walt06] C. van der Walt and E. Barnard, Data Characteristics that Determine Classifier Performance, Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, 2006, 160-165

[Wong00] A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, Local Compilation: A Novel Paradigm for Multilanguage-Based and Reliable Distributed Computing over the Internet, Special Issue: Mobile and Wireless Communications and Information Processing, Journal of Simulation, Vol. 75, No. 1, July 2000, 18-31

[Yeung98] D.S. Yeung and A.K.Y. Wong, The OORHS: A Conceptual Framework that Provides Easy and Reversible Distributed Programming, International Journal of Computer Systems, Science and Engineering, Vol. 13, No. 15, 1998, 289 - 301

[Yu06] S. Yu, H. Qiang and S. Jing, A Framework of XML-Based Geospatial Metadata System, Proceedings of the APWeb Workshops, 2006, 775-778

Chapter 5 Essential Software Engineering Support

5.1 Introduction

The focus of this research is to discover meaningful and useful herbal ingredients as well as prescriptions from a TCM (Traditional Chinese Medicine) knowledge base. This is precisely reflected in the project title: "***Web-based Data Mining and Discovery of Useful Herbal Ingredients (****WD^2UHI****)***". Yet, meaningful discoveries are possible only if the knowledge base is a consistent, classical consensus-certified product in the TCM domain. For this reason, my research makes use of the knowledge base of the extant Nong's deployed MC D/P system (MCS), which has treated hundreds of patients daily in the Hong Kong SAR for the last three years. This implies that the automatic MCS/WTS/CTSS system variant customization process is at least part of the original Nong's MC system knowledge base.

The original MI (meta-interface) concept proposed by Nong's to automate MCS for individual customers was driven by need. It is, however, difficult to follow through the Waterfall software engineering procedure, which was used in the development of the successful, extant, deployed Nong's MCS, for every new MCS customization. The original MI concept automates the whole software engineering customization process using icons. The argument is that if every icon represents a distinctive semantic path in the WTS, then gluing different icons together would make the final target system. The semantic path represented by every icon represents a piece of the relevant classical TCM

knowledge. Yet, the original MI proposal is short of explaining how icons could be constructed and integrated into the final WTS variant, which is interoperable with its cognate variants and other TCM systems that work with the canonical TCM knowledge to a varying degree.

Yet, the original MI concept by Nong's is useful in this research to automate the construction of a prototype to verify the findings at different stages in this research. The main argument is that if all the prototypes were customized from the same knowledge base, then subsequent findings would form a continuum of relevant results. The secondary argument is that if the consensus-certified knowledge base of the deployed Nong's MCS provides the basis for customizing the prototypes in my research, the experimental results derived from a real clinical environment would be meaningful and credible. These arguments led to my proposal of the new MI paradigm, which is a serious extension of the high-level, skeletal MI concept original proposed by Nong's. In fact, Nong's has also accomplished an important step, which is totally relevant to this research. That is, Nong's has produced an ontology version for its TCM master or enterprise onto-core. This means that part of this master onto-core can be used as the basis, with permission, for different semantically constructed prototypes from the same ontological basis for the experiments in the PhD research.

If the enterprise TCM onto-core was deemed as a sample space Ω of formal knowledge items (e.g. concepts, terms, attributes, and their associations with pre-defined axioms to constrain interpretations), which were extracted

exhaustively and consensus-certified by domain experts, Ω is logically a TCM ontology. In this light, Nong's master TCM onto-core, in effect, is $\Omega_{Nong's}$ or $\Omega_{PP/N}$ for the whole PP/N enterprise. It supports WTS clinical practice anytime, anywhere. If $\Omega_{PP/N}\{p_1, p_2, p_3, p_4....p_k\}$ is assumed, $\{p_1, p_2, p_3, p_4....p_k\}$ is the set of unique elements/events in $\Omega_{PP/N}$. If the following subsets, $\Omega_A\{p_1, p_2\}$ and $\Omega_B\{p_2, p_3, p_4\}$ are created from $\Omega_{PP/N}$, then they are the local TCM onto-core for the customized variants, WTS_A and WTS_B respectively.

The new MI paradigm is a formal software engineering technique that automates the generation of new software application from the given enterprise ontology core such as the Nong's master TCM onto-core. It is formal because the ontological contents are formal and therefore the icons that embed different unique explicit semantic paths are formal entities. This paradigm can effectively address different contemporary issues [Osterweil08] that deal with problem domains, varied operational environments, and cultural differences (e.g. natural languages). These issues usually complicate successful development of web-based applications, which are distributed over diverse geographic locations that cover various cultures and inevitably involve complex ICT (information communication technology) technologies and MSPM (multi-site project management) activities. It was repeated observed that even with the same project requirements the MSPM linguistic variations could still cause ambiguity, resulting in incorrect system implementation or non-interoperable software modules. One effective answer [Cheah07] to disambiguate enterprise-wide requirements is to set up a single vocabulary for all the enterprise software

activities – effectively the enterprise ontology or lexicon. In effect, the $\Omega_{PP/N}$ is an enterprise vocabulary example. In the Nong's case, it contributes to support more precise development and customization of mobile-clinic systems in the future.

5.2 Improved MI Paradigm for Automatic System Generation

The new MI paradigm that helps generate/customize cognate system variants from the given ontology is a more effective software engineering technique than the Waterfall model. In the paradigm, the user needs to provide only the specification of icons that represent the target system, the MI mechanism will generate an immediately deployable target system from the indicated master/enterprise ontology core. The key element is the ontology core, which is the vocabulary automatically nullifies multi-representations of the same word or statement [Cheah07]. The lack of an effective enterprise vocabulary to coordinate and disambiguate software development activities means a high cost-effectiveness ratio, as confirmed by the following surveys:

1) *High cost*: In Australia and USA, it is common for 50% or more of enterprise expenditures to be spent on software development and maintenance [Ausi00].
2) *Prone to failure*: Less than 50% of software development projects in the Western world were completed successfully [Standish04].

3) *Same trend*: The trend of roughly 70% software project failures will continue in the future, remaining the same as it was three decades ago [Coplien04].

The weakness of the Waterfall model is possible superfluous human interventions. For example, in a large software engineering project this kind of intervention could be uncontrollable due to the large number of possible collaborating development centers for the same project. From a high-level point of view, the new MI paradigm proposed in this PhD thesis still can be abstracted by four generic Waterfall phases (Figure 5.2.1):

1) Phase 1 - Requirement specification and analysis. Its goal is to analyze and accurately extract the following elements from the narrated requirements: the necessary and sufficient number of functions for the target system; formal parameters for each of these functions; and the execution serializability (logical control flow) among the identified functions to ensure coherent and meaningful results. A function normally performs only one application-specific task of transforming the actual parameters into the expected result. For example, if $f(x_1, x_2)$ is a function, (x_1, x_2) are two formal parameters that would assume actual values/parameters before execution (i.e. the transformation process). The functions and their intertwined logical relationships form the *functional specification*; constraints specified for these relationships form the *constraints specification* to govern the ambit of system behavior/dynamics. The *functional* and *constraints* (F&C) *specifications*

together form the domain of semantics for the system to know exactly *what to do*. In the MI paradigm the requirement specification is the ***iconic specification***.

2) Phase 2 – Design specification. Details of *how the final system should work* are addressed by: i) organizing the system semantics into small manageable modules (modularization) by the principle of *information hiding*; ii) specifying how the modules should synchronize and associate; iii) proposing the subsumption hierarchy for the modules that can be separated into two basic groups by their nature: control-oriented (CO) and data-oriented (DO); while the higher-level CO does little computation but controls the timely invocation of other modules, the objective of the lower-level DO is to produce useful information from actual parameters for use by the higher-level modules; iv) proposing the system architecture to support the final system operation; and v) evaluating data structures and algorithms/protocols to support information retrieval and inter-modular synchronizations for coherent operations. ***Design in the MI paradigm means: i) representing the semantic paths, which are extracted from the master ontology core, as functionally distinctive icons; and ii) integrating some icons to represent the desired functionality for the target/customized system.***

3) Phase 3 – Implementation. This phase aims to correctly translate the design specification into an intermediate form for: i) human understanding and manipulation, and ii) conversion into the machine-

executable representation. The intermediate form is a program or software of a specific language (e.g. C++ or Visual Basic). To humans, the program syntactically represents the system semantics; the machine executes its compiled form (executable code). ***In the MI paradigm, the MI mechanism instantiates the icons with the corresponding executable codes automatically to form the target system – automatic system generation.***

4) Phase 4 – Testing and debugging. Test cases are created to validate and verify that the implemented system prototype indeed fulfils all the functions indicated in the requirement specification. Debugging a distributed application is more an art than a science, for we can rarely apply traditional approaches. From the literature, the only recognized technique to debug distributed software effectively is program visualization (e.g. [Wong00, Katifori07]). ***Since the system generated by the MI mechanism is immediately deployable, it goes into the system validation phase right always for test runs.***

The feedback loops (Figure 5.2.1) show that if errors are found in the software engineering process then changes have to be repeatedly made in the upper source(s). Therefore, too many loop-backs make this software engineering process expensive. As a result many experts emphasized the importance of producing logically correct specifications, and this can be achieved by adopting appropriate formal methods (e.g. Petri net). Then, the chance for errors to occur during the translation of the verified user

specifications into the corresponding design specification can be reduced by using semi-formal, semi-automatic tools (e.g. DBDesigner (DBD) by Microsoft). For example, the extant/old Nong's MC D/P system was developed by using the Waterfall model augmented by the technique of fast prototyping and adoption of the DBD. The fast prototyping process allows user participation throughout the whole development process so that the system functionality can be monitored continuously. The DBD helps the engineer draw the "semantic representation" of the TCM knowledge base, which in the MI concept, is the TCM ontology core (onto-core), in the form of a subsumption hierarchy. (In the MI paradigm this hierarchy is called the semantic net of DOM (document object code) tree. It is the machine-processable form of the TCM onto-core underneath.) The format of the DBD drawing matches the XML-annotated knowledge base. The SQL system (also by Microsoft) converts the annotated code directly into a usable SQL database. If the CO and DO modules are programmed in VB.net (Visual Basic for the Internet), they interact readily with the SQL database. The DBD, SQL, VB.net together form the suite that underlies the Nong's Waterfall based software engineering process at the time, namely, the *congruent automation principle* (CAP). The CAP, in effect, is the basis for the Nong's *configuration control* framework of the Waterfall era. With respect to Figure 5.2.1, the rest of the configuration control scheme would take care of other development aspects such as system migration, software changes, system versioning, and maintenance.

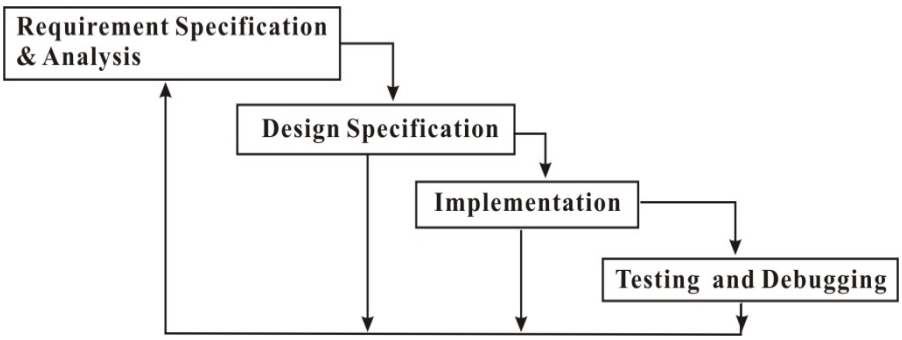


Figure 5.2.1 Generic waterfall development life cycle

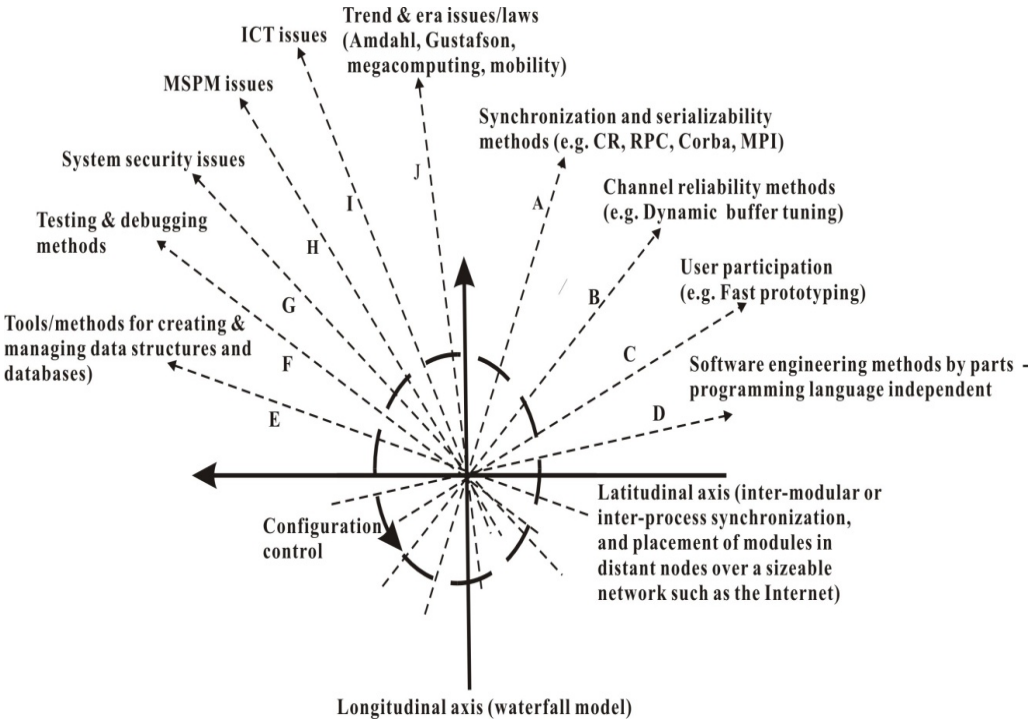


Figure 5.2.2 Ten external forces that affect software system success

In fact, the new MI paradigm for automatic system generation provides a solution to overcome some of the formidable challenges in software engineering of the 21st century. These formidable challenges include cultural diversity that leads to ambiguous understanding of the target system. The MI paradigm naturally avoids THWADI ("*that's how we've always done it* [Boehm08]"). It is an effective answer to successful development of

immediately, remotely deployable pervasive web-based telemedicine systems, which is an emerging phenomenon of this century. In reality, the THWADI guideline is useful because computing requirements evolve rapidly in different eras, governed by the Moore's Law [Lewis96, Bardram07]: i) Amdahl's era (early 1960s) – focusing on how to synchronize sequential processes correctly; ii) Gustafson-Barsis era (mid-1980s) – emphasizing parallel computing (i.e. High Performance Computing (HPC)) to yield speedup; iii) megacomputing era (mid-1990s) – concentrating on forming distributed systems on the Internet; and iv) pervasive era (early 2000) – concerning about mobility of hardware and software entities as well as the supporting location-aware capability. Accumulated statistics and recent history tell us that the intensive traditional configuration control endeavour alone can hardly improve software quality. Otherwise, we would not have stuck with the same situation as 30 years ago today – roughly 70% software project failures. Sadly, such failure rate will continue in the future [Coplien04] unless new and powerful paradigms such as the MI emerge. A hard fact for quality software engineering today is that traditionally usable tools fail to keep up with the speed of the evolutionary requirements.

Figure 5.2.2 summarizes the difficulty of devising a configuration control scheme that will work effectively all the time. The difficulty lies in the changing forces, and this means a static configuration control scheme would hardly be able to cope with the dynamic magnitude changes in these forces. In the preliminary investigation of my research I have identified from the literature the ten most influential forces that determine the success of a software

engineering project anytime. These are represented as entries in the set $F = \{A, B, C, D, E, F, G, H, I, J\}$ as depicted in Figure 5.2.2:

A. *Synchronization and serializability methods*: These govern how entities in the system interact coherently. Examples include CR (critical region), RPC (remote procedure call), Corba, and MPI. The method used depends on the problem domain and the intended environment of operation.

B. *Channel reliability methods*: These shorten the service roundtrip time in client/server interaction. Usually dynamic or adaptive methods are more effective than static methods [Lin06].

C. *User participation*: This is a necessity for effective fast prototyping so that immediate user feedback improves the prototype. It is ideal if the user participates in all stages of the Waterfall model.

D. *Software engineering methods by parts*: This is integration of software parts (modules/artefacts) built by other groups into the system being built. It can be physical code inclusions (into the system software) or logical remote invocation via pre-defined linkages. The parts can be in various programming languages but do not affect the final system performance [Wong00].

E. *Tools/methods for creating/managing data structures and databases*: These represent the paradigm that data structures on the blueprint are realized automatically into physical databases; for example, converting a DBD drawing directly into a physical SQL database (i.e. Microsoft environment).

F. *Testing and debugging methods*: These support different testing and debugging situations. For example, program/system behavior visualization is suitable for monitoring distributed agent-based software in which agents are mobile in a real-time sense [Wong00].

G. *System security issues*: The aim is allow a system to run smoothly without unnecessary interruptions.

H. *MSPM (multi-site project management)* [Cheah07]: Usually teams based on different geographic locations are involved in the development of a successful enterprise software system. To eliminate ambiguity, a vocabulary to bridge cultural and language differences among working groups needs to be created. The creation of such a vocabulary is regarded by many researchers as an ontological approach (i.e. the vocabulary is the "enterprise ontology") [Uschold07].

I. *ICT (information communication technology)*: This discipline combines appropriate technologies to build an efficient web application.

J. *Trend and era issues/laws*: Inevitably, as the computing industry advances through various trend-setting eras and laws into today's mobility era with mobile hardware for location-aware networks and mobile software agents that migrate at will, some of the older methods and tools will be invalidated.

The longitudinal and latitudinal axes in Figure 5.2.2 form the backbone of the configuration control to achieve equilibrium among these ten forces. Although the Waterfall model was the basis for the configuration control, the

two key issues of modular task placement into network nodes and ensuring correct task synchronization to achieve coherent results would still need to be addressed. Unfortunately, no previous experience on devising an effective configuration control scheme for pervasive telemedicine system development has been found in the literature. The Nong's in-house experience, which is the result of trial-and-error, is the only useful clue to this research so far.

From the above discussion it is clear that effective software engineering support is essential for customizing interoperable telemedicine system variants successfully. The 10-force influence indicates that it is hard to guarantee quality of the target system simply through the configuration control scheme. By discarding THWADI philosophy, it becomes obvious the MI paradigm is a new direction to follow. The essence and first step of this paradigm is the construction of the underlying domain/enterprise ontology core by consensus certification. Once the onto-core is present, the next step is to construct the reusable semantic icons that each represents a distinctive semantic path in the onto-core. The third step then is to construct the MI mechanism to automate the generation of the target system. In this thesis, these steps form the novel *enterprise ontology-driven information system development* (EOD-ISD) approach – a synonym of the improved primordial MI paradigm originally proposed by Nong's. The EOD-ISD, however, adds a new conceptual element to the MI paradigm, the *living onto-core*. This consideration appeared after some important new research findings, which also led to following observations:

i) Though the enterprise ontology is a consensus-certified knowledge base, it could become stagnated with information from the past – up to the time of when the consensus certification was confirmed.

ii) The knowledge in a domain is evolving; new scientific findings and experience of individuals can be absorbed continuously to enrich the enterprise ontology – *the living ontology issue*.

5.3 Inspiration for the EOD-ISD Approach

The EOD-ISD (*enterprise ontology-driven information system development*) inspirations came from one need and two main sources of information. The need is to create *WD²UHI* prototypes for meaningful experimental results in light of discoveries of herbal ingredients. The two main sources are as follows:

a) *Original MI paradigm concept*: Nong's proposed this original framework to facilitate the customization of interoperable cognate telemedicine system variants, namely the MC D/P systems. The core concept is to generate the cognate systems variants from the same TCM onto-core. The 3-layer architecture of the extant MC D/P system developed by the Waterfall approach has been working functionally well, except that any customization from the master TCM knowledge base is prone to error because too much human intervention on the process. The master TCM knowledge base was built from selected medical knowledge items that are suitable for clinical operation by a team of

domain/medical experts via a lengthy "pruning & agreeing" process – consensus certification. In the original MC D/P system the three layers are: i) the bottom layer – the consensus-certified knowledge base, which later became the master/enterprise TCM ontology core; ii) the middle parsing mechanism that works with the semantic subsumption hierarchy of the bottom layer; and iii) the query system that enables the user to interact with the system. In effect, the bottom layer is the repertoire of knowledge with entities, concepts and their associations defined. The middle layer is the form of the bottom layer for machine understanding – machine-processable, and the top query layer abstracts the bottom and middle layer for human understanding and interaction. *The knowledge base only elicits facts that pertain to clinical operations.*

b) *Semantical UMLS (Unified Medical Language System)* [UMLS]: This was developed by the US National Library of Medicine to resolve the differences in allopathic clinical terminology due to regional/national disparities. It makes use of ontology in the bottom layer that represents the medical knowledge included. The 3-layer architecture of UMLS are: i) top syntactical layer – query system modules (i.e. semantic groups) for human understanding; every group is a subset of the ontology in the bottom layer; ii) middle logical layer – semantic net to logically represent the semantics in the bottom ontological layer (machine-processable form of the ontology below; this is the basis from which the parser in the middle layer draws logical conclusions by inference); and iii) the bottom ontology layer – an overall integrated ontology (of

various interconnected sub-ontologies), which is a subsumption hierarchy of the medical semantics organized by consensus certification. *Unlike the Nong's MC D/P system, the UMLS is a consultation system rather than a clinical system.*

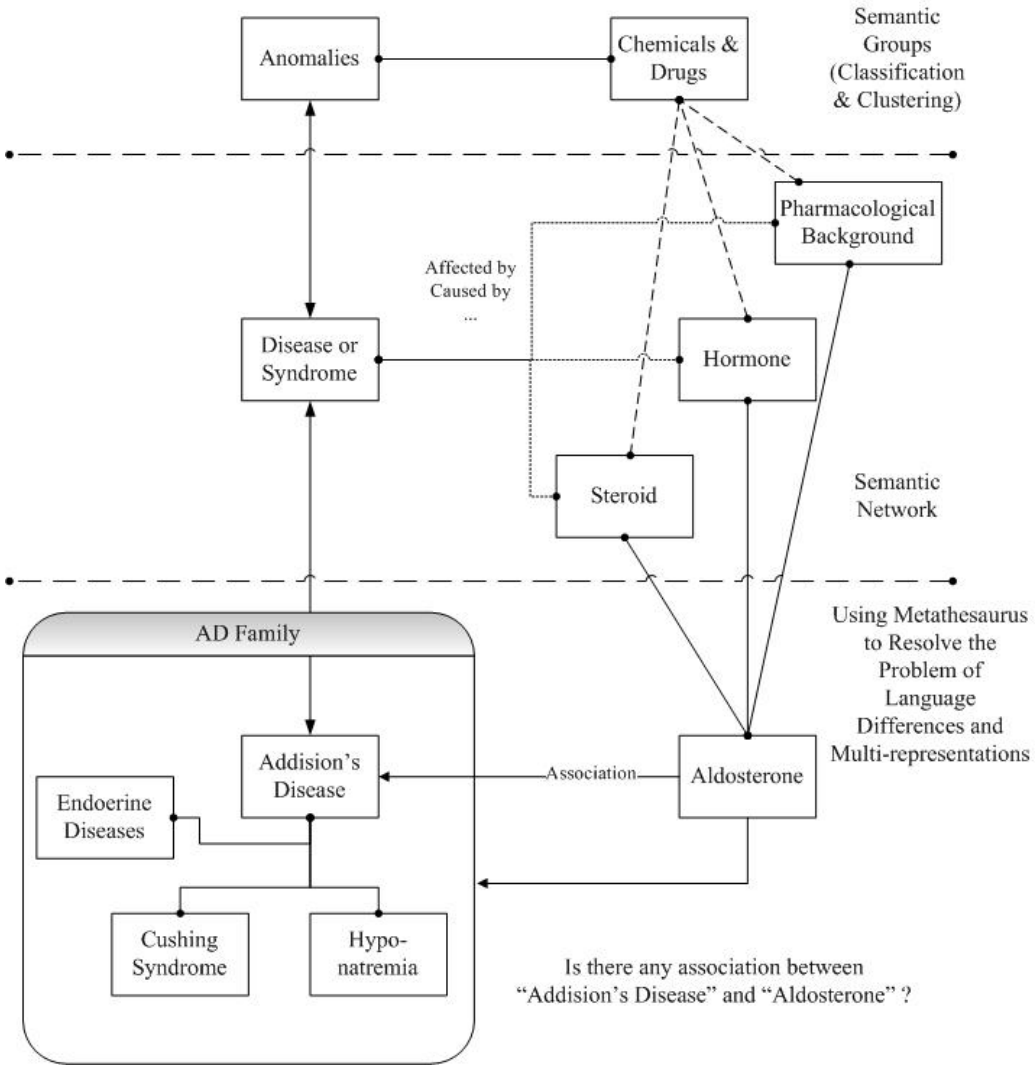


Figure 5.3.1 The 3-layer UMLS architecture [UMLS]

Figure 5.3.1 shows the 3-layer UMLS architecture. The AD family is part of the annotated ontology (bottom layer), which would be transformed into the semantic net in the middle layer so that the parser can use it to decipher the query "*Is there any association between Addision's Disease and Aldosterone?*".

This query comes from the top syntactical layer, and the different queries can be classified into semantic groups with respect to the nature of their functions.

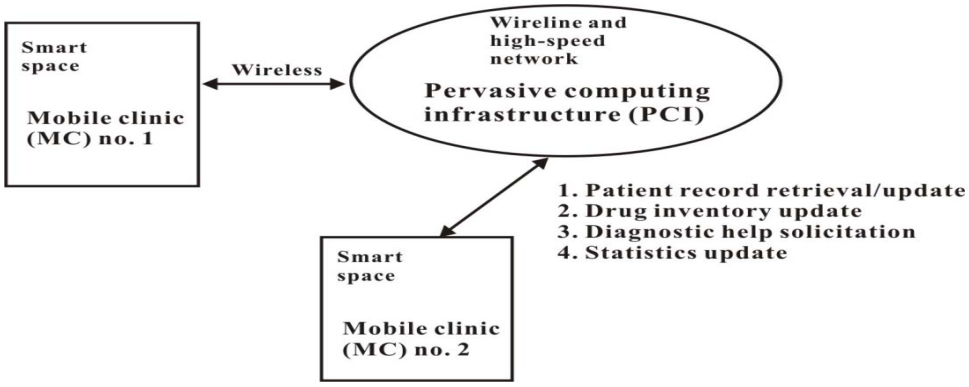


Figure 5.3.2 A Nong's pervasive MC-based telemedicine D/P system

The desire to combine the good aspects of the successful Nong's MC D/P system for clinical telemedicine and the UMLS allopathic consultation system gave rise to the new MI paradigm. Later the motivation to enliven the bottom ontology had metamorphosed the MI paradigm to the EOD-ISD approach, which is equipped with the **novel** STV (Semantic TCM Visualizer) for semantic transitivity checking (among the three layers that will be explained later). From my point of view the EOD-ISD is urgently required for quickly generating prototypes for the WD^2UHI experiments, which should be conducted in the Nong's MC clinical environment.

Telemedicine refers to administering medicine over the mobile Internet, and this involves wireless and wireline communications [JWong08b]. A typical example is the Nong's TCM telemedicine system consisting of collaborating mobile clinics (MC) over the mobile Internet; this web-based system has been

deployed with success in the Hong Kong SAR. Each MC is equipped with a diagnosis/prescription (D/P) system with which the physician on board the MC treats its local patients – computer-aided diagnosis, and electronic prescription and dispensing [Lin08]. The software of the original D/P system, however, was developed based on the traditional Waterfall model with fast prototyping. That is, the D/P system development starts with a functional specification, which was then converted into the design specification for verification, implementation and testing. Although the old Nong's D/P system works well in its present form, it is not easy to customize it to suit the needs of different clients who have distinctly varied telemedicine D/P requirements and operating environments. In the old Nong's D/P system, every system customization requires a repeat of every step in the whole Waterfall model, and debugging can be messy and prone to errors for two reasons: i) the D/P system is intrinsically distributed and therefore traditional debugging tools are not applicable, and ii) the semantics of the system are implicit.

The Nong's telemedicine D/P system setup is particularly relevant for this research, and this research helps verify and validate the EOD-ISD approach in the Nong's environment. For this reason Nong's permitted us to build our prototypes with the Nong's proprietary TCM enterprise ontology core (TCM onto-core) as the master/basis. From this master enterprise ontology different local D/P systems prototypes for experiments are customized for verification and testing purposes. Figure 5.3.2 depicts the Nong's D/P system operating environment in which mobile clinics (MC) collaborate over the pervasive computing infrastructure (PCI). An MC has to move into a smart space (a

wireless communication cell) before it can communicate with its peers, as well as its dedicated surrogate server. The surrogate server, which helps the MC seek central support, collaborates with others over the central high-speed wireline network. Other than treating the local patients in a computer-aided manner, the MC also carries out other remote tasks, including: i) patient record retrieval and update; ii) updates of the local drug MC inventory as well as the remote central one within the PCI; iii) diagnostic help solicitation from remote physicians, and iv) central update of the statistics for effective MC scheduling and disease control (as required the Hong Kong SAR government). An MC vehicle is normally manned by a physician, a dispenser, and a paramedic, and supported by a customized telemedicine D/P system.

With the help of the EOD-ISD, a *WD*^{2}*UHI* prototype is constructed in the form of a MC D/P system. Then, physicians will be invited to use and validate the prototype in a true clinical environment before it is used for conducting herbal ingredients discoveries experiments.

The *WD*^{2}*UHI* prototype is basically a CTSS (customized telemedicine software system), derived from the given iconic specification. The 2-dimensional view of the 3-strata CTSS is shown in Figure 5.3.3. A physician interacts with the system through the GUI (1st stratum). The operation of the system is semantic and supported by the ontology in the 2nd stratum, and this ontology evolves by absorbing new knowledge from open sources (3rd stratum).

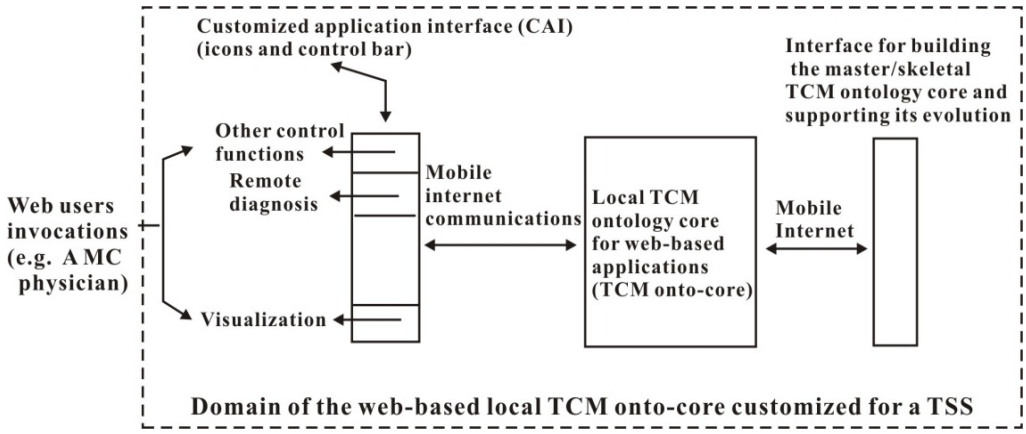


Figure 5.3.3 The 2-dimenional view of a *WD²UHI* prototype

The *WD²UHI* prototype, basically a CTSS (customized telemedicine software system), is derived from the given iconic specification. From this specification, the following are automatically constructed by the EOD-ISD mechanism:

a) *CAI (customized application interface)*: This has the same appearance as iconic specification (to be explained later). This interface enables the physician in the mobile clinic to interact with the telemedicine system for computer-aided diagnosis and prescription preparations. In effect, this is the top query layer of the MC system.

b) *Local TCM onto-core*: This is automatically created from the enterprise TCM onto-core, from a blueprint embedded in the iconic specification. This local TCM onto-core does not evolve by itself. Meanwhile, the semantic net and the corresponding parsing mechanism are also created. In effect, this is the middle semantic-net layer for machine processing; the STV is also hooked up automatically for semantic transitivity checking (if invoked).

c) *Support for evolution*: A special mechanism, which can be invoked or inhibited by the user, is made ready to support and enliven the local TCM onto-core. This mechanism is supported by the following: i) automatic real-time text mining over the open web; and ii) automatic semantic aliasing for prescription discovery and similarity measurement of the discovered prescriptions by computing their relevance indices. This special mechanism does not change the skeletal local TCM onto-core generated from the given iconic specification. Rather, new discoveries are part of the evolutionary process and are only temporarily appended to the skeletal onto-core. Eventually the discoveries can be incorporated into the master/enterprise TCM onto-core if the subsequent consensus certification allows this to be done.

Figure 5.3.3 is the 2-dimensional view of a *WD*^{2}*UHI* prototype of three operational strata: i) CAI or GUI (graphical user interface) as first stratum; ii) local TCM onto-core in the middle second stratum; and iii) support for evolution at the rear or third stratum. Figure 5.3.4 is the corresponding 3-dimensional view of more details.

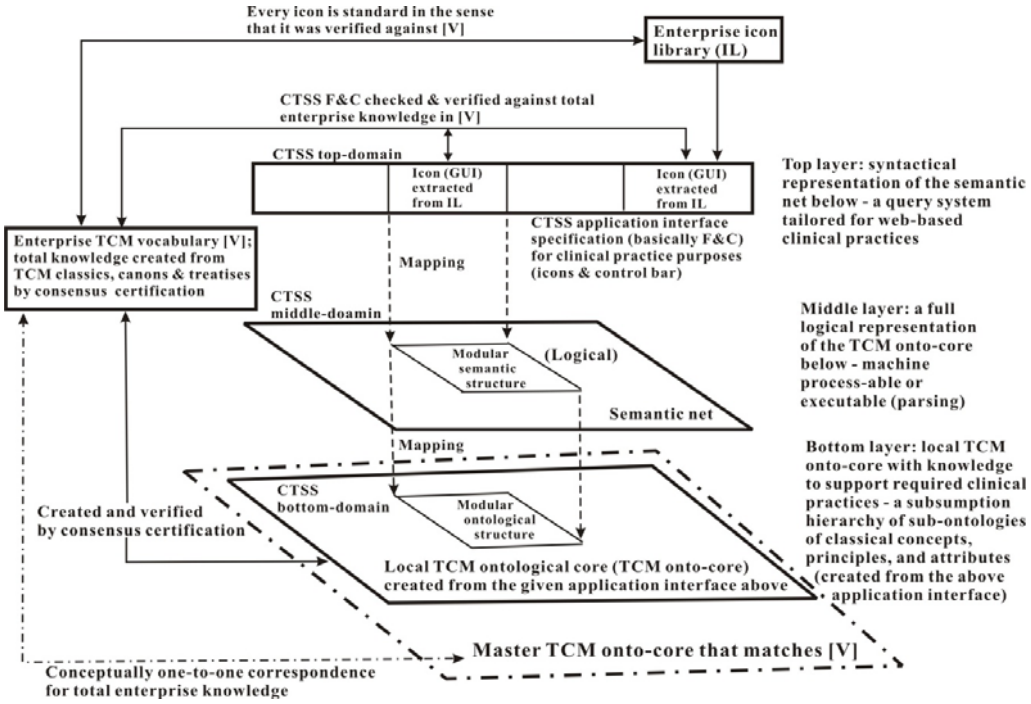


Figure 5.3.4 The 3-dimensional view of the CTSS

Figure 5.3.4 also provides details for the EOD-ISD mechanism in the process of customizing a telemedicine software system (TSS) from the given CAI/MI/iconic specification. What happens here involves the three layers are as follows:

a) *Bottom layer*: This shows how the local TCM onto-core for the target TSS is customized from the enterprise time-honored total knowledge or master/enterprise TCM onto-core. The lexicon of this layer is logically represented by [V]. The lexicon or vocabulary in [V] supports all the Nong's clinical practice in a necessary and sufficient manner. The modular ontological structure (bottom layer) in Figure 5.3.4 matches its logical representation for machine understanding and processing – the modular semantic structure (middle).

b) *Middle layer*: This is the semantic net (network) that fully and logically represents the local customized TSS – the machine-processing form. The parsing mechanism (parser) is the software that draws the logical conclusion for the queries input from the top layer (e.g. $Q\{p_1, p_2, p_3\}$; $\{p_1, p_2, p_3\}$ are parameters to drive the parsing mechanism).

c) *Top layer (i.e. CTSS top-domain)*: This is the customized application interface (CAI) specification for the target CTSS (i.e. functional & control (F&C) specifications together) to syntactically represent the local CTSS semantic net for human understanding. The CAI specification is made up of icons selected from icon library (IL); new icons can be created and added to IL anytime. The terms in an icon are standardized by [V]. The whole CTSS is automatically realized from the given iconic or CAI specification by EOD-ISD mechanism. The physical GUI of the target TSS has the same appearance as the given CAI specification.

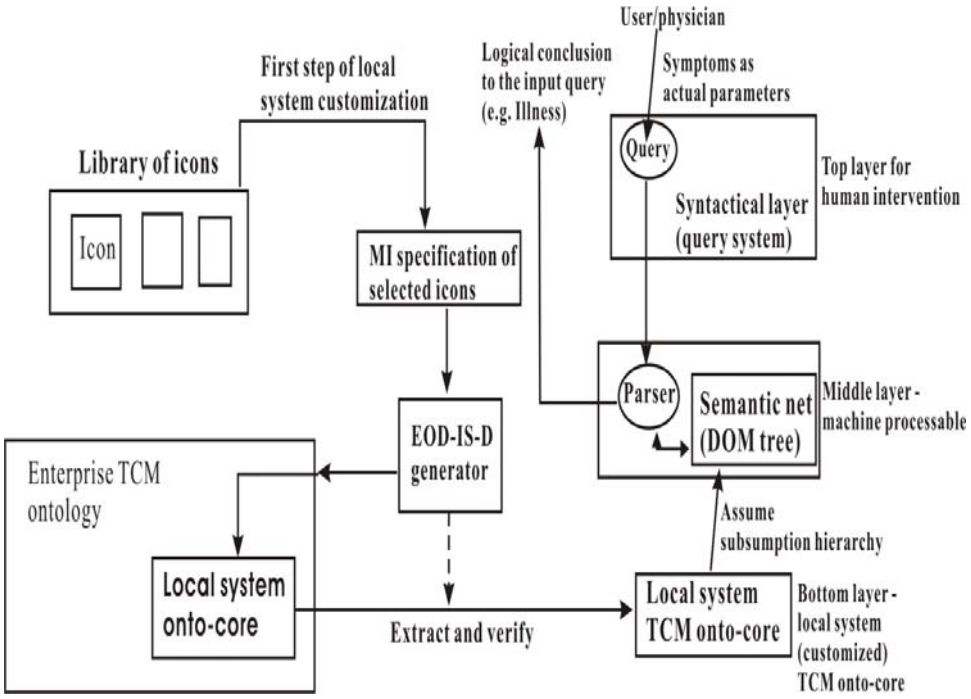


Figure 5.3.5 EOD-ISD approach overview (excerpt of Figure 4.3.1.1)

In the whole EOD-ISD process, *the most engineered CTSS part and also the only part* that involves human intervention is the top layer – CAI specification. Once the iconic/CAI specification is confirmed, the rest of the system is automatically generated. Working together, the three layers effectively realize the philosophical arguments of Gruber [Gruber93a] and Guarino [Guarino95] in an integrated fashion. Gruber's ontology is an explicit consensus-certified conceptualization, which is understandable to humans and machine-processing at the same time. Guarino deepened this ontology concept by arguing that it should have a subsumption hierarchy consisting of sub-ontology constructs with axiomatic associations to constrain interpretation, as exemplified by the UMLS.

Figure 5.3.5 (excerpt of Figure 4.3.1.1) is the EOD-ISD approach overview. The user needs to provide only the iconic/CAI/MI specification for the EOD-ISD generator to automatically generate an immediately deployable system. The actual graphical user interface (GUI) created from the iconic specification has the same appearance as the specification (to be explicated later). This lets a physician treat patients in a computer-aided manner. All the symptoms keyed-in via the GUI by the physician (e.g. s_1, s_2, s_3) are captured as actual parameters for the query (e.g. $Q(s_1, s_2, s_3)$) to be implicitly (i.e. user-transparently) constructed by the GUI system as input to the parser. The parsing mechanism draws the logical conclusion from the DOM tree (i.e. the corresponding illness for the $Q(s_1, s_2, s_3)$ query). The iconic specification is by made up of a set of icons selected from the icon library. Icons can be defined, re-defined and added to the library at will. The basis for the target system to work semantically correctly is the master/enterprise TCM onto-core. The EOD-ISD mechanism extracts only those semantic paths embedded in the MI specification to create the onto-core for the target system. For this reason semantic transitivity [Ng08] always exists among the three layers in the customized TSS. To summarize, the key elements in the EOD-ISD include:

a) ***Enterprise TCM vocabulary:*** All CTSS terms are verified against it (i.e. [V] in Figure 5.3.4).
b) ***Unique icon library (IL):*** This contains all the graphic icons that Nong's accumulated over time. Any new icons created for customers will be added to IL as evolution. An icon in the context of the MI paradigm is a modular semantic structure backed up by its modular

ontological structure (Figure 5.3.4). For machine processing, every icon is supported by a group of "control-oriented" and "data-oriented" object classes. An application interface to be customized is physically a collection of selected icons from IL that meet specific clinical functions of stated constraints. Icon creation is a formal process, for its terminology is checked and verified against the standard enterprise vocabulary [V]. This disambiguates communications within the Nong's enterprise, between Nong's and the global TCM community, and among the Nong's customers and collaborators.

c) ***Customized application interface (CAI):*** In its business plan, Nong's would customize MC telemedicine packages and then remotely install them at the individual client sites [JWong08b]; completely in an automated manner. The customization process is basically fast prototyping that involves only the CAI/iconic/MI specification. With the final CAI specification the customization and remote installation of the TSS artifacts (at the client's site) are automated. Verification and validation of the target system can be conducted anytime and anywhere via the *semantic TCM visualizer* (STV) – a mandatory element in the proposed EOD-ISD approach.

d) ***Annotated master/enterprise TCM onto-core blueprint:*** This is the huge piece of annotated code (or blueprint) for the subsumption hierarchy of the entire enterprise TCM onto-core to match the formal knowledge in the enterprise vocabulary [V]. The blueprint creation is semi-automatic to quicken rectification of errors by the group of TCM

domain experts who perform consensus-certification. This semi-automatic process has two phases:

i) ***Manual phase:*** The DOM (document object model) tree for the master TCM onto-core has to be drawn manually. The drawing helps experts visualize and verify the necessary facts quickly against the canonical information in [V]. In fact, there are usable commercial tools in the field that can support such drawing; the DBDesigner (DBD) by Microsoft is an example.

ii) ***Automatic phase:*** Firstly, the annotated blueprint is automatically generated from a drawn DOM tree. Annotation can be achieved by different metadata systems. For example XML, RDF, and OWL metadata systems are popular because the codes generated for them are interoperable [Rifaieh06]. In fact, the DBD system can generate the corresponding XML-annotated codes from its own drawings. Secondly, the GUI (graphical user interface) subsystem is automatically generated for the final WTS system for human interaction.

e) ***Automatic CTSS/WTS database generation:*** A physical CTSS/WTS is generated from the given CAI specification that indicates what portion of the enterprise TCM onto-core blueprint to be extracted automatically by the MI paradigm. The extraction, in the form of a piece of annotated code (blueprint), is then automatically instantiated into the respective local TCM onto-core. For example, this automation can be achieved

with the SQL system by Microsoft. Previous experience with Microsoft showed that the following sequence was viable: i) produce a DBD drawing manually; ii) generate the XML-annotated code (blueprint) for the DBD drawing automatically; and iii) create the physical SQL database automatically for the blueprint.

f) ***Appropriate programming language(s) for the logical object classes:*** The executable forms of those functions in an icon in the IL are object classes. In the MI, paradigm functions in an icon are instantiated as object classes selected from the main enterprise object library; the MI paradigm is object-based. Many languages can effectively support the object-based programming paradigm. Yet, the *congruent automation principle* (CAP) in the context of this research dictates the choice of the programming language. The CAP rules that the final customized WTS should be automatically generated for the given CAI specification. As a result the iconic object classes should be able to seamlessly interact with the final physical database. The commercial package offered by Microsoft – "DBD, SQL, and VB.net (Visual Basic for Internet)" – is in fact, one way to realize the CAP, for the translational process from the DBD drawing of a DOM tree to the corresponding physical SQL database is coherent. If the iconic functions are programmed in VB.Net, they can seamlessly interact with the final SQL database immediately.

g) ***Semantic TCM visualizer (STV):*** This novel contribution by the thesis converts an XML-annotated code into the matching DOM tree and traces the parsing mechanism on-line. In this way, it verifies and validates any parts of the physical CTSS anytime and anywhere.

h) ***Remote CTSS installation:*** The CTSS package contains: the GUI for human interaction; wireless communication capability for the MC; the CTSS database; object classes; and other auxiliary software tasks. It is sent via the web to remote sites for installation.

Using CAI specification as the input to automate the TSS generation process eliminates some serious MSPM (multi-site project management) problems (force H in Figure 5.2.2). This is achieved for the component icons in CAI which were derived from the formal knowledge in the master TCM vocabulary [V] that standardizes the enterprise terminology to disambiguate communication both inside the enterprise and among the collaborators of the enterprise.

5.4 Experimental Results

Many prototypes were generated for testing with the EOD-ISD approach. The correctness of these prototypes was verified by invited physicians, who actually used the customized D/P system to treat patients. So far, the correctness of the EOD-ISD based customization was meticulous. In section 9.2.2 of Chapter 9, some experimental results for demonstration are presented.

5.5 Recap

To generate accurate WD^2UHI prototypes quickly for experiments, support by the EOD-ISD based software engineering is essential. In the process, the user needs to provide only the MI or iconic specification for the EOD-ISD generator to customize the corresponding D/P system automatically. The correctness of the customized system for the experiments was verified by the invited physicians, who would actually use it to help treat patients. Figure 5.3.5 is the overview of the EOD-ISD approach, which is the enriched version of the skeletal MI paradigm shell proposed by Nong's. It was considered as skeletal/shell, for it had only the following items: i) a usable enterprise ontology for TCM clinical practice; and ii) simply the conceptual argument that a TCM mobile-clinic system could be customized directly from the given specification of icons (iconic specification) using the enterprise ontology as the basis – however, it was devoid of prototypes, verifications and validations. The need to generate credible prototypes quickly for verifying algorithms/methods/concepts proposed in the thesis for mining and discovering useful herbal ingredients and prescriptions instigated my research pursuit of realizing the shell MI argument into a usable methodology. The EOD-ISD approach is the outcome of this pursuit, and it is a very important stage in the course of the WD^2UHI research.

5.6 Conclusion and Connective Statement

In order to support the customization of accurate WD^2UHI prototypes quickly for experiments in the research, the novel EOD-ISD approach is

proposed. In this approach, the user provides the MI/iconic specification for the EOD-ISD generator to produce the target system in a single step. The verifications of the customized D/P systems were carried out by invited physicians, who actually used the prototypes to help treat the patients; this makes the prototypes trustworthy. In fact, the D/P system generated by the EOD-ISD approach also includes other novel elements proposed in this thesis to support useful and meaningful herbal discoveries. Therefore, it is logical to present these elements in the next chapter, under the "living ontology" concept, together with the definitions for *automatic semantic aliasing* and *relevance index*.

5.7 Key References

[Ausi00] Australian Bureau of Statistics, Finance, Australia 2000 Special Article – Information Technology and Telecommunications in Australia, 2000, http://www.abs.gov.au/Ausstats/abs@.nsf/0/9053E0EB512D0DDC4CA256F2A0007346F?Open

[Bardram07] J.E. Bardram and H.B. Christensen, Pervasive Computing Support for Hospitals: An Overview of the Activity-Based Computing Project, IEEE Pervasive Computing, Vol. 6, No. 1, January 2007, 44-51

[Boehm08] B. Boehm, Making a Difference in the Software Century, IEEE Computer Society, Vol. 41, No. 3, March 2008, 32-38

[Cheah07] C. Cheah, Ontological Methodologies - From Open Standards Software Development to Open Standards Organizational

Project Governance, Computer Science and Network Security, Vol. 7, No. 3, March 2007

[Coplien04] J. Coplien, Organizational Patterns: Beyond Technology to People, Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, 2004

[Gruber93a] T.R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, No. 2, 1993, 199-220

[Guarino95] N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, 1995, 25-32

[JWong08b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon**,** An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008

[Katifori07] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis and E. Giannopoulou, Ontology Visualization Methods – A Survey, ACM Surveys, Vol. 39, No. 4, October 2007

[Lewis96] T. Lewis, The Next 10000 Years: Part 1, IEEE Computer Society, Vol. 29, No. 4, 1996, 64-70

[Lin06a] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, Application of Soft Computing Techniques to Adaptive User Buffer Overflow

Control on the Internet, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 36, No. 3, May 2006, 397-410

[Lin08] W.W.K. Lin, J.H.K. Wong and A.K.Y. Wong, Applying Dynamic Buffer Tuning to Help Pervasive Medical Consultation Succeed, Proc. of the 1st International Workshop on Pervasive Digital Healthcare (PerCare), Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications, Hong Kong, 17–21 March 2008, 675-679

[Ng08] S.C.S. Ng and A.K.Y. Wong, RCR – A Novel Model for Effective Computer-Aided TCM (Traditional Chinese Medicine) Learning over the Web, Proceedings of the International Conference on Information Technology in Education (CITE), Wuhan, China, July 2008

[Osterweil08] L.J. Osterweil, C. Ghezzi, J. Kramer and A.L. Wolf, Determining the Impact of Software Engineering Research on Practice, IEEE Computer Society, March 2008, 39-49

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[Standish04] The 3rd Quarter Research Report: Chaos Demographics, The Standish Group International, 2004, http://standishgroup.com/sample_research/ darts_sample.php

[UMLS] http://umls.nlm.nih.gov/

[Uschold07] M. Uschold, M. King, S. Moralee and Y. Zorgios, The Enterprise Entology, Artificial Intelligence Applications Institute, University of Edinburg, UK, 2007, http://citesee.ist.psu.edu/cache/papers/cs/11430/ftp:zSzzSzftp.aiai.ed..ac.ukzSzpubzSzdocumentszSz1998zSz98-ker-ent-ontology.pdf/uschold95enterprise.pdf

[Wong00] A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, Local Compilation: A Novel Paradigm for Multilanguage-Based and Reliable Distributed Computing over the Internet, Special Issue: Mobile and Wireless Communications and Information Processing, Journal of Simulation, Vol. 75, No. 1, July 2000, 18-31

Chapter 6 Living Ontology, Semantic Aliasing and Relevance Index

When a *WD*^{2}*UHI* based prototype is generated, it has the 3-layer architecture (as shown by the customized D/P system in Figure 6.1). The bottom layer of this system is the TCM onto-core, which is customized by the EOD-ISD mechanism from the master or enterprise TCM onto-core for the MI/iconic specification given. The customized TCM onto-core is called the *skeletal* or *intrinsic* onto-core that is not equipped to evolve. If not augmented, this skeletal onto-core is stagnated with the "built-in" knowledge. The STV (Semantic Transitivity Visualizer) is a novel contribution and part of the EOD-ISD mechanism. It lets the user visualize the semantic transitivity of the three layers: query at the top, semantic net in the middle, and TCM onto-core at the bottom.

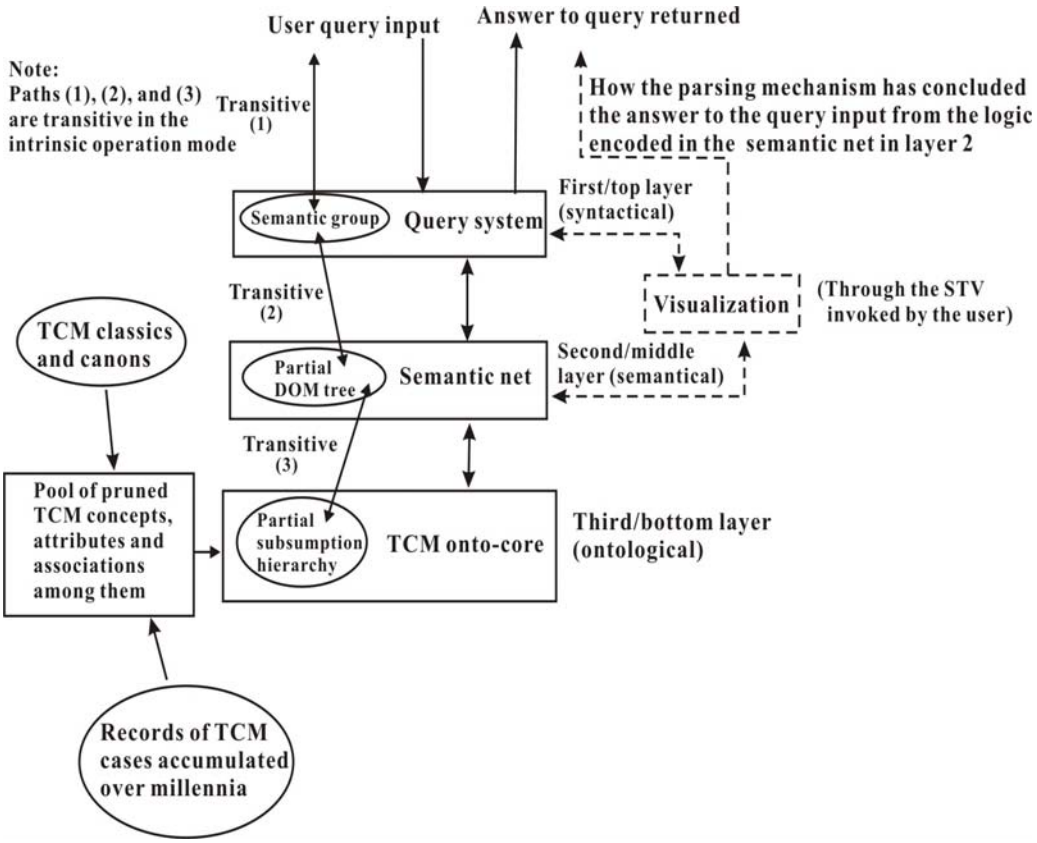


Figure 6.1 Customized D/P system (or *WD²UHI* prototype) - 3-layer architecture (Figure 2.3.1 excerpt)

It is a requirement in this research to pry open the customized intrinsic/skeletal TCM onto-core and let it evolve with time with the help of text mining – a form of data mining [JWong08a]. This is achieved by the OCOE part of the novel OCOE&CID (***O**n-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery*) approach proposed in the thesis [JWong09a] and was verified successfully in the Nong's clinical environment. An open evolvable TCM onto-core is conceptually the combination of "*a closed skeletal TCM onto-core plus the master aliases table (MAT) contents*". **The novel OCOE&CID mechanism, as in Figure 6.2, is a "*sub-concept*" within the EOD-ISD approach**. It automatically invokes (unless manually

inhibited) its text miner WEKA [WEKA] to plough through the open sources that include the open web to look for new scientific findings and cases. It then prunes the findings and adds the result to enrich the MAT contents. This achieves on-line knowledge acquisition, which is essential to the real-time incessant TCM onto-core evolutionary process. The MAT contents are stored in four related special data structures, which can be appended/disconnected to/from the skeletal TCM onto-core manually. As a result, the new knowledge updates in the open evolvable onto-core affect only the MAT contents and do not alter any of the skeletal/intrinsic onto-core information that was customized from the master enterprise onto-core. The MAT contents are necessary for supporting the unique CID part of the OCOE&CID mechanism - the semantic aliasing operation. This operation serves two purposes: i) it standardizes new TCM knowledge, including those handwritten information, with respect to the established standard enterprise vocabulary/lexicon; and ii) it discovers clinical intelligence by two principles, namely, SAME, and logical axiomatic relationship.

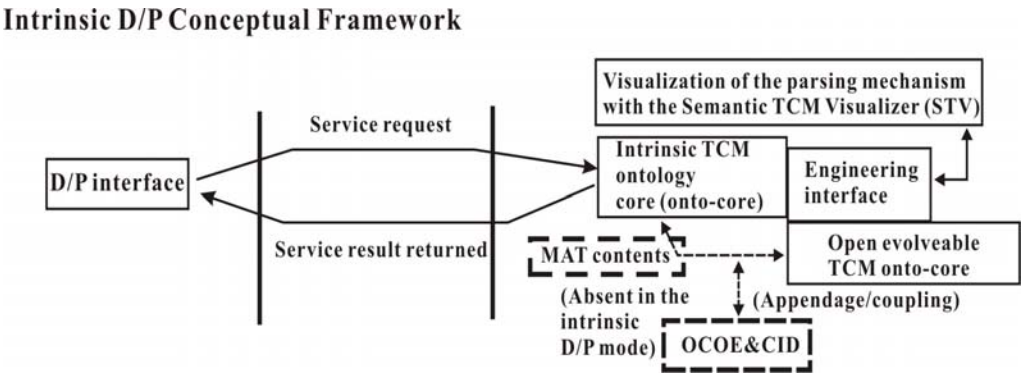


Figure 6.2 Intrinsic diagnosis/prescription (D/P) framework and OCOE & CID

The discussion above becomes clearer if we walk through the details embedded in Figure 6.2 as follows:

a) **Mobile-clinics (MC) based:** Mobile clinics collaborate over the mobile Internet in a pervasive manner. A MC must move into a wireless smart space (SS) before it can communicate with its peers as well as the surrogate server(s) assigned to serve the SS by the telemedicine system (TS) central control. The surrogate lets the physician on board the MC retrieve/verify patient records and collaborate with remote physicians in medical diagnosis. It cooperates with its distributed peers on the same wireline high-speed network in the TS management center.

b) **Essential MC TS elements:** They are summarized in Figure 6.2 as follows:

   i) **D/P interface** (Figure 6.2)**:** The physician treats patients clinically in a computer-aided fashion by interacting with the graphical user interface (GUI), normally by the standard "key-in" operation. The key-in terms are all standard since they already exist in the Nong's enterprise TCM vocabulary {V}.
   ii) Every keyed-in term is verified before it is echoed in the right place. The D/P GUI interface is the top syntactical layer of the D/P system's 3-layer architecture.
   iii) **Intrinsic TCM onto-core:** This is the Nong's proprietary enterprise closed skeletal TCM onto-core that does not evolve

automatically. The following terms are used interchangeably with the same meaning hereafter: intrinsic, closed and skeletal, when referring to the TCM onto-core.

iv) **Semantic TCM Visualizer (STV):** This lets the physician visualize the inference process by the parser in the middle semantic layer of the 3-layer architecture of the Nong's TS. The parser works with the semantic net, which is the machine-processable form of the TCM onto-core below. The shortcoming of the original STV is its lack of ability to verify the existence of semantic transitivity among information items in the three different layers of the TS system architecture – cross-layer referencing. This shortcoming should be resolved through the ***O**n-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery* (OCOE&CID) research. The rationale is that if knowledge items in the three different layers of the 3-layer TS architecture do correspond, their semantic transitivity should exist; giving one item the STV should pinpoint the others consistently.

v) **Engineering interface (EI):** This is where various accessory operations of the Nong's MC D/P system are invoked (e.g. the (Semantic TCM Visualizer (STV)). For example, in the open evolvable Nong's D/P system version, the user can manipulate the OCOE&CID pointers to couple/detach the master aliases table (MAT) contents from the skeletal TCM onto-core. This interface also allows: i) setting up experiments to verify the

OCOE&CID mechanism, and ii) starting/stopping the OCOE&CID text miner WEKA anytime [JWong09a].

6.1 The OCOE&CID Approach

The OCOE&CID (***O**n-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery*) approach, which is a novel conceptual element in the EOD-ISD framework, is aimed at achieving the following objectives:

a) **Real-time continuous and automatic TCM onto-core evolution:** By default a text miner (i.e. WEKA) is running at the background to incessantly plough through the open web for related scientific findings. After pruning, useful findings will be added to the contents of the special OCOE&CID data structures. These additions enrich the open evolvable TCM onto-core, namely, the combination "closed skeletal TCM onto-core + master aliases table (MAT) contents" of the running telemedicine system in a real-time manner. The MAT is the entry point to the special OCOE&CID data structures. The OCOE&CID mechanism only updates the data in the MAT domain and does not alter the skeletal onto-core knowledge at all. In the OCOE&CID conceptual framework this kind of selected evolution is called the "*logical-knowledge-add-on*" technique.

b) **Real-time automatic semantic aliasing**: The skeletal TCM onto-core is built from classical information enshrined in relatively ancient canons,

treatises, and case histories by consensus certification. Since it does not evolve automatically, it risks the danger of stagnating with old knowledge. The OCOE&CID neutralizes this danger by opening up the closed skeletal TCM onto-core with the help of continuous text mining and automatic semantic aliasing (ASA). The ASA weights the similarity between two terms (e.g. Ter_1 and Ter_2; $Ter_1 = Ter_2$ indicates that they are synonyms). For the $P(Ter_1 \cup Ter_2) = P(Ter_1) + P(Ter_2) - P(Ter_1 \cap Ter_2)$ expression, $\cup$ / $\cap$ stands for union/intersection. If Ter_1 and Ter_2 are similar and $Ter_1 \neq Ter_2$ is logically true, they are known as aliases (not synonyms). Then, $P(Ter_1 \cap Ter_2)$ represents the probability or degree of their similarity; $P(Ter_1)$ and $P(Ter_2)$ are probabilities of the multi-representations (other meanings). For example, the English word "errand" has two meanings (multi-representations): "a short journey", and "purpose of a journey". In Figure 6.1.1, Illness (A, a) and Illness (A, b) are aliases because they have only two common defining attributes x_1 and x_3, and their difference includes the set $\{x_2, x_4, x_5, x_6\}$. If the attributes are weighted, the degree of similarity between the two illnesses can be estimated by computing the relevance index (RI) between them. In the OCOE&CID convention one of the illnesses (e.g. Illness (A, a)) is the referential context (RC) to which its alias (e.g. Illness (A, b)) is compared (weighted against); $RI = 0.7$ means that Illness (A, b) is 70% similar to Illness (A, a) and the prescriptions for treating Illness (A, b) therefore has 70% curative efficacy for the RC or

Illness (A, a). If the attributes in Figure 6.1.1 are categorized into: primary attributes (PA) of weight 0.5, secondary attributes (SA) of weight 0.3, tertiary attributes of weight 0.2, and nice-to-know attributes/ones (NKA/O) of weight 0.0, the RI scores of all the aliases (i.e. Illness (A, b), Illness (A, c) and Illness (X, x) for the Illness (A, a) chosen as the RC) can be computed (see Table 6.1.1).

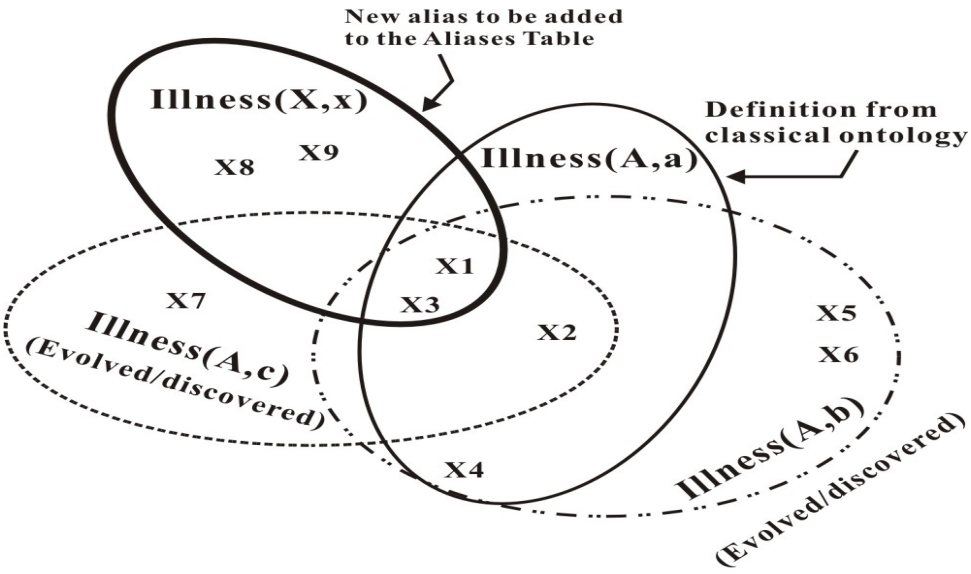


Figure 6.1.1 Illness (X, x) is a new alias for the referential context (RC) Illness (A, a)

Illness / context	Attributes, corpus or alias's set	Attribute classes (for the referential context)	RI scores, { 50% - PA, 30% -SA, 20%- TA, 0%- NKA}	Remarks
Illness (A, a); **Common Cold**	*corpus* : $\{x_1, x_2, x_3, x_4, x_8, x_9\}$	PA - $\{x_1, x_2\}$, SA- $\{x_3\}$, TA - $\{x_4\}$, NKA - $\{x_8, x_9\}$	$RI = \frac{0.5}{2}(1+1) + 0.3(1) + 0.2(1) + \frac{0}{2}(1+1) = 1$	**Referential context (RC)**
Illness (A, b)	*alias's set:* $\{x_1, x_2, x_3, x_4, x_5, x_6\}$	PA - $\{x_1, x_2\}$, SA- $\{x_3\}$, TA - $\{x_4\}$, NKA - $\{x_5, x_6\}$	$RI = \frac{0.5}{2}(1+1) + 0.3(1) + 0.2(1) + \frac{0}{2}(1+1) = 1$	**Alias of 100% relevance**
Illness (A, c)	*alias's set:* $\{x_1, x_2, x_3, x_7\}$	PA - $\{x_1, x_2\}$, SA- $\{x_3\}$, NKA - $\{x_7\}$	$RI = \frac{0.5}{2}(1+1) + 0.3(1) + 0.2(0) + \frac{0}{2}(1) = 0.8$	**Alias of 80% relevance**
Illness (X, x)	*alias's set:* $\{x_1, x_2, x_8, x_9\}$	PA - $\{x_1, x_2\}$, NKA - $\{x_8, x_9\}$	$RI = \frac{0.5}{2}(1+1) + 0.3(0) + 0.2(0) + \frac{0}{2}(0) = 0.5$	**Alias of 50% relevance**

Table 6.1.1 Computations of aliases' RI scores for the referential context (RC) Illness (A, a) in Figure 6.1.1

c) **Special elements to support real-time ontology evolution:** They include the MAT (master aliases table) and the OCOE&CID mechanism. The MAT is the directory through which the special data structures that catalyze continuous TCM onto-core evolution can be managed. Via the Engineering Interface (Figure 1b) the user can disconnect MAT from the skeletal TCM onto-core to revert the clinical practice back to the intrinsic mode (i.e. with original closed onto-core). The disconnection does not stop the evolution of the OCOE&CID contents in the special data structures, however, unless the text miner is stopped. When the OCOE&CID mechanism is invoked, its special data structures will be first initialized with the knowledge in the skeletal TCM onto-core, and then the contents in these data structures (called the "foundation") grow

with evolution. Every referential context (RC) (which is identified/defined by the domain experts) in the TCM onto-core associates with a set of special OCOE&CID data structures (Figure 6.1.2):

i) *Contextual Aliases Table (CAT)*: This contains all the aliases (e.g. Pneumonia) for the RC (e.g. Pneumonia and Influenza for the Common Cold RC in Figure 6.1.2).

ii) *Contextual Attributes Vector (CAV)*: This contains all the classically enshrined, canonical attribubes for the RC. For example, for the **Illness (A, a)** RC, its canonical attributes is the set $\{x_1, x_2, x_3, x_4\}$.

iii) *Relevance Indices Table (RIT)*: This lists all the RI scores of the aliases for the RC – one RI for every alias. For example, in Figure 6.1.2, Pneumonia is 70% similar to Common Cold (for those weights assigned to PA, SA, TA, and NKA). How RI scores are computed is illustrated in Table 6.1.1.

iv) *Possible Prescriptions Table (PPT)*: This lists the enshrined prescriptions for every alias in CAT. The weightings of the prescriptions are the same as the RI computed for the alias; the RI for RC prescriptions is 1 (i.e. 100%).

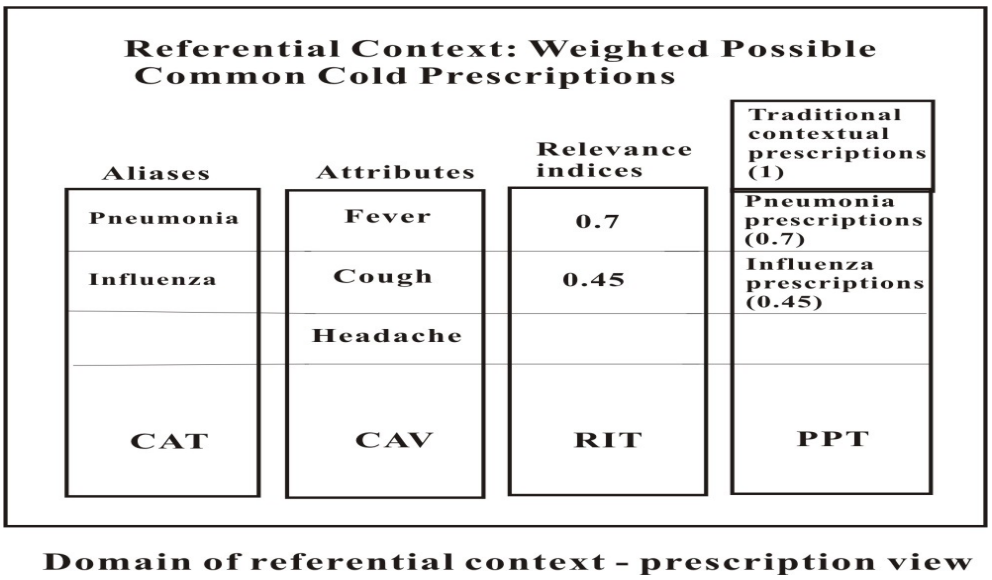


Figure 6.1.2 A set of our catalytic data structures for the referential context (RC) Common Cold

d) **Text mining and RI computation:** Text mining is the driving force behind the continuous evolution of the open TCM onto-core, namely, the combination: "closed skeletal TCM onto-core + MAT contents". MAT contents are the totality of those in CAT, CAV, RIT, and PPT together. The text miner WEKA finds new contexts/names (e.g. such as Illness (X, x)) and extracts their important attributes (e.g. $\{x_1, x_2, x_8, x_9\}$). After pruning the new findings by: i) constraining the context against the standard enterprise vocabulary {V}; and ii) computing context RI for the RC (similar to Table 6.1.1), they will be added to the MAT contents as part of the continuous evolutionary process. In the OCOE&CID framework RI computation is based on the adapted WEKA parameters. Using Figure 6.1.2 as an example, the Common Cold RC has Pneumonia and Influenza as aliases. If Pneumonia was the i^{th} alias to Common Cold, its RI value

is $RI_i = SW_i = \frac{SAC}{\{MAS \in \{V\}\}}$. The SW_i value (adapted form WEKA) is "size of the attribute corpus (SAC) of the Common Cold context over the mined attribute set (MAS) for Pneumonia that overlaps; MAS is constrained by the enterprise vocabulary {V}". Assuming **Illness (A, a)** in Figure 6.1.1 was Common Cold of four standard PA, x_1, x_2, x_3 and x_4, and **Illness (A, c)** was Pneumonia defined by the PA, x_1, x_2 and x_3, the RI for Pneumonia is $RI_i = WA_i = 3/3$.

e) **Clinical intelligence discovery with the "SAME - 同病異治, 異病同治" principle** [WHO07]: Automatic semantic aliasing standardizes any text-mined TCM term and absorbs it into the open TCM onto-core. The SAME principle finds new prescriptions for treating the RC. If PAa, PAb, PAc and PXx respectively were prescriptions for the illnesses in Figure 6.1.1 (i.e. **Illness (A, a)**, **Illness (A, b)**, **Illness (A, c)**, and **Illness (X, x)**), the total set of prescriptions for treating the **Illness (A, a)** RC is $P_{A,a}^{total} = PAa \cup PAb \cup PAc \cup PXx$, where $\cup$ is union. The prescription efficacies in $P_{A,a}^{total}$ are rated by the RI scores; for example, conceptually PXx has 25% efficacy of treating **Illness (A, a)** (see Table 5.1).

f) **Semantic transitivity:** This supports cross-referencing across the three layers of the telemedicine system (Figure 5.1) in the STV visualization

process. The original Nong's STV does not have this "cross-layer" cross-referencing capability.

6.2 Experimental Results

The OCOE&CID (***O**n-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery*) approach, which is an constituent element within the EOD-ISD conceptual framework, is aimed at achieving the following objectives: i) real-time continuous and automatic TCM onto-core evolution; ii) real-time automatic semantic aliasing; iii) real-time ontology evolution support using special elements; iv) text mining and RI computation; and v) intelligence discovery with the "SAME - 同病異治, 異病同治" principle. The experimental results demonstrate how these objectives are achieved in the section *9.4.1 The OCOE&CID Approach* of Chapter 9.

6.3 Recap

The OCOE&CID approach is within the conceptual EOD-ISD framework, which enables the following: i) a trustworthy prototype can be generated for experiment; ii) this prototype is automatically customized in a single step from the named master/enterprise TCM onto-core for the MI specification given; and iii) the customized prototype is equipped with the following capabilities: ontology evolution, automatic semantic aliasing, text mining and relevance index (RI) computation. Then, herbal discoveries are achieved by the "SAME" principle and logical axiomatic relationship. The

OCOE&CID shows how herbal ingredients can be discovered through the use of the novel automatic semantic aliasing mechanism. These discoveries are considered Type 1. From the angle of an enterprise, any entity, which has never existed (Type 1) before in its operational ontology or dictionary, it is a discovery. For example, if set P is not part of the knowledge base K (i.e. $P \notin K$) but it does possess all the attributes of one of the pre-defined classes in *CL*, for $CL \notin K$, then the logical phenomenon, "$< P \in CL > \wedge < P \notin K >$", where $\wedge$ for logical "AND", indicates ***P as a new occurrence*** with respect to the knowledge base K. Thus, P is a Type 1 discovery (*it differs from the Type 2 discovery, which focuses on unfolding new ways for any extant element in the current knowledge base being applied elsewhere*).

6.4 Conclusion and Connective Statement

The novel EOD-ISD framework proposed in this thesis indeed provides support for achieving the two main research objectives: i) development of a usable WD^2UHI platform, and ii) discovery of useful herbal ingredients. Yet, the automatic semantic aliasing mechanism is intrinsically algorithmic, and it is therefore tedious to program it accurately. After some careful exploration, I propose a faster and more trustworthy AI approach for knowledge classification and discovery - the backpropagation neural network (NN). The details for this NN proposal will be presented in the next chapter.

6.5 Key References

[Bloehdorn05] S. Bloehdorn, P. Cimiano, A. Hotho and S. Staab, An Ontology-based Framework for Text Mining, LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, May 2005, 87-112

[Holzman03] L.E. Holzman, T.A. Fisher, L.M. Galitsky, A. Kontostathis and W.M. Pottenger, A Software Infrastructure for Research in Textual Data Mining, The International Journal on Artificial Intelligence Tools, Vol. 14, No. 4, 2004, 829-849

[JWong08a] J.H.K. Wong, T.S. Dillon, A.K.Y. Wong and W.W.K. Lin, Text Mining for Real-time Ontology Evolution, Data Mining for Business Applications, Springer, 2008, ISBN: 978-0-387-79419-8, 143-150

[JWong09a] J.H.K. Wong, A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), 2009

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[WEKA] http://www.cs.waikato.ac.nz/ml/weka/

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

Chapter 7 Knowledge Classification for Herbal Discovery

7.1 Introduction

In the original Nong's TCM (Traditional Chinese Medicine) diagnosis/prescription (D/P) telemedicine system for mobile-clinic (MC) applications, there is an underlying database containing raw clinical data. Figure 7.1.1 shows a UML (Unified Modeling Language) conceptual organization of the raw clinical data. UML is suitable for outlining the structure of the ontology [Kogut02]. In this UML example, there are only ten clinical entities uniquely identified. In the set {0/咳嗽/Cough} the unique symbols, "0", "咳嗽" and "Cough" have the same connotation - "0" as the identifier. Solid-line arcs (e.g. between "3" and "4") indicate between entities, they are logically transitive. The entities can be placed anywhere within the database but their retrievals depends on the "***retrieval algorithm*** (RA)" implemented as part of the D/P software system.

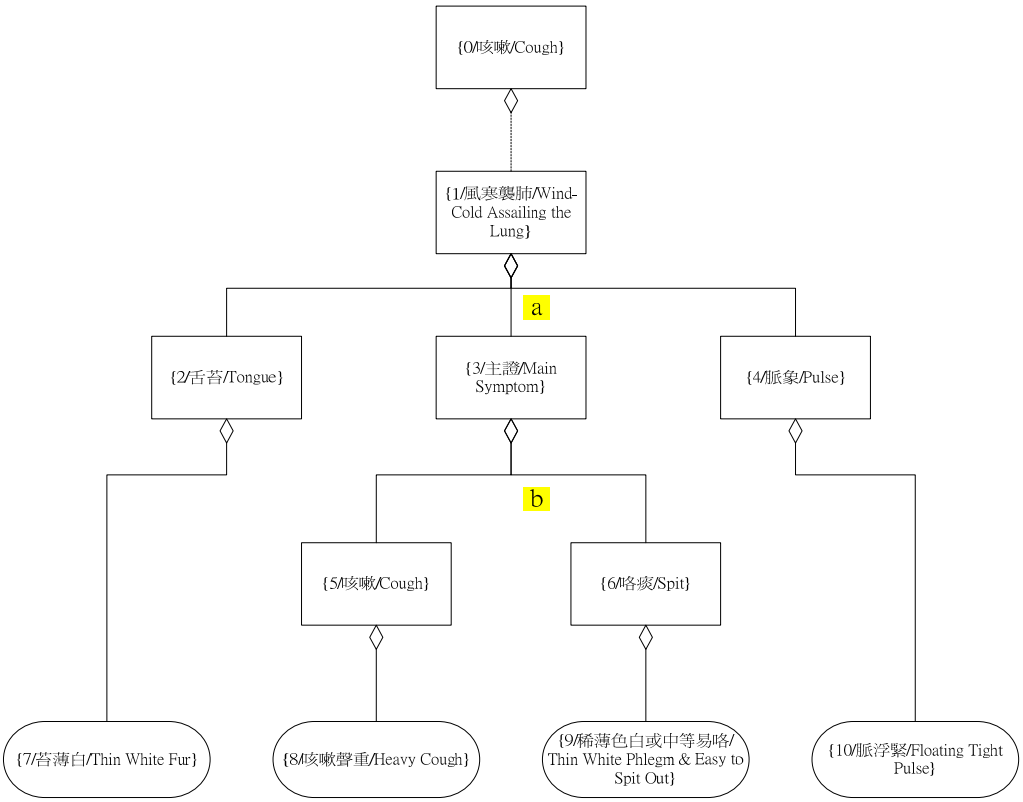


Figure 7.1.1 UML organization of raw clinical data in the original Nong's D/P system ("0" and "5" are logically different - "0" is an illness name and "5" a symptom) (excerpt of Figure 2.7.1)

In light of UML, the aggregation relationship in Figure 7.1.1 can also be construed as a "*part-of*" relationship, indicating that a "whole" object consists of "*part*" objects [Dillon93]. This kind of relationship also exists commonly in XML documents. For example, {1/風寒襲肺/Wind-Cold Assailing the Lung} is part-of {0/咳嗽/Cough}, and {8/咳嗽聲重/Heavy Cough} is part-of {5/咳嗽/Cough}. The "*part-of*" relationship is for high-level semantic visualization rather than logical implementation.

The problem with the UML representation [Cranefield01, Kogut02] is a lack of sufficient information for correct implementation. Some UML parts (e.g.

logical points *a* and *b* in Figure 7.1.1) depend on the interpretation of the implementers (e.g. *a* can be AND, OR or EXCLUSIVE OR). This would lead to multi-representations of one concept and many incompatible implementations and eventual system failures. Thus, a data-oriented system is preferred, and this would allow the points *a* and *b* to converge to their true meaning with respect to the data set. In fact, this is one advantage offered by the neural network (NN) approach. The convergence of the points *a* and *b* is achieved by training with a given dataset. This convergence is the result of classification by artificial intelligence (AI), namely, NN. Therefore in this chapter, I aim to establish that the classification using NN indeed provides a solid basis for herbal discoveries.

As a matter of fact, the NN approach provides the following advantages:

a) It ***converges*** naturally to the logical requirement, which is defining the true meaning for logical points such as *a* and *b* in Figure 7.1.1, with respect to the training dataset.
b) The NN is a ***simple***, ***generic***, ***reusable***, ***less error-prone*** API, which can be invoked anytime, anywhere to be trained with the given set of data so that it can ***predict*** the outcome from new data. The outcome can be different from one application to another (e.g. in one case it is illness and in another it is herbal ingredient).
c) The API approach enhances the **software reliability** because it reduces the amount of possible errors that could be introduced inadvertently in the programming process tremendously.

d) The NN performance in light of speed can be improved by ***automatic real-time logical pruning***, in which the computation excludes those unimportant NN arcs [Lin04].

In Figure 7.1.1 the elements 0 and 5 are synonyms. The *retrieval algorithm* or RA_1 is working by predicate logic; for example: (i) if "8" is true then "5" is true; (ii) if "9" is true then "6" is true; and (iii) if "5" and "6" are true then "3" is true (i.e. the logic for point "***b***" is a logical "AND" function). For the same Figure 7.1.1, if another RA, namely RA_2, interprets point "***b***" as a logical OR (i.e. if "5" or "6" is true then "3" is true), then RA_1 and RA_2 are logically incompatible (i.e. $RA_1 \neq RA_2$). In our research, if one has to scrutinize the RA code in order to find out the exact meaning of "b", then the RA code has ***implicit*** semantics. Diagrammatically Figure 7.1.1 does not differentiate the exact semantics between RA_1 and RA_2 - the two systems look similar superficially. In the same light, the implementation for the logical point "a" may vary from one system to another – incompatible variants. Only when the predicates for the different logical points are axiomatically defined formally in a variant that we can say whether two modules are actually clones.

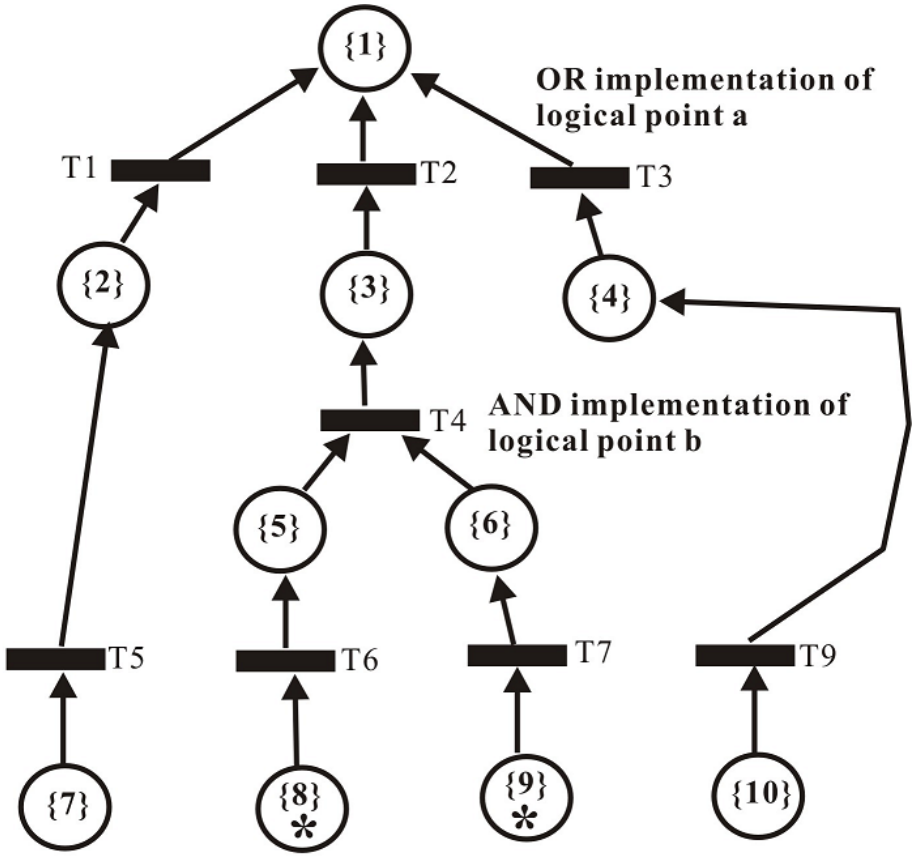


Figure 7.1.2 Formal Petri net representation of Figure 7.1.1

Figure 7.1.2 is the formal Petri net representation of the "part of" hierarchy in Figure 7.1.1, where logical point "a" (T2 should be fired) and "b" (T4 should be fired) have the OR and AND implementations respectively. In fact, this Petri net can be used to simulate/check/specify the logical correctness of Figure 7.1.1. This Petri net (PN) is bipartite and is made up of three sets of symbols arcs (A), transitions (T), and places (P) as follows:

a) PN = [P, T, A]

b) P = [{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}]

c) T = [T1, T2, T3, T4, T5, T6, T7, T8, T9]

T1, T2 and T3 represent a logical OR operation and T4 is the logical AND operation. Places {8} and {9} are initialized with two tokens (i.e. "*"). When a token is in a place then the place assumes the "true" state. T4 fires as long as both places {5} and {6} have token(s). The transition firing is atomic, and this means that at anytime only one transition can fire even it takes zero time to complete. In Figure 7.1.2, with the tokens in places {8} and {9}, the successive firing will generate the final token in place {3}, which represents the intermediate logical conclusion. In fact, the traversals starting from the places {8} and {9}, through the places {5}, {6} and {3}, finally for the token to reach place {1} represents a "***parsing***" process. The traversals can be depicted clearly by the reachability graph in Figure 7.1.3. In this graph the state vector changes with time when the transitions fire in an in-deterministic fashion. Every vector (e.g. M0) indicates the current state at the time, with a "1" to indicate the presence of a token – a "logically true" state. The leftmost bit position in the vector is for the place {1} and the rightmost for the place {10}.

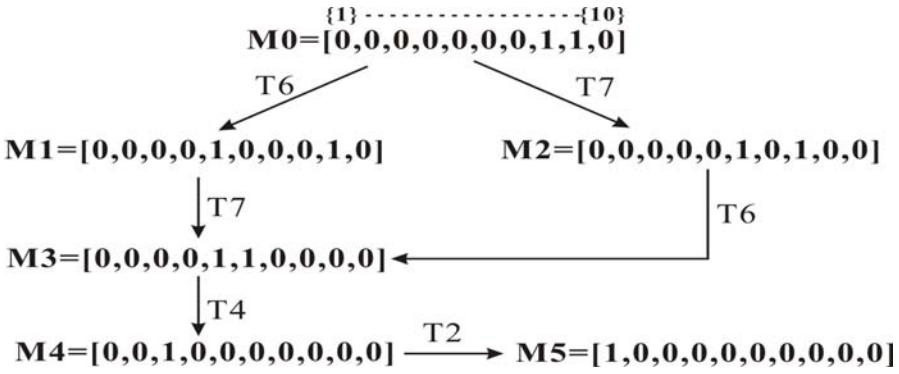


Figure 7.1.3 Reachability graph for the Petri net in Figure 7.1.2

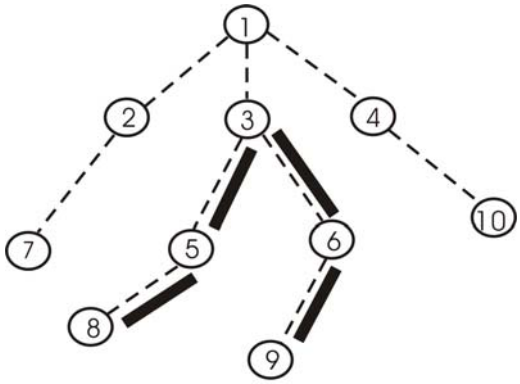
If we reverse the logical operations for the point "a" (i.e. to become logical AND) and "b" (i.e. to become logical OR), then the corresponding Petri

net and its reachability graph will be very different. Therefore, logically speaking Figure 6.2 differs from the Petri net with the reversed (i.e. "a" becomes "AND" and "b" becomes "OR") variant. Both variants, which are not clones, however, are still formal and axiomatic, but not logically compatible.

7.2 Ontology Viewpoint

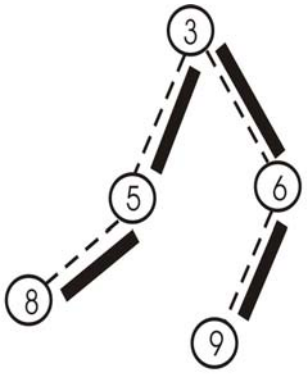
In the last section it was pointed out that the same high-level hierarchy of entities and their associations can be informal (when logical points "a" and "b" are undefined" or formal if the logical points are axiomatically defined. In the latter case the formalism can be inspected by transforming it into the preferred representation such as the Petri net. Yet from the ontological point of view, two logically incompatible variants, which consists of the same set of entities but somewhat different associations, can be regarded as two different sub-ontologies explicitly isolated from the same "larger ontology construct". If Figure 7.2.1 represents the larger ontology, the Figure 7.2.2 is a subontology isolated from it. In these diagrams the nodes/entities are uniquely identified circles; the dotted-line arcs are associations/relationships among the entities, and the bold lines are the arcs for the isolated subontology. In Guarino's view [Guarino95], Figure 7.2.1 is the subsumption hierarchy of the larger ontology, and the Figure 7.2.2 is the subsumption hierarchy of the smaller isolated sub-ontology. If Figure 7.2.2 is isolated for a company or establishment to fulfill a specific operation purposes (e.g. the Nong's mobile-clinic TCM telemedicine D/P system), then is an enterprise ontology. Inside the ontology there are many possible operation paths. For example, the set of operation paths for Figure

7.2.1 include: “1,2”; “1,3”; “1,4” ; “1,2,7”; “1,3,5”; “1,3,6”; “1,3,5,8”; “1,3,6,9”; and “1,4,10”. The traversal of an operation path is event-driven, and therefore a stochastic process. From the ontology point of view Figure 7.2.1 is a semantic net or DOM (document object model) tree that represents the entire ontology for machine processing. In this light every operation path is also a semantic path, which is the outcome by the parsing mechanism for the given set of input parameters.



A complete tree

Figure 7.2.1 A larger ontology



A partial tree partitioned from the complete tree

Figure 7.2.2 An isolated subontology

In contrast to algorithmic programming, the final software system contains the logical predicates an implicit manner. That is, other than the programmer others cannot decipher the exact semantics correctly. In contrast, the ontology-based approach is completely explicit, and as a result anybody (manager, programmers and users) can clearly visualize a function, for it is simply the realization of a unique semantic path in the ontology. Therefore, in ontology-based software engineering, the first logical step is building the “ontology basis” by consensus certification. In this research three levels of ontology constructs are differentiated:

a) *Global* – This contains the total formal knowledge of a domain. For example, the global TCM ontology is made up of all the classics, treatises and case histories accumulated over a few thousand years.
b) *Enterprise* – This is a subset of the global ontology to suit the purpose of a company, enterprise and/or establishment. The current Nong's proprietary TCM ontology for clinical practice is a "local" (to Nong's) standard/vocabulary with the aim to support correct communications and interoperability within the company and with collaborating partners.
c) *Local of the local* – Nong's creates different D/P variants with respect to the specifications of the clients. In the creation process every customized target system would have only a different subset of the Nong's enterprise ontology for its in-situ operation. These systems would be interoperable to a varying degree depending on the contents of the local TCM ontology.

The advantage of the ontology-based approach is that there is less chance of error for the target TCM D/P system to be customized from the Nong's enterprise TCM ontology. This is derived from the fact that every function in the target system can be verified against a specific semantic path, with the concept of semantic transitivity in mind.

To recap, in this research all the D/P prototypes for the experiments in the Nong's clinical environment were customized from the Nong's enterprise TCM ontology core (onto-core). To customize a target system only the corresponding meta-interface (MI) needs to be built. The MI is a collection of

reusable icons selected from the standard icon library. With MI as the input, the EOD-ISD (***Enterprise-Ontology-Driven Information System Development***) mechanism generates the target or customized system automatically. The TCM onto-core of the target system is a subset of the Nong's enterprise onto-core. The target system has a 3-layer architecture: i) the bottom local ontology as the operation standard; ii) the middle semantic net that exactly represents the bottom ontology for machine processing (by the parser to be exact); and iii) the top GUI where the user and input parameters for the parser to draw the logical conclusion from the DOM tree. The three layers conceptually are semantically transitive and logically the same. Yet, the parsing mechanism is algorithmic and works with predicate logic as shown by Figure 7.1.2.

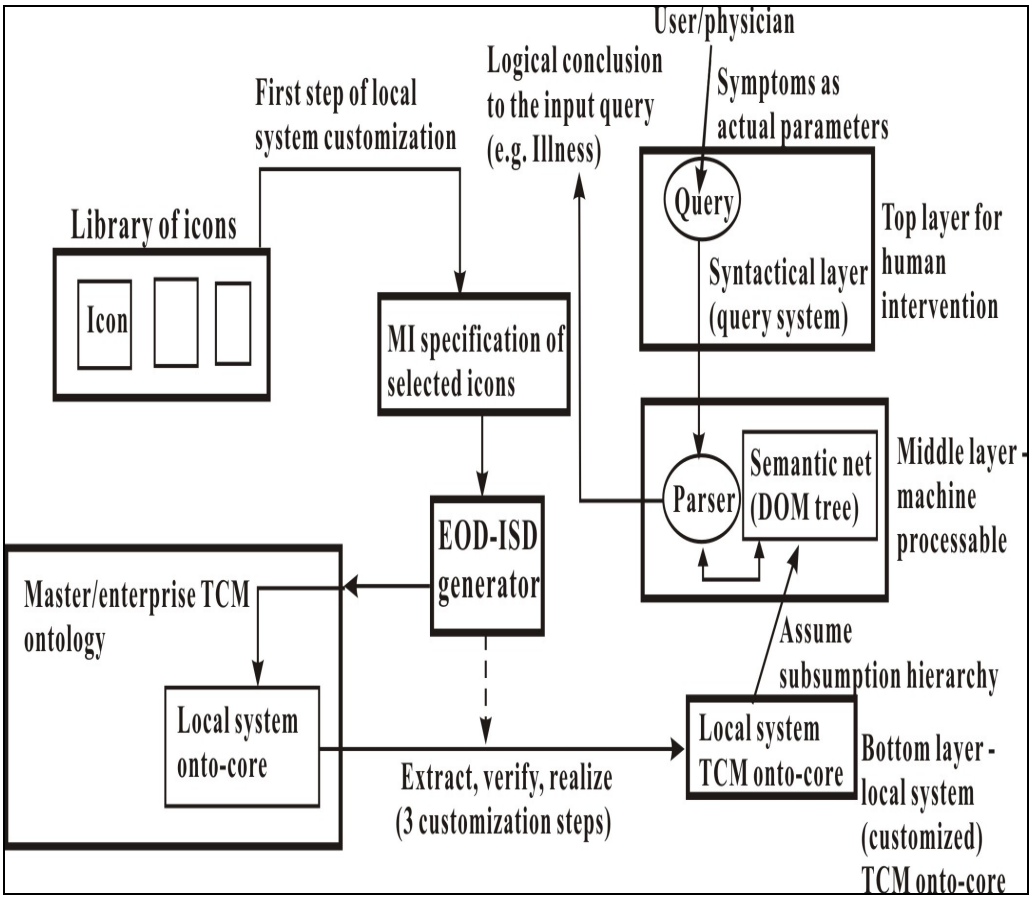


Figure 7.2.3 Local/target system customization flow (excerpt of Figure 5.3.5)

7.3. Shortcoming of the Algorithmic Approach

From the predicate logic point of view, the parser may work with Figure 7.1.2 in the following algorithm:

T6 →{5}; /* if place {8} is true then T6 fires and puts a token in {5} */

T7 →{6};

T4 →{3};

T2 →{1} /* end */

The problem in the above approach is the interpretation of a "***true***" state. The presence of a token implies a binary "**1**" state for a place and its absence implies "**0**". Having a toggle state as such is not appropriate for clinical application because it is difficult to specify the degree of importance (or weight) for a parameter. To explicate this point let us assume that the place {8} in Figure 7.1.2 represents one symptom and place {9} another. For regional and/or epidemiological reasons, the contributions of places {8} and {9} are respectively 0.3 and 0.7 in Hong Kong, but it should be 0.5 for both in Canada. *In order to accommodate the use of weighted parameters naturally and in an adaptive manner, it has become a necessity to find a suitable alternative to replace the deterministic algorithmic approach.*

After a thorough search, the backpropagation neural network (NN) and supervised learning [Bishop95, Sarle97] are seemingly an equitable alternative for the following reasons:

a) The inputs can be weighted according to their degree of significance to the target result.
b) The teacher signal is the reference/target/class with which the NN can learn (or be trained) to attain. If the sample size for training is large enough, then the NN produces the same result as the algorithmic approach.
c) It is cost effective in the sense that the same NN architecture can be trained to satisfy different purposes, with respect to the given teacher signal in each case – that is, the architecture is reusable.
d) The size of the NN architecture can be controlled because the only mandatory architectural constraint is that the number of neurons in the hidden layer must be greater than or equal to the number of neurons in the input layer.
e) The NN architecture can be logically pruned on the fly to suit real-time applications [Lin04].

7.4 Neural Network by Backpropagation as an Alternative

Figure 7.4.1 shows how a given subsumption hierarchy is mapped into the generic neural network (NN) by backpropagation. This NN architecture has three layers: i) a number of input neurons (to receive the input data) or N_i; ii)

the hidden layers of N_h neurons; and iii) one output neuron or N_o. The usual operation condition is $N_h \geq N_i$. The output neuron in this case represents the illness, Cold, or 感冒. From another angle the mapped backpropagation NN has become a classifier because with the symptoms received by the input neurons Cold is the intelligent conclusion by inference.

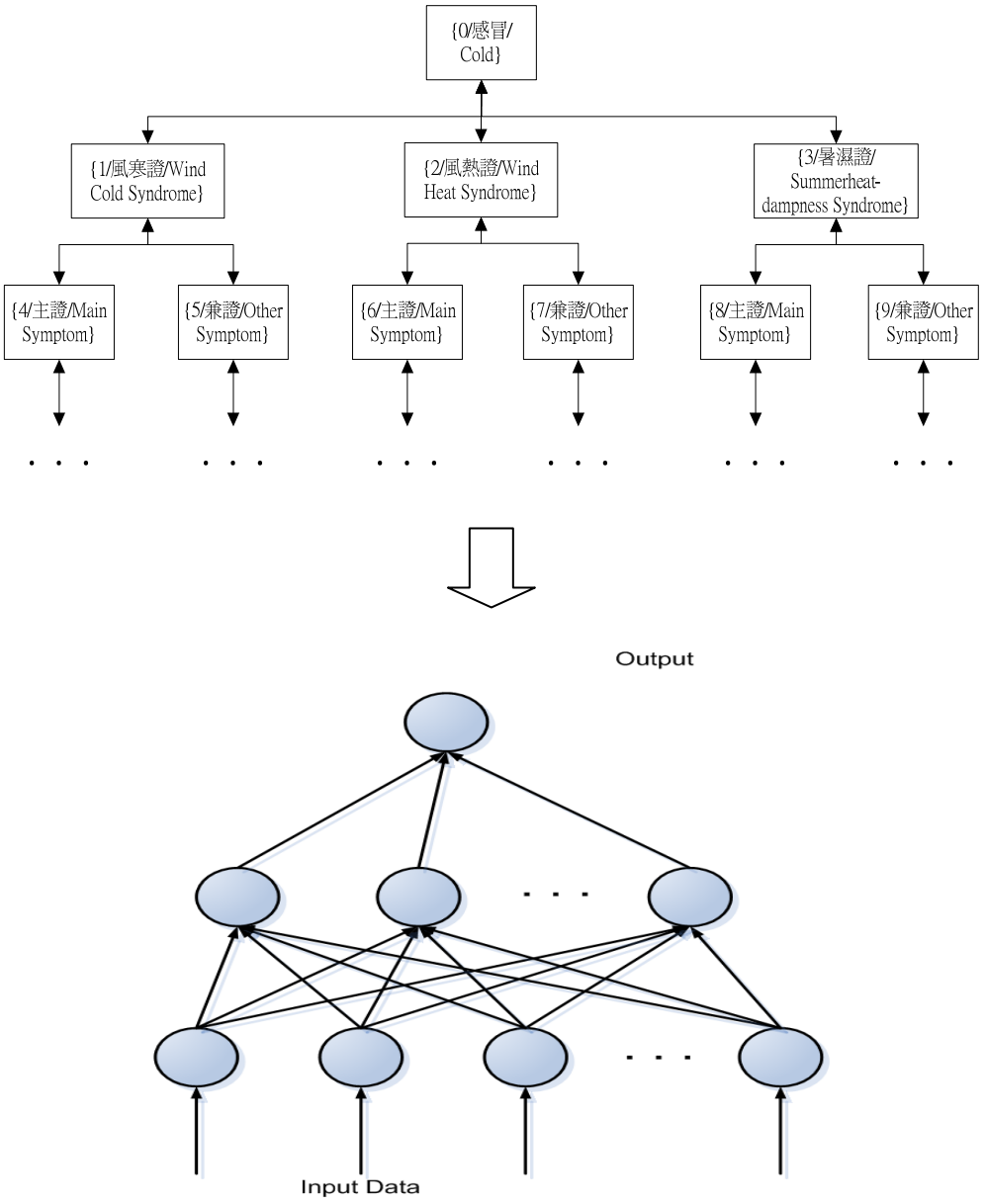


Figure 7.4.1 Mapping subsumption hierarchy and backpropagation NN

The ***first*** obvious, significant advantage from the artificial intelligence (AI) approach, in terms of the backpropagation NN architecture, is that *it obviates the necessary axiomatic definitions of the logical points in the subsumption hierarchy (e.g. "a" and "b" in Figure 7.1.1) as in the algorithmic approach*. Conceptually such axiomatic definitions converge intrinsically according to the set of training samples used; in effect, the NN learns and determines what is equitable for an axiomatic definition. The convergence associates with the weights of the connecting arcs (among the neurons) in the NN. Unless it is specified, the backpropagation NN architecture in our research is "fully-connected". If the training set of samples represents a variable (v) population (P) or v^P, then the set/matrix of trained NN arc weights w^S would be "perfect" (i.e. $w^S = w^P$; the superscript *S* indicates the set/matrix). Being "perfect" means that any selected number of variables as input would consistently yield the perfect/expected answer (e.g. 感冒 (Cold)). That is, the "perfect" NN works like a deterministic algorithm. If the variable set for training the NN is limited in size v^T (T for training), then $v^P >> v^T$ would produce an imperfect w^S. Working with an imperfect w^S, any variable not included in the training set would lead to an imperfect answer; the imperfection is indicated by the percentage error ε. Conceptually we accept the inferred answer for $\varepsilon \le \zeta$, where ζ is the maximum percentage error to be tolerated.

From the discussion above, we can easily point out that the ***second*** significant advantage from the backpropagation NN approach is that one generic NN architecture can be used for different classification purposes because the core idea is to associate whichever input set is used with the

designated output/teacher signal. We can generalize the backpropagation NN output in Equation 7.4.1 by letting; i) O as NN output; ii) x^m as the set/matrix/vector of input variables of size m; iii) w^S as the set of NN arc weights obtained by training with the training set v^T. If the NN were trained with the population of variables, its output would be consistently perfect, as shown by Equation 7.4.2 (i.e. the same as the deterministic algorithmic approach).

$$O = f\{x^m, w^S\} \ldots\ldots Equation 7.4.1$$

$$O^{perfect} = f\{x^m, w^P\} \ldots\ldots Equation 7.4.2$$

If we scrutinize the operation of a neuron (or node j) in the hidden layer, we would obtain a picture similar to Figure 7.4.2. The input set to this node is $x^m = x^n = (x_0, x_1, \ldots, x_n)$; x_0 is the bias/offset to give the input set flexibility. All the input variables at the j^{th} level (for node j) are weighted (e.g. x_1 by w_{1j} and x_2 by w_{2j} etc.). The collective effect of the x^n input vector is obtained by $u_j = x_0 + \sum_{i=1}^{n} x_i w_{ij}$. The output o_j from node j is a function of u_j as in Equation 7.4.3; this <u>*activation function*</u> can take any form (e.g. sigmoid function $f(x) = \frac{1}{(1+e^{-x})}$ and the hyperbolic tangent function $f(x) = \tanh(x)$ [Yann98]). In our research the non-linear $f(x) = \frac{1}{(1+e^{-x})}$ sigmoid function is adopted to

deal with the NN non-linear property. Figure 7.4.3 is the sigmoid function, which monotonically asymptotes for $x \to \pm\infty$.

$o_j = f\{u_j\}$.....*Equation*7.4.3

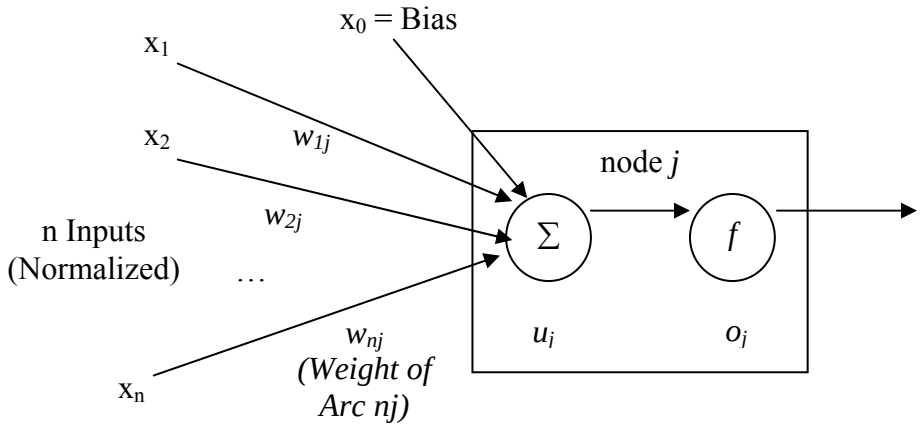


Figure 7.4.2 Input, weights and output of a neuron (node j) in the hidden layer

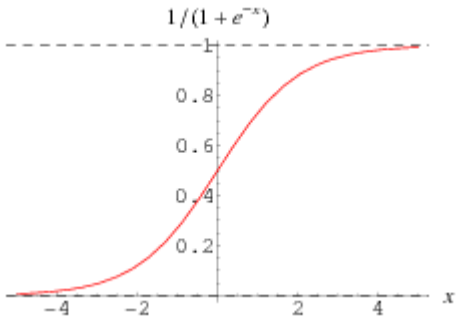


Figure 7.4.3 Sigmoid function

7.4.1 Training/Learning

Backpropagation, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. The term, an abbreviation of "***backward propagation of errors***", was first described by Paul Werbos in 1974 and gained recognition until 1986. It is a supervised learning method, and is an implementation for the delta rule (a gradient descent learning rule for updating the weights of the artificial neurons in a single-layer perceptron). It requires a teacher that knows, or can calculate, the desired output for any given input, and is most useful for feed-forward networks (networks that have no feedback, or simply, connections that do not loop back). Backpropagation requires that the activation function used by the artificial neurons is differentiable.

The following is an actual algorithm for a 3-layer network (only one hidden layer):

Initialize the weights in the network (often in random)

Do

For each example e in the training set

O = neural-net-output (network, e); forward pass

T = teacher output for e

Calculate error (T - 0) at the output units

Compute delta_wi for all weights from hidden layer to output layer; backward pass

Complete delta_wi for all weights from input layer to hidden layer; backward pass continued

Update the weights in the network

Until all examples classified correctly or stopping criterion satisfied

Return to the network

The equation for the weight update of back propagation is as follows:

$$\Delta w_{jk}(t+1) = \gamma \delta_k y_j + \alpha \Delta w_{jk}(t)$$

where:

j = The j^{th} neuron of the $(p-1)^{th}$ layer

k = The k^{th} neuron of the p^{th} layer

t = The t^{th} epochs

δ_k = The error signal at the k^{th} neuron = (d-y)y(1-y)

where d=desired output and y=actual output

y_j = The input of the j^{th} neuron

α = Learning rate

γ = Momentum coefficient

In light of our research, the main advantage of backpropagation over traditional methods of error minimization is that it reduces the cost of computing derivatives by a factor of N, where N is the number of derivatives to be calculated. Furthermore, it allows higher degrees of non-linearity and precision to be applied to problems.

7.4.2 Probability Theory and Normalization

Probability theory is the branch of mathematics concerned with the analysis of random phenomena. It is essential to many human activities that involve quantitative analysis of large sets of data. The definition of the discrete probability distribution starts with a set called the sample space, which relates to the set of all possible outcomes in classical sense, denoted by $\Omega = \{x_1, x_2, ...\}$. It is then assumed that for each element $x \in \Omega$, an intrinsic probability value *f(x)* is attached, which satisfies the following properties:

1. $f(x) \in [0,1]$ for all $x \in \Omega$
2. $\sum_{x \in \Omega} f(x) = 1$

This means that the probability function *f(x)* lies between zero and one for every value of *x* in the sample space Ω, and the sum of *f(x)* over all values *x* in the sample space Ω is equal to 1. Therefore, the probability of the entire sample space is 1, and the probability of the null event is 0.

In our experiments, all input values are normalized before feeding into the network for training so that it is valid under the probability theory.

7.4.3 Reasons for Choosing NN Backpropagation

The reasons for choosing NN backpropagation are as follows:

i) It is a generic approach, and once it is implemented, it can be applied to other areas of the research project. Unlike the algorithmic approach, the same NN can be trained to have logical convergence. For example, the definition of point "b" in Figure 6.1 can be a logical AND or OR operation depending on the set of training data. This reduces the amount of programming effort as well as the number of inadvertent errors of human origin introduced in the programming process. In fact, the same NN architecture can be reused for different sets of inputs with different applications and outcomes in mind.

ii) The input can be normalized, which can satisfy the epidemiology requirements (by setting different tolerance level of the RMSE, e.g. RMSE < 0.1.).

iii) The relative importance of any input parameters can be easily specified by the assigned weight.

iv) Dynamic NN pruning is possible so that the execution time can be reduced for time-critical cases. The pruning process does not affect the accuracy of the result, which should stay within a tolerance band [Lin04]. Besides, computation time speedup can be gained easily by parallel executions of the different NN modules working on different input sets.

v) It has a long history and thus rich user experience.

vi) It is suitable for Type 2 discoveries because the same set of parameters can be used to excite different named/dedicated NN modules (e.g. the NN named after the herb X (i.e. the X-NN module, or an illness Y (then it is the dedicated Y-NN). If the outputs of the named NN modules for the same input set are computed relevance indices, then the "names" of the dedicated NN modules can be correlated according to the indices. This correlation or "hidden association" would lead to Type 2 discoveries. The rationale is the following probability expression, where P_r indicate the probability and U and V are two different subsets sets arbitrarily chosen from the common knowledge base K: $P_r(UV) = P_r(U \cap V) = P_r(U) + P_r(V) - P_r(U \cup V)$. If K is the population $\Omega = \{s_i\}$ and $i = 1,2,..., j,..,k,...,n$ the following are logically true: $CL_j = (s_j)$; $CL_k = (s_k)$; $CL_j \neq CL_k$; and $< CL_j, CL_k > \in K$. $CL_j = (s_j)$ says that the class CL_j is defined by the set of attributes represented by s, where the subscript j (i.e. s_j) marks the particular of s entry. Conceptually $P_r(UV)$ or $P_r(U \cap V)$ is the relevance index (RI) or the degree of similarity between U and V. The actual RI value, however, depends on the "***referential host***" (details in Chapter 6).

7.5 Suitability of the Backpropagation NN Approach

The backpropagation NN approach is immensely suitable for classification and discovery operations in the TCM domain. When entities and their associations are put together with equal importance in the TCM domain,

the result could be a 2-dimensional mesh as shown by *part (1)* in Figure 7.5.1. The entities (e.g. A, B, C, etc.) appear as nodes that can represent anything (e.g. node E represents {1/風寒襲肺/Wind Cold Assailing the Lung}; node A represents the {10/脈浮緊/Floating Tight Pulse} symptom; node G represents {7/苔薄白/Thin White Fur} symptom – as shown in Figure 7.1.1). If node E is lifted as the output node, then nodes A and G are the input nodes as shown by *part (2)* in Figure 7.5.1. In fact, *part (2)* can be mapped into the generic backpropagation NN architecture shown in Figure 7.4.1 easily. Likewise, if node B presents an herbal ingredient (e.g. {Ephedra/麻黃}) and nodes G and I represent {7/苔薄白/Thin White Fur} and {10/脈浮緊/Floating Tight Pulse} symptoms respectively, then the NN that represents *part (3)* in Figure 7.5.1 can be trained to yield Ephedra/麻黃 as the output.

Strictly speaking *part (2)* and *part (3)* are two different AI classifiers; the former is for identifying the illness and the latter is for identifying the herbal ingredient that treats the input symptoms. The same set of symptoms can be used as the inputs to excite the two different AI classifiers to generate very different results. Assuming the following we may make some herbal discoveries: i) v1 = {7/苔薄白/Thin White Fur} ; ii) v2 = {10/脈浮緊/Floating Tight Pulse}; iii) v3 = {Ephedra/麻黃}; (iv) v4 = {1/風寒襲肺/Wind Cold Assailing the Lung}. This can be explained by using the set theory. Part (2) of Figure 7.5.1 can be represented by the set $Z_a = (v1, v2, v4)$, and part (3) of Figure 7.5.1 can also be represented by $Z_b = (v1, v2, v3)$. Then, $Z_a \cap Z_b = (v1, v2)$ implies that both sets have the same two symptoms ($\cap$ for

intersection), and *v*3 (i.e. {Ephedra/麻黃}) should treat *v*4 (i.e. {1/風寒襲肺/Wind Cold Assailing the Lung}). Conceptually, if the same set of symptoms is the input to two different NN classifiers, the outputs from the two different classifiers would have an association of the pre-defined nature, as shown in Figure 7.5.2.

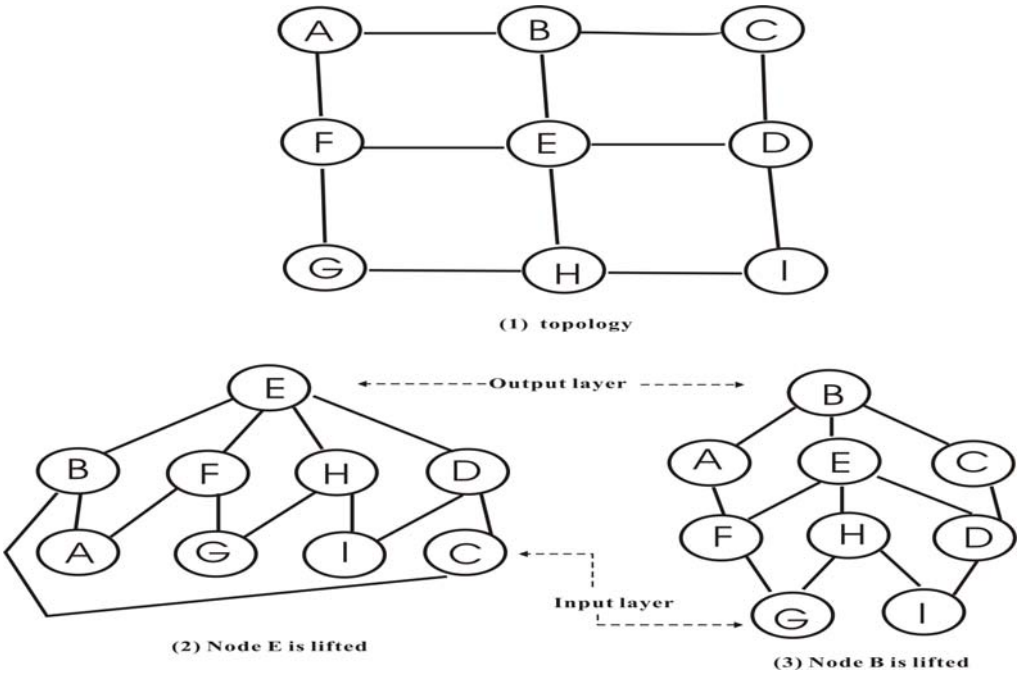


Figure 7.5.1 Creating subsumption hierarchies by lifting nodes

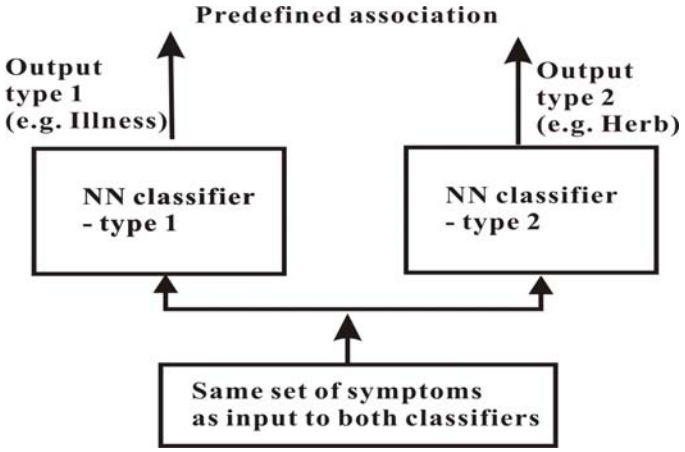


Figure 7.5.2 Discoveries by the pre-defined associations

Figure 7.5.3 is the 4-way association mesh that allows any side (e.g. symptoms) to be the input for the backpropagation NN artificial intelligence to yield one of the predefined results (e.g. prescriptions). The convention in this thesis is to name the classifier after the expected output versus the input). For the mesh there are at least 12 possible classifiers as follows: 1) illness versus symptoms; 2) prescriptions versus symptoms; 3) herb versus symptoms; 4) illness versus prescriptions; 5) herb versus prescriptions; 6) symptom versus prescriptions; 7) herb versus illnesses; 8) symptom versus illnesses; 9) prescription versus illnesses; 10) illness versus herbs; 11) symptom versus herbs; and 12) prescription versus herbs. If a set of symptoms $Z_s = (S1, S2, S3)$: i) can be treated by a set of herbal ingredients or herbs Z_h ; ii) exhibited by a set of illnesses Z_{il} ; and iii) treated by a set of prescriptions Z_{pp} , then logically their intersection looks like Figure 7.5.4.

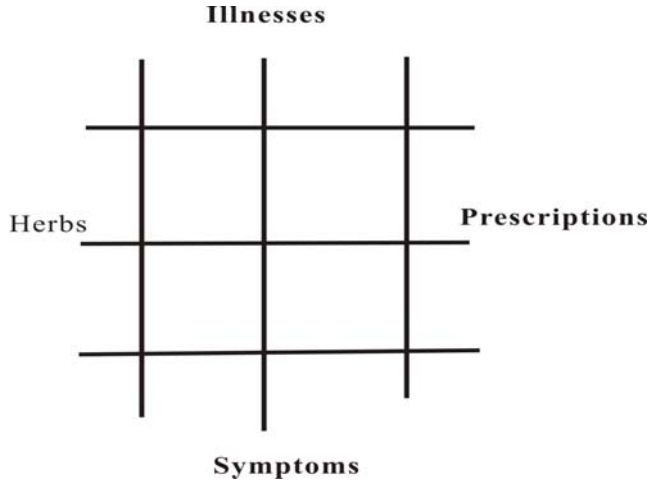


Figure 7.5.3 The 4-way association mesh topology

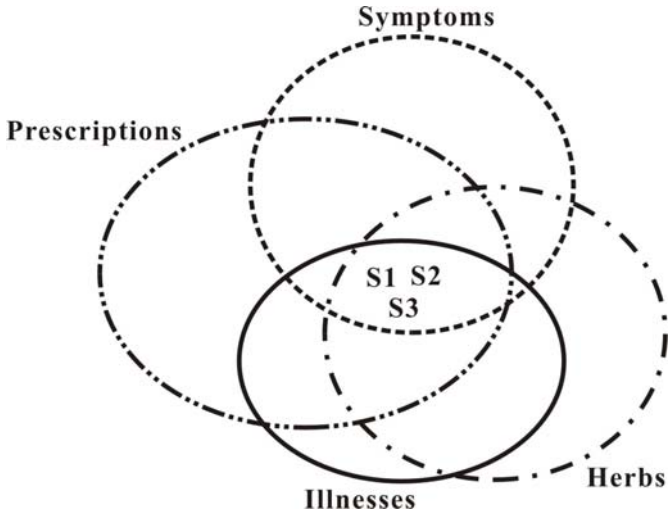


Figure 7.5.4 The intersection among four different classes/sets

Figure 7.5.4 actually fulfils the TCM SAME or "同病異治, 異病同治" (in Chinese terminology) principle, which was enshrined in the core of TCM since its dawn. In English this principle is defined as: "*If the symptoms are the same or similar, different conditions could be treated in the same way medically, independent of whether they come from the same illness or different ones* [WHO07]." By applying the SAME principle, we can "***discover***" the following, even if the illness name exhibiting the symptoms is not known: i) useful herbs that treat it; and ii) useful prescriptions that treat it [JWong08c].

7.6 Sensitivity Analysis and Real-Time NN Pruning

Sensitivity Analysis is the study of how the value change in a variable x would affect the output of the associated function, namely $f(x)$. In real-life applications the resultant change in $f(x)$ may invoke some necessary actions, which are defined with respect to the problem domain. For example, in the area

of neural network (NN) applications $f(x)$ may be regarded as the output of a neuron or the output of the entire NN, depending on the problem of interest. In this light the induced changes in $f(x)$ due to change in x may form the basis for optimizing the NN structure. If the NN structure is continuously optimized by pruning or eliminating inert arcs due to changes in x on the fly, it is real-time NN optimization or pruning. The NN application of such is well discussed in various publications (e.g. [Wong08, Lin04, Gallant92].

One of the many formal techniques that can be used in sensitivity analysis is the Hessian matrix [Wong08]. The matrix simply indicates if $f(x)$ has a maxima or minima or both. The functional interpretation of the maxima/minima point is an application which has to be pre-defined. That is, the Hessian matrix provides a generic formal background for the interpretations to adhere to, however, trials. For example, it was proved by previous research results that if $f(x)$ converges so will its derivatives [Gallant92]. This is the basis for the HBP (Hessian-based pruning) optimization of a NNC (neural network controller) dynamic buffer tuner [Lin04]. In fact, the variable x can assume any parametric characters (e.g. weights, biases…) in a real application, and the speed for $f(x)$ to converge to the target value depends on the coefficient(s) chosen. In effect, we can make $f(x)$ converge faster by adjusting its coefficient(s), in light of "learning laws" [Hagan96]. Learning laws are basically forecast methods such as the famous Moore's law in the PB^t form, where P is the curve fitting parameter, B the learning rate (that is, how much has been learned, in percentage, from the last operation that would contribute to

improve the next one), and *t* the time units involved per "learning & improvement" cycle [Lewis96].

The Hessian matrix can be derived from the Taylor Series [Finney94], as shown by Equation 7.6.1; where $F(w)$ is the function of variable w (e.g. to represent the NN connection/arc weight) ; Δw for the change in w ; T superscript for *transpose*; O for higher order terms. Then, the gradient matrix for $F(w)$ is defined by Equation 7.6.2; where n for the n^{th} term and $\partial / \partial w_1$ for the 1^{st} order partial differentiation. The 2^{nd} order partial differentiations form a Hessian matrix, as shown by the expression Equation 7.6.3.

$$F(w+\Delta w) = F(w) + \nabla F(w)^T \Delta w + \frac{1}{2} \Delta w^T \nabla^2 F(w) \Delta w + O(||\Delta w||^3) + \ldots$$

Equation7.6.1

$$\nabla F(w) : \left[\frac{\partial}{\partial w_1} F(w) \quad \frac{\partial}{\partial w_2} F(w) \ldots \frac{\partial}{\partial w_n} F(w) \right]^T$$

Equation7.6.2

$$\nabla^2 F(w) : \begin{bmatrix} \frac{\partial^2}{\partial w_1^2} F(w) & \frac{\partial^2}{\partial w_1 w_2} F(w) & \cdots & \frac{\partial^2}{\partial w_1 \partial w_n} F(w) \\ \frac{\partial^2}{\partial w_2 \partial w_1} F(w) & \frac{\partial^2}{\partial w_2^2} F(w) & \cdots & \frac{\partial^2}{\partial w_2 \partial w_n} F(w) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2}{\partial w_n \partial w_1} F(w) & \frac{\partial^2}{\partial w_n \partial w_2} F(w) & \cdots & \frac{\partial^2}{\partial w_n^2} F(w) \end{bmatrix}$$

Equation7.6.3

We will walk through the semantics of the expressions, Equation 7.6.1, Equation 7.6.2 and Equation 7.6.3. Before that $F(w)$ is instantiated into the following form first: $F(w) = F(x) = F(x_1, x_2) = (x_2 - x_1)^4 + 9x_1x_2 - x_1 + 2x_2 - 1$. Conceptually, the generic variable w is now defined by two real-life variable x_1 and x_2. The walkthrough will cover:

a) Start with the given extreme (stationary) point, $x^1=[0.57, -0.57]^T$, as the example; $\nabla F(x^1) = 0$.

b) Determine the global minimum with the extreme point.

c) Find the 2^{nd} Taylor series expansion for the minimum.

d) Plot $F(x)$ the Taylor series expansion in step c) above.

The phases in the walkthrough example are as follows:

Phase 1: From $F(w) = Fx) = F(x_1, x_2) = (x_2 - x_1)^4 + 9x_1x_2 - x_1 + 2x_2 - 1$ we get:

$$\nabla F(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} F(x) \\ \frac{\partial}{\partial x_2} F(x) \end{bmatrix} = \begin{bmatrix} -4(x_2 - x_1)^3 + 9x_2 - 1 \\ 4(x_2 - x_1)^3 + 9x_1 + 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Phase 2: Compute the Hessian matrix as follows:

$$\nabla^2 F(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(x) & \frac{\partial^2}{\partial x_1 \partial x_2} F(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(x) & \frac{\partial^2}{\partial x_2^2} F(x) \end{bmatrix} = \begin{bmatrix} 12(x_2 - x_1)^2 & -12(x_2 - x_1)^2 + 9 \\ -12(x_2 - x_1)^2 + 9 & 12(x_2 - x_1)^2 \end{bmatrix}$$

Phase 3: Compute the Eigen values to check if the Hessian matrix is positive definite; if the Eigen values are positive the likelihood for the matrix to have a minimum is strong; substituting $x^1=[0.57, -0.57]^T$ into the $\begin{bmatrix} 12(x_2-x_1)^2 & -12(x_2-x_1)^2+9 \\ -12(x_2-x_1)^2+9 & 12(x_2-x_1)^2 \end{bmatrix}$ matrix to yield $\nabla^2 F(x^1)=\begin{bmatrix} 15.59 & -6.59 \\ -6.59 & 15.59 \end{bmatrix}$.

Phase 4: Compute the Eigen values (λ) by transforming $\begin{bmatrix} 15.59 & -6.59 \\ -6.59 & 15.59 \end{bmatrix}$ into the new form $\begin{bmatrix} 15.59-\lambda & -6.59 \\ -6.59 & 15.59-\lambda \end{bmatrix}$, which yields $(15.59-\lambda)^2-(6.59)^2=0$; implying $\lambda_1=22.18$ and $\lambda_1=9$ λ_2=9 ; therefore x^1 is a strong minimum (minima).

Phase 5: Plot $F(w)=F(x)=F(x_1,x_2)=(x_2-x_1)^4+9x_1x_2-x_1+2x_2-1$ (as shown in Figure 6.6.1) to visualize the minimum at $x^1=[0.57, -0.57]^T$

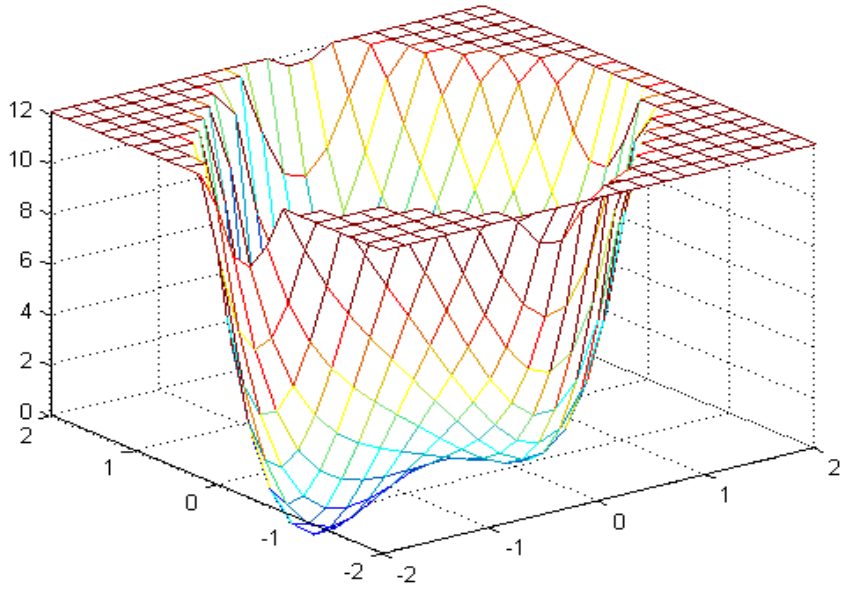


Figure 7.6.1 The minimum at $\mathbf{x^1}$=[0.57, -0.57]^T for $F(w)$

7.6.1 Neural Network Pruning in the WD^2UHI Research

The ***formal background*** for NN pruning, which depends on the nature of the chosen function $f(w)$ that controls the behavior of the neuron, is the Hessian matrix. The argument is that if $f(w)$ has a minimum, then its output can be skipped/pruned it does not has impact on the output of the next function downstream in the cascade.

Figure 7.6.1.1 shows the relationship between input neurons (Node#1 and Node#2) and the neuron, which takes $w3 = f1(w1, w2)$ as its input to produce its own output $O2 = w4 = f2(w3)$. In reality, $f1$ (e.g. addition) and $f2$ (e.g. Sigmoid) are pre-defined. Conceptually, $O2 = F(w1, w2)$ can summarize the combined effect of $f1$ (e.g. addition) and $f2$ (e.g. Sigmoid) in between the

input (i.e. $w1$ and $w2$) and the output $O2$. Conceptually, $F(w1, w2)$ has the similar property to the previous example, $F(x_1, x_2) = (x_2 - x_1)^4 + 9x_1x_2 - x_1 + 2x_2 - 1$. In theory, $O2 = F(w1, w2)$ would have minimum points, which can be checked by computing the Eigen values from the corresponding Hessian matrix. With the Sigmoid function the minimum should be the zero value. From the application point of view $O2 = w4 = 0$ is an inert arc (for its 0 weight) and does not contribute to the computation of the Sigmoid output of the next layer above (e.g. $O3 = f3(w4) = 0$; assuming $f3$ is also Sigmoid. Therefore, $f3$ can be skipped (i.e. *pruned logically*) in the subsequent NN computations to save time. In fact, this argument has a formal Hessian background to prop it up. The interpretation of $O2 = w4 = 0$, however, depends on the application case and the function chosen for the action (e.g. Sigmoid). In my PhD research the NN (backpropagation) computation works by the Sigmoid function only, and the inputs can be added because they are normalized.

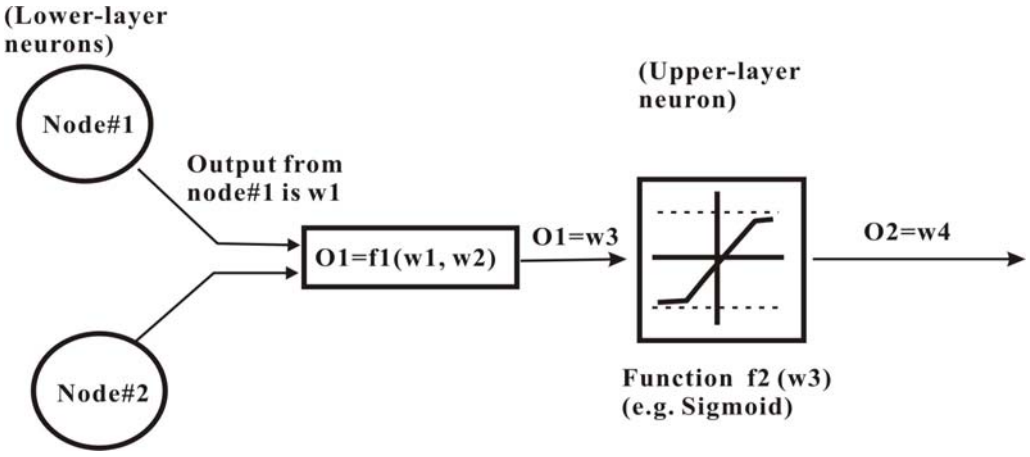


Figure 7.6.1.1 Functional relationship between two layers of neurons

$$O2 = F(w1, w2)$$

Many experiments were conducted to verify the Hessian-based approach for real-time NN pruning. The experimental results unanimously confirm that the execution time of the NN (backpropagation) modules can indeed be reduced by applying this approach. Some of the experimental results are presented in section 9.4.2.3 in Chapter 9.

7.6.2 Termination of Pruning

Equation 7.6.2.1 is the Taylor series for $f(x)$ with the Lagrange form of the remainder (or error term), defined by Equation 7.6.2.2. It approximates $f(x)$ by a polynomial, as long as $f(x)$ is continuous and differentiable in the interval $(x - x_0)$; $f'(x)$ is the 1st order differentiation and $f^{(n)}(x)$ the n^{th} order. This power series $f(x)$ generally converges for all values of x for an interval (i.e. the *interval of convergence*), for example $(x - x_0)$ and diverges outside this interval. In fact, the independent variable x in Equation 7.6.2.1 can be single or multiple in nature.

$$f^{(n)}(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots + \frac{f^{(n)}(x_0)(x - x_0)^n}{n!} + \frac{f^{(n+1)}(\xi)(x - x_0)^{n+1}}{(n+1)!} \quad \text{.....Equation\_7.6.2.1}$$

$$R_n = \frac{f^{(n+1)}(\xi)(x - x_0)^{n+1}}{(n+1)!} \quad \text{..... .... Equation 7.6.2.2}$$

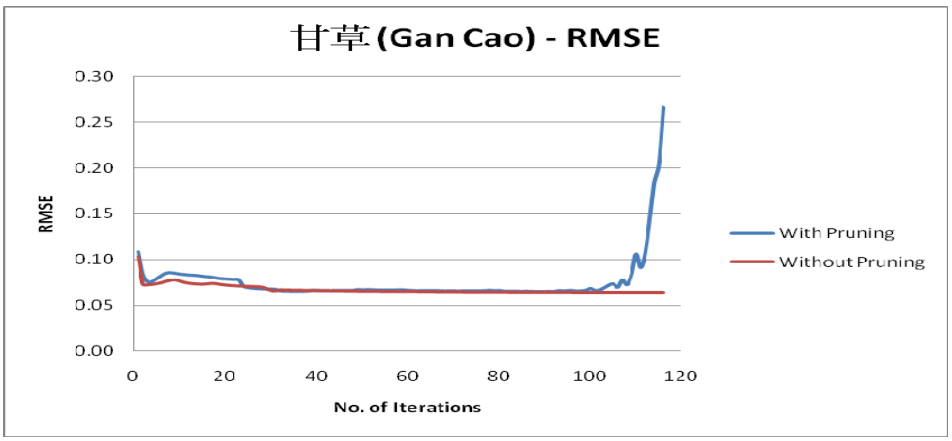


Figure 7.6.2.1 RMSE decay during NN training or pruning

Figure 7.6.2.1 shows how two real-life RMSE (root mean square error) curves decay gradually. The "un-pruned curve" decays for the whole interval of convergence, but the decay of the "pruned" curve suddenly stopped and increased exponentially. In fact, the example in Figure 7.6.2.1 is the result of one of the NN experiments in my research. It is shown here as the example for discussion. The trend of RMSE (without pruning) is conceptually $f(z)$, where z is the independable variable. In this case z is the training episodes or the pruning iterations (both to be defined in Chapter 9 later). Conceptually, the single-variable curve $f(z)$ fits the philosophy defined by both equarions Equation 7.6.2.1 and Equation 7.6.2.2.

The "curve with pruning", however, fits the equations Equation 7.6.2.1 and Equation 7.6.2.2 only up to the point when "No. of iterations" is roughly equal to 105. After that this curve becomes unstable. The reason for this is that the pruning process has changed the very nature of the original $f(z)$ to the new $f_P(z)$, which has a shorter interval of convergence. The *formal adherence*

to explain this phenomenon is the ***Lagrange remainer*** R_n, which says that $R_n \to 0$ for $n \to \infty$. This logical adherence can be shown by using $f^{(n)}(x)$ alone because the value of $f^{(n)}(x)$ decays when n increases. The example below in the sequel walks through this point step by step.

In this example a simple neural network (NN) (Figure 7.6.2.2) is used. The *activation functions* for the neurons are assumed Sigmoid, $f(x) = \frac{1}{(1+e^{-x})}$. Then, we can represent the HO (Hidden-Output) layer of this NN model by the $f(x_1, x_2)$ matrix, shown by Equation 7.6.2.3.

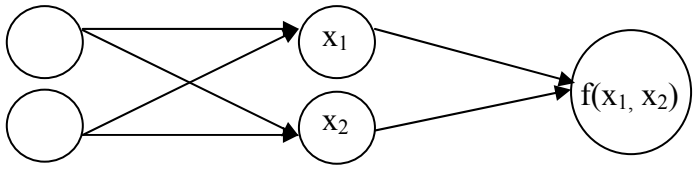


Figure 7.6.2.2 A simple neural network

$$f(x_1, x_2) = \begin{bmatrix} \frac{1}{(1+e^{-x_1})} \\ \frac{1}{(1+e^{-x_2})} \end{bmatrix} \text{..... } \textit{Equation } 7.6.2.3$$

From Equation 7.6.2.3 we can the matrices of differentials as follows:

i) 1st order differentiation of the 2-variable $f(x_1, x_2)$:

$$f^{(1)}(x_1, x_2) = \begin{bmatrix} \dfrac{e^{-x_1}}{(1+e^{-x_1})^2} & 0 \\ 0 & \dfrac{e^{-x_2}}{(1+e^{-x_2})^2} \end{bmatrix}$$

ii) 2nd order (Hessian matrix):

$$f^{(2)}(x_1, x_2) = \begin{bmatrix} \dfrac{2(e^{-x_1})^2}{(1+e^{-x_1})^3} - \dfrac{e^{-x_1}}{(1+e^{-x_1})^2} & 0 \\ 0 & \dfrac{2(e^{-x_2})^2}{(1+e^{-x_2})^3} - \dfrac{e^{-x_2}}{(1+e^{-x_2})^2} \end{bmatrix}$$

iii) Examples of higher orders:

$$f^{(3)}(x_1, x_2) = \begin{bmatrix} \dfrac{6(e^{-x_1})^3}{(1+e^{-x_1})^4} - \dfrac{6(e^{-x_1})^2}{(1+e^{-x_1})^3} + \dfrac{e^{-x_1}}{(1+e^{-x_1})^2} & 0 \\ 0 & \dfrac{6(e^{-x_2})^3}{(1+e^{-x_2})^4} - \dfrac{6(e^{-x_2})^2}{(1+e^{-x_2})^3} + \dfrac{e^{-x_2}}{(1+e^{-x_2})^2} \end{bmatrix}$$

$$f^{(4)}(x_1, x_2) = \begin{bmatrix} \dfrac{24(e^{-x_1})^4}{(1+e^{-x_1})^5} - \dfrac{36(e^{-x_1})^3}{(1+e^{-x_1})^4} + \dfrac{14(e^{-x_1})^2}{(1+e^{-x_1})^3} - \dfrac{e^{-x_1}}{(1+e^{-x_1})^2} & 0 \\ 0 & \dfrac{24(e^{-x_2})^4}{(1+e^{-x_2})^5} - \dfrac{36(e^{-x_2})^3}{(1+e^{-x_2})^4} + \dfrac{14(e^{-x_2})^2}{(1+e^{-x_2})^3} - \dfrac{e^{-x_2}}{(1+e^{-x_2})^2} \end{bmatrix}$$

$$f^{(5)}(x_1, x_2) = \begin{bmatrix} \dfrac{120(e^{-x_1})^5}{(1+e^{-x_1})^6} - \dfrac{204(e^{-x_1})^4}{(1+e^{-x_1})^5} + \dfrac{150(e^{-x_1})^3}{(1+e^{-x_1})^4} - \dfrac{30(e^{-x_1})^2}{(1+e^{-x_1})^3} + \dfrac{e^{-x_1}}{(1+e^{-x_1})^2} & 0 \\ 0 & \dfrac{120(e^{-x_2})^5}{(1+e^{-x_2})^6} - \dfrac{204(e^{-x_2})^4}{(1+e^{-x_2})^5} + \dfrac{150(e^{-x_2})^3}{(1+e^{-x_2})^4} - \dfrac{30(e^{-x_2})^2}{(1+e^{-x_2})^3} + \dfrac{e^{-x_2}}{(1+e^{-x_2})^2} \end{bmatrix}$$

iv) Summary of up to the 5th order for $x = 0.5$:

Order n	$f^{(n)}(x)$ fore x=0.5
1	0.222222
2	0.074074
3	0.074074
4	0.123457
5	0.057613

Table 7.6.2.1 Differentiation results

v) The plot (Figure 7.6.2.3) for Table 7.6.2.1 shows that the absolute values for $f^{(n)}(x)$ (i.e. $|f^{(n)}(x)|$) decay quickly, as marked by the trend-line. This implies that $f(x_1, x_2)$ adheres to the philosophy expressed by Equation 7.6.2.1 and Equation 7.6.2.2. Since the minimum of $f(x) = \frac{1}{(1+e^{-x})}$ is zero, the outputs from the neurons (in the range of [0,1]) in the hidden layer can be pruned because they do not affect the computation results in the next stage in the NN cascade.

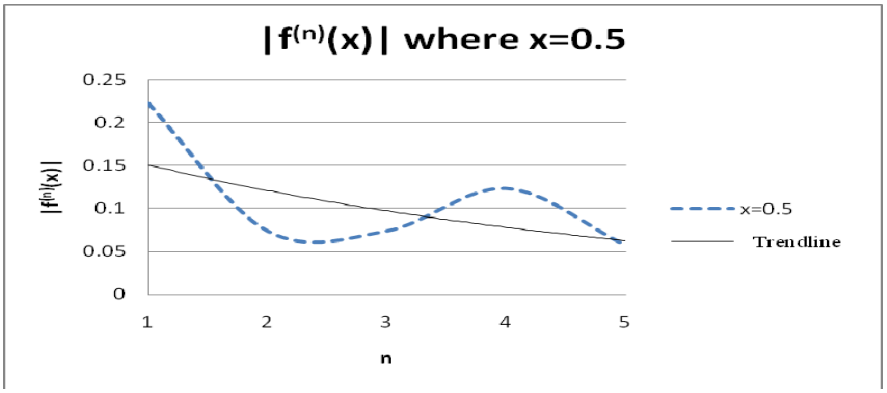


Figure 7.6.2.3 The absolute values of $f^{(n)}(x)$ decay quickly

The pruning of the original $f(x_1, x_2)$, however, changes its original property, similar in effect to what is shown in Figure 7.6.2.1. We can decide when pruning can be stopped by checking the sudden rise in the RMSE value.

7.6.3 Special Meaning for the RMSE

The NN (backpropagation) modules in the *WD²UHI* platform are ***named*** and trained in the "*supervised*" fashion [Callan03]. The NN convergence is conceptually toward the target name/class/label. The target value, however, not controlled by any objective function in the form of $\{0, \Delta\}$, where "0" is the teacher signal or reference and Δ the error limit/tolerance about "0". For any "*supervised*" training with $\{0, \Delta\}$, learning is regarded as completed only if the difference between the trained NN output and the reference "0" is consistently less than $|\Delta|$. In the *WD²UHI* case the final convergence of the NN output depends on the set of training data rather than "0" ; Δ is ignored.

For the *WD²UHI* the "*name*" of the NN module conceptually means 1 or 100%. This is conceptually true for a perfectly trained NN. But, in reality, the learned NN might have converged to any value in between 0 and 1 (i.e. the final "target"). This convergence depends on the size and the "*completeness*' of the training dataset. For this reason, the statement, "*training has completed*" for *WD²UHI* has the following possible connotations:

a) The output of the named NN module would have a value between 0 and 1.
b) The training stops because the set of training data has been depleted; the number of training episodes (repeated training session with the same data set) has decided attainment of the final NN output.
c) The named NN has settled to a specific RMSE (root means square error) value.

When the trained NN module is invoked with a single raw input set (i.e. not in the original training data set), then it would produce an output (O) anywhere in between 0 and 1. As a matter of fact, this is similar in effect to the *automatic semantic aliasing* mechanism (Chapter 6) that produces the relevance index (i.e. the degree of similarity) from the angle of the referential "host". The NN output O associates the input data to the "name" of the NN module; this association is the degree of similarity of the input data set to the very nature of the NN "name" (e.g. Flu/感冒). *Therefore, O can be treated as a relevance index of the "input data set" to the NN name/class/label*.

7.7 Recap

Using the algorithmic approach to build to classifier is not appropriate because it is difficult to specify the degree of importance (or weight) for a parameter (the interpretation of a "true" state). As the algorithmic approach is implicit, the ontology based approached is used. Its explicit nature has its

advantages. Combining neural network with the ontology based approach, the advantages are: i) it is generic; ii) input can be normalized; iii) the network can be trained; iv) pruning is possible and v) its nature avoids traditional engineering problems. This conceptual approach is named as LCD&BNN (lift-classify-define backpropagation neural network).

7.8 Experimental Examples for Demonstration

Many experiments were set up to verify that the backpropagation NN approach can indeed deliver TCM classifications effectively and lead to discoveries by the SAME principle. More detailed experimental results are selected and presented in section 9.5.1 in Chapter 9. The experimental examples here are aimed at making the background of Chapter 7 clearer. All the experiments relevant to this chapter and section 9.5.1 had made use of the Nong's proprietary enterprise TCM ontology. Since this enterprise ontology is very big, only the Influenza sub-ontology (also sizeable) was isolated for my planned experiments; the isolation is similar in concept to Figure 7.2.1 and Figure 7.2.2. Figure 7.8.1 underlines the three types of causes that underline the Influenza sub-ontology (bilingual - Chinese/English). The subsumption hierarchy of this sub-ontology will be mapped into the corresponding backpropagation NN architecture as shown in Figure 7.4.1.

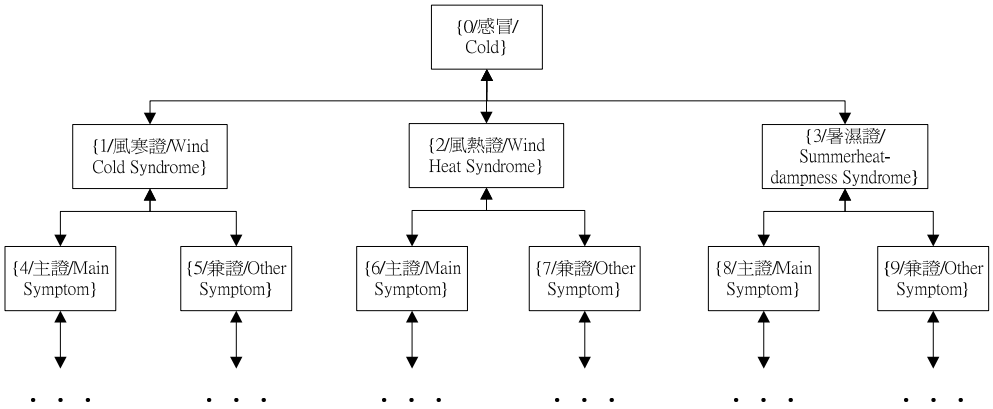


Figure 7.8.1 Three types of causes that underline the Influenza sub-ontology

The partial Influenza sub-ontology isolated from the Nong's enterprise TCM ontology as an XML metadata system is listed below (Figure 7.8.2) (the full sub-ontology can be found in Appendix III):

<?xml version="1.0" encoding="Big5" ?>

<感冒> (<Cold>)

<風寒證> (<Wind Cold Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defense-exterior >)

<怕冷> (<Fear of Cold>)

<怕冷重>id="1"</怕冷重>

(<Seriously Fear of Cold>id="1"</Seriously Fear of Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<發熱輕>id="1"</發熱輕>

(<Mild Fever>id="1"</Mild Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<無汗>id="1"</無汗>

(<Absence of Sweating>id=”1”</Absence of Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭痛四肢痠痛>id="1"</頭痛四肢痠痛>

(<Headache and Limbs Pain>id=”1”</Headache and Limbs Pain>)

</頭身> (</Head and Body>)

</衛表> (</Defence-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻塞流清涕多嚏>id="1"</鼻塞流清涕多嚏>

(<Nasal Congestion, Clear Sniffle and Profuse Sneezing>id=”1”

</Nasal Congestion, Clear Sniffle and Profuse Sneezing>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽癢>id="1"</咽癢> (<Throat Itching>id=”1” </Throat Itching>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽聲重>id="1"</咳嗽聲重>

(<Profuse Coughing>id=”1”</Profuse Coughing>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰稀薄色白>id="1"</痰稀薄色白>

(<White Clear Phlegm>id=”1”</White Clear Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<口不渴或渴喜熱飲>id="1"</口不渴或渴喜熱飲>

(<Not Thirsty Nor Fancy Hot Drinks>id=”1”</Not Thirsty Nor Like Hot Drinks>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌苔薄白>id="1"</舌苔薄白> (<Thin White Fur>id=”1”</Thin White Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈象浮或兼緊>id="1"</脈象浮或兼緊>

(<Floating Tight Pulse>id=”1”</Floating Tight Pulse>)

</脈象> (/<Pulse>)

</風寒證> (</Wind Cold Syndrome >)

…

</感冒> (</Cold>)

Figure 7.8.2 The partial isolated Influenza sub-ontology in XML

7.8.1 Finding a Suitable Mature NN Tool

In order to avoid inadvertent errors in process of building the backpropagation NN architecture, mature tools (freeware) were selected in the market for evaluation and adoption. The preference is Java-based tools, for they usually have richer user experience and support. Eventually the WEKA tool, in which the NN is a function, was chosen. It was decided that the backpropagation NN architecture should have three layers, namely *one input layer*, *one hidden layer* and *one output layer*. This decision is based on

published previous experience in the literature search [Funahashi89, Hornik89]. This previous experience confirms that any continuous function can be approximated with an arbitrary accuracy using a 3-layered network. A concise survey of popular and free Java neural network tools has revealed the following: a) NNWJ (Neural Network with Java); b) JOONE (Java Object Oriented Neural Engine); and c) WEKA (Waikato Environment for Knowledge Analysis)

7.8.1.1 NNWJ (Neural Network with Java)

NNWJ was introduced by Jochen Frohlich's team in 1996 [NNWJ] with the aim to provide different types of neural networks such as Multilayer Perceptron, Backpropagation Net (BPN) and Kohonen Feature Map. BPN has the same structure as the Multilayer Perceptron and uses the backpropagation learning algorithm.

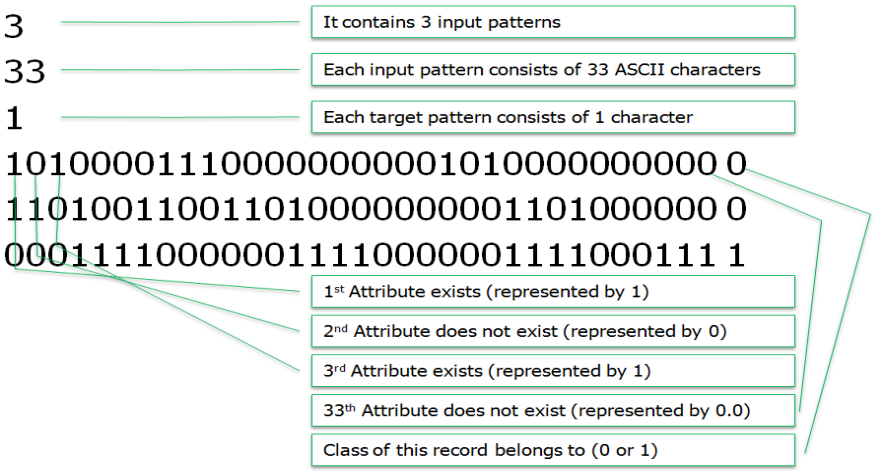


Figure 7.8.1.1.1 Input data format of NNWJ

7.8.1.2 JOONE (Java Object Oriented Neural Engine)

JOONE was introduced by Paolo Marrone's team in 2004 [JOONE]. The aim is to provide an environment to easily train many different neural networks, which are initialized with different weights, parameters or different architectures, in parallel.

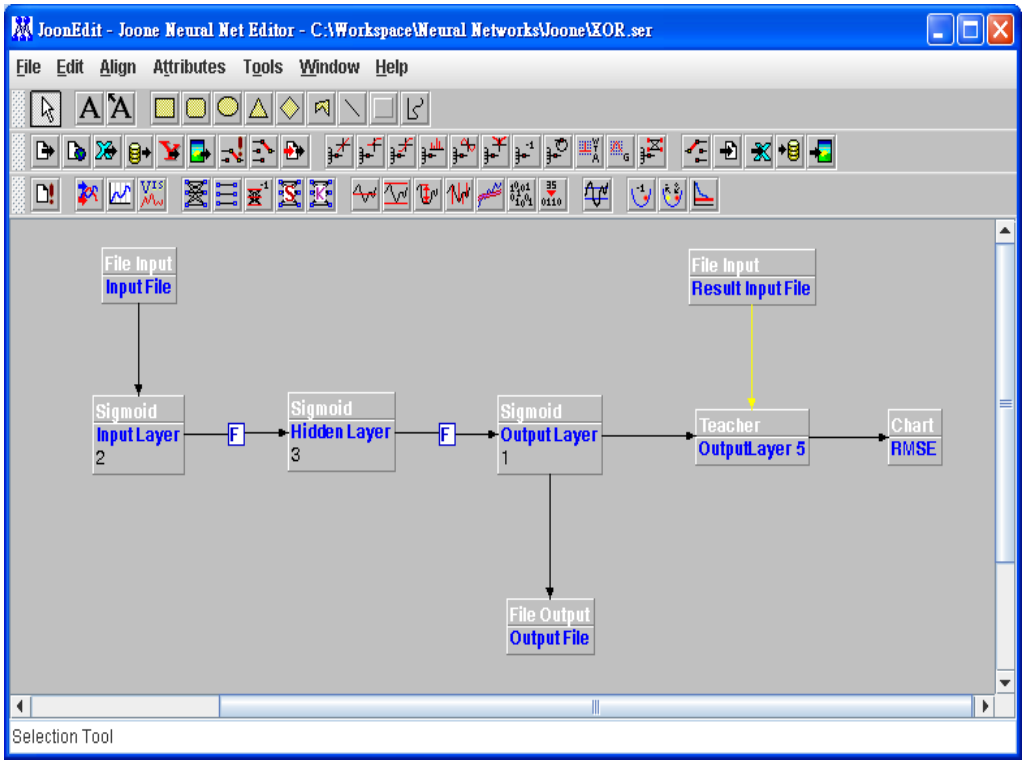


Figure 7.8.1.2.1 JOONE GUI editor

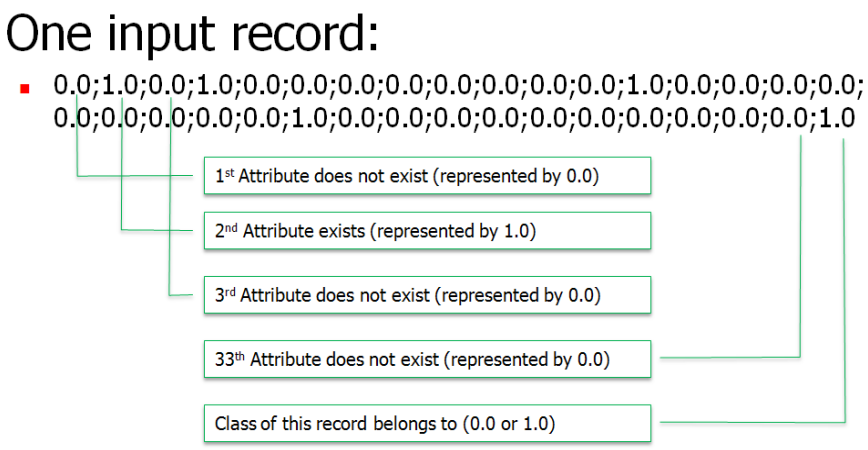


Figure 7.8.1.2.2 Input data format of JOONE

$$y_n = x_{(t-N)} \quad \text{where } 0 < N \leq taps$$

Figure 7.8.1.2.3 The activation function

After training the user simply chooses the fittest neural network to use. It is a Java framework that facilitates building and running applications on neural networks. JOONE applications can be built on a local machine, and then trained on a distributed environment and eventually run on whatever device. It has a modular architecture based on linkable components that can be extended to build new learning algorithms and neural network architectures. All the components have some basic specific features, including persistence, multithreading, serialization and parameterization. These features would guarantee scalability, reliability and expansibility. The JOONE provides a graphical user interface (the GUI Editor) to visually create, modify and train a neural network.

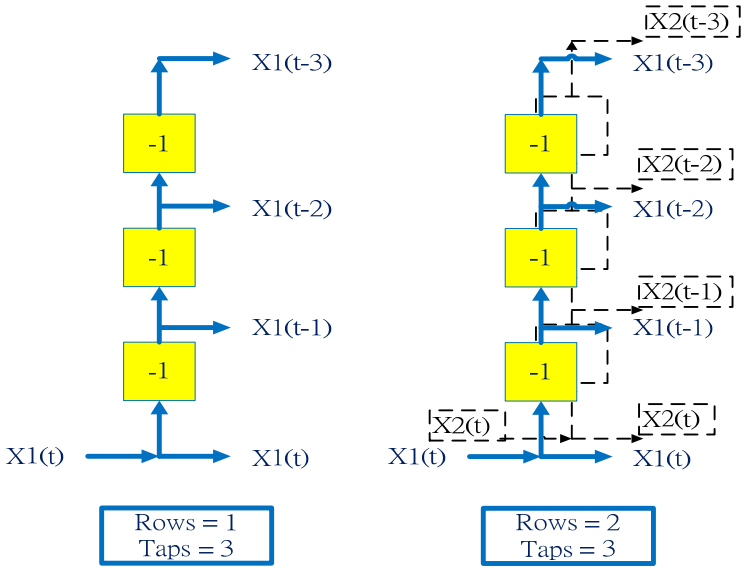


Figure 7.8.1.2.4 Taps parameter – delay layer in JOONE

There is a delay layer in JOONE that uses the sum of the input values to delay the number of iterations specified by the taps parameter [Marrone07]. An example is shown in Figure 7.8.1.2.2 in which two different delay layers are present; one with 1 row and 3 taps, and another with 2 rows and 3 taps.

The structure of the delay layer is as follows: a) number of inputs equal to the rows parameter; and b) number of outputs equal to rows * (taps +1). The taps parameter indicates the number of output delayed cycles for each row of neurons (plus one because the delayed layer also presents the actual input signal Xn (t) to the output). During the training process, error values will be fed backwards through the delay layer as required. This layer is popular to train a neural network for time-series prediction, giving it a “temporal window” of the input raw data.

7.8.1.3 WEKA (Waikato Environment for Knowledge Analysis)

WEKA [WEKA] was introduced by Eibe Frank, Mark Hall and Len Trigg in 1996. It is a collection of machine learning algorithms for data mining tasks and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes. MultiLayerPerceptron (MLP) is one of the classifiers of WEKA for machine learning (with backpropagation algorithm integrated).

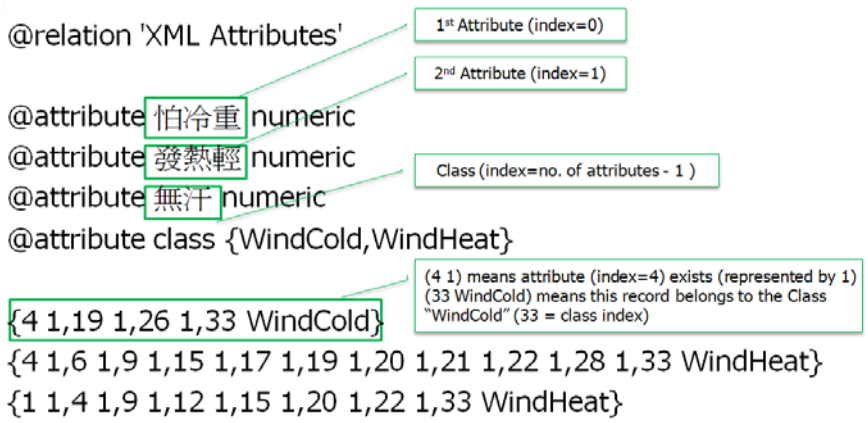


Figure 7.8.1.3.1 Input data format of WEKA

Decay in WEKA:

Weight decay is a method to constrain the training process so that a simple solution for a problem can be obtained [Kramer89]. In the machine learning literature, this type of constraint is defined as the "bias", which is necessary for the inductive leap from training data to future events. The weight decay method allows connection weights in a network to differentially decay

towards zero. This will divide the starting learning rate by the number of episodes to determine what the current learning rate should be. The purpose of using decay in neural network is to help stop the network from diverging from the target output, and improve general performance. If a connection weight is driven to zero during the training process, the input source for that connection weight will have no influence on the activation level of the output nodes. This will happen if the network determines that the input source is useless for the classification purpose. When the WEKA GUI is used, the decay learning rate will not be shown other than the original learning rate.

Similarly to JOONE, WEKA has the Java Class “Random” for generating pseudorandom numbers. The class uses a 48-bit seed, and works with the Linear Congruential Formula [Knuth97].

7.8.1.4 Comparison of the Three Tools

The three tools are compared concisely in Table 7.8.1.4.1.

	NNWJ (BPN)	JOONE	WEKA (MLP)
Number of Neurons Layer	Minimum=2 Maximum=N	Minimum=2 Maximum=N	Minimum=2 Maximum=N
Number of Neurons in Each Layer	Minimum=1 Maximum=N	Minimum=1 Maximum=N	Minimum=1 Maximum=N
Weight Matrices	Automatically created and initialized	Automatically created and initialized	Automatically created and initialized
Bias Values	Automatically used	Automatically used	Automatically used
Number of Input/Target Patterns	Minimum=1 Maximum=N	Minimum=1 Maximum=N	Minimum=1 Maximum=N
Activation Function	Sigmoid	Linear, Sigmoid, Tanh etc.	Sigmoid (or self-developed)
Delay Layer	Not Supported	Supports	Not Supported
Decay	Not Supported	Not Supported	Supports
Random Seed	Not Supported	Supports	Supports
Open Source	No	Yes	Yes

Table 7.8.1.4.1 Comparison of three common Tools of neural network in JAVA

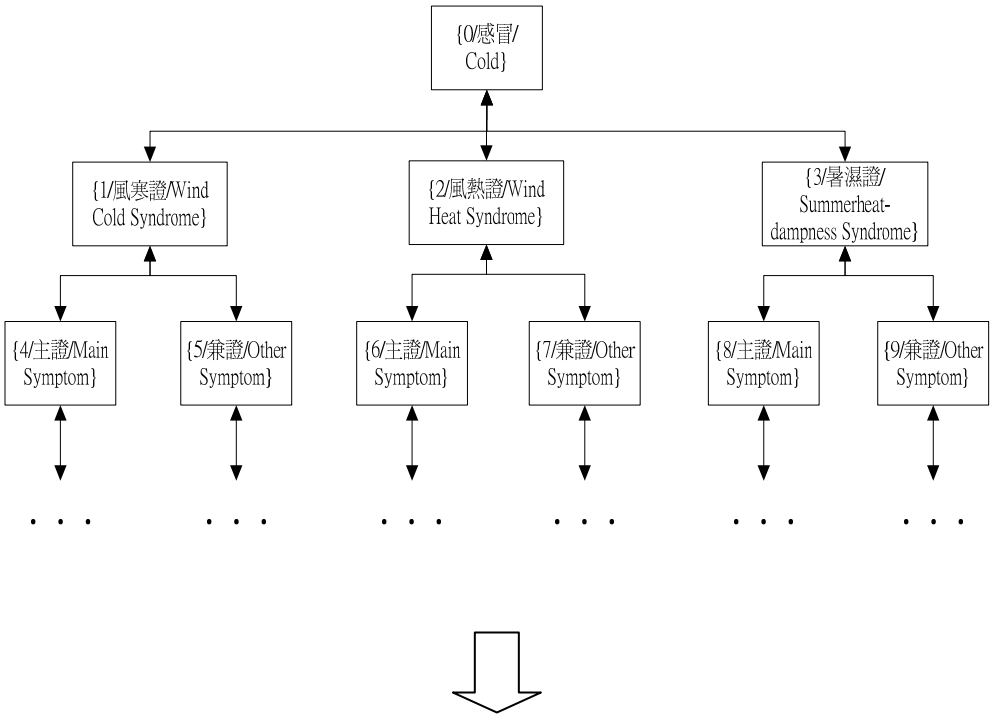
The comparison in Table 7.8.1.4.1 shows that WEKA has the advantage over NNWJ for being open-source (under the GNU General Public License). This enables the use of and/or modification of the original source code at will. It supports the decay function, which is needed in the relevant experiments. Most important of all, we have rich in-house experience with WEKA. For the above reasons WEKA was chosen.

7.8.2 Two Models for the Experiments

Two models should be tried and compared. From the experimental results I would identify the future direction for the implementation of the backpropagation NN approach that supports TCM classification and discoveries.

7.8.2.1 Model 1 – One Network Tree

This is the same as Figure 7.4.1 (excerpt below), with the following characteristics: a) number of networks = 1; b) input parameters = symptoms of all illness names and types in the DOM tree (i.e. 1 = exist, 0 = not exist); c) number of input neurons equal to the number of symptoms; d) number of hidden neurons equal to twice the number of the input neurons; and e) number of output neurons equal to one (i.e. the classification result).



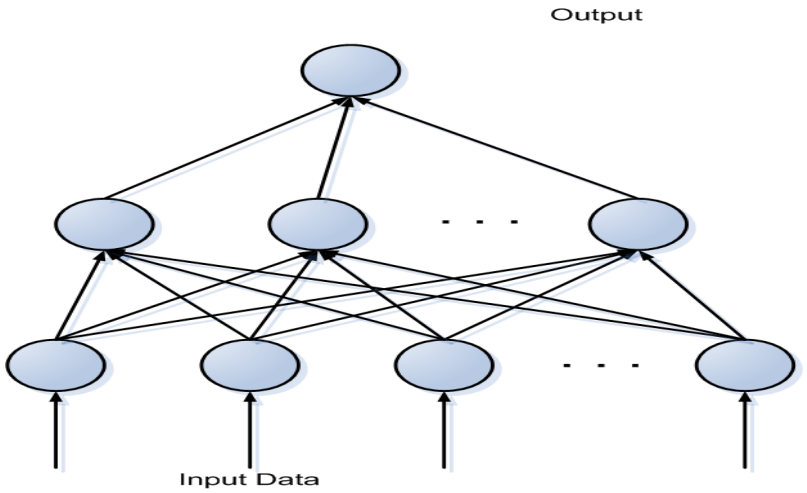
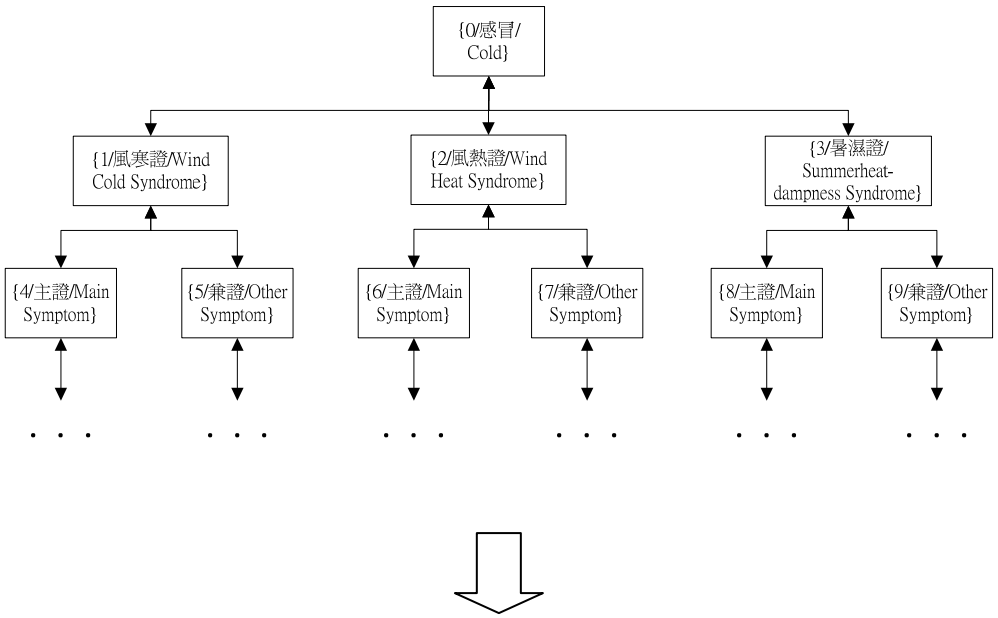


Figure 7.8.2.1.1 Mapping subsumption hierarchy and backpropagation NN

(Excerpt of Figure 7.4.1)

7.8.2.2 Model 2 – Separate Network Tree

In this model the DOM tree is partitioned into several smaller DOM trees as shown in Figure 7.8.2.2.1.



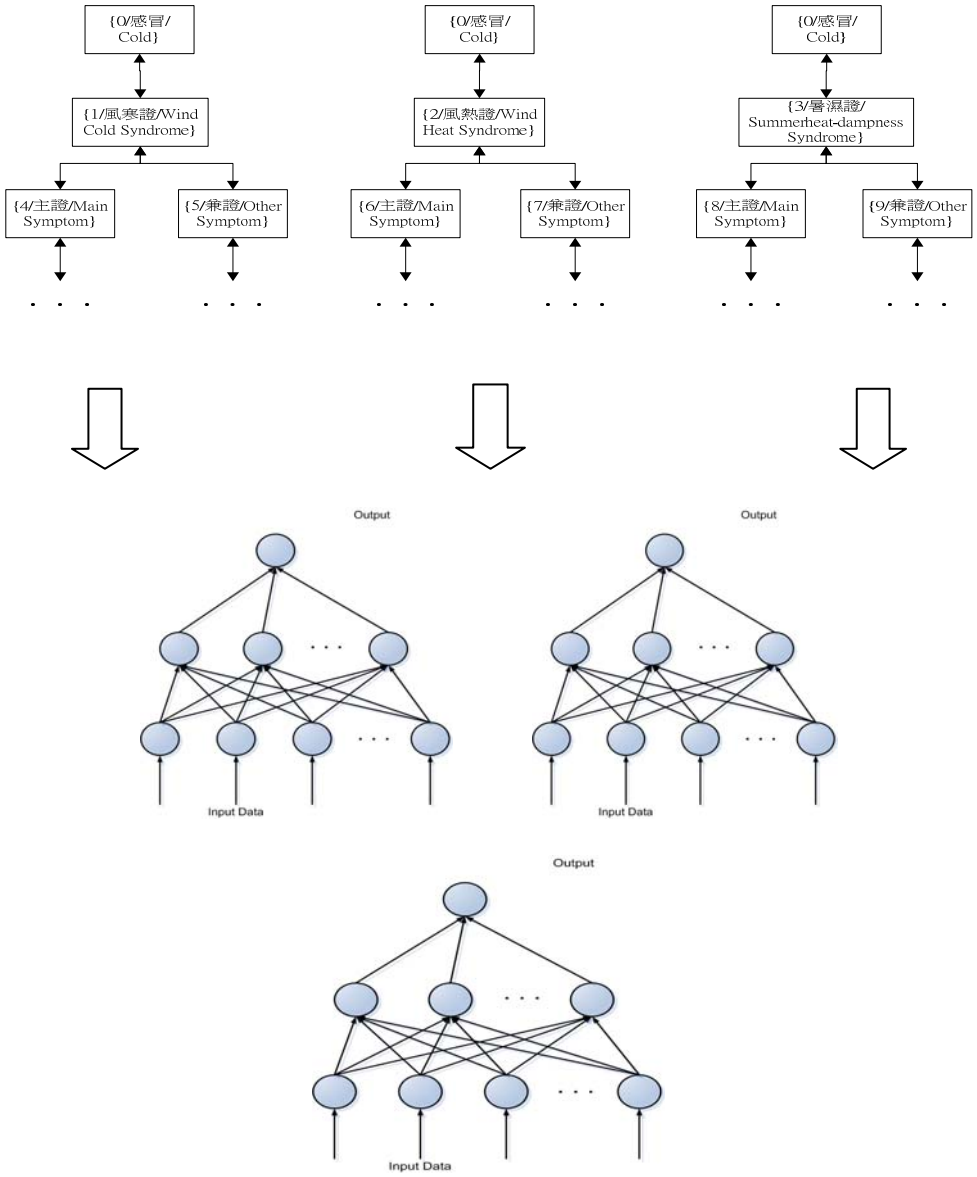


Figure 7.8.2.2.1 Partitioning the NN

The partitions in Figure 7.8.2.2.1 have the following characteristics: i) each tree contains only one illness name and type; ii) number of networks equal to number of illness types; iii) input parameters of each network equal to the symptoms of each illness type (i.e. 1 = exist, 0 = not exist); iv) number of input neurons equal to the number of symptoms of each illness type; iv) number of hidden neurons equal to the twice the number of the input neurons; and v) number of output is one (relevance index).

7.9 Detailed Experimental Results

Many experiments were conducted to verify that the NN (backprogation) is indeed an effective way to reduce amount of effort that we experienced in the algorithmic programming approach. The experimental results unanimously indicate that the NN approach is the right direction to pursue. Some of the results are selected, presented and discussed in details in section 9.5.1 in Chapter 9.

7.10 Recap

In order to gain from the successful experience of Nong's in the implementation of the YOT mobile-clinic systems, the programming of the automatic semantic aliasing (ASA) algorithms in the experiments was based on the Waterfall model. It was found that the repetitive nature of the experiments required tremendous of time in debugging and cross-checking even though the ASA is in the API form. The need to reduce the amount of time and effort in preparing for the experiments the AI approach was explored. The main arguments for the AI investigations are as follows:

a) *Less error prone:* If the NN (backpropagation) architecture is programmed correctly, the logical points (e.g. points "a" and "b" in Figure 7.1.1) will automatically converge to the necessary logical operation (e.g. AND or OR) depending on the training data set.

b) *Generic use:* Since the TCM ontology is made up of many categories of elements, the associations (e.g. light of the relevance indices (RI)) among them are of the combinatorial nature. If the same named NN modules are working on the same set of input parameters to produce their distinctive RI values (i.e. each named element versus the same input set), then the associations among of the name elements versus the same input set can be profiled and examined. This is particular useful for a sizeable knowledge base (e.g. the Nong's enterprise TCM ontology for clinical practice) that has many elements and associations among them.

c) *Weighted input:* The relative significance of every entry in the input set can be weighted to produce the desired outcome.

d) *Pruning:* The execution time of the NN module can be pruned in a dynamic manner, and this is very useful for those time-critical applications.

e) *Rich user experience:* The NN (backpropagation) approach is mature and has very rich user experience that this research can borrow from. In particular, this approach has become a generic tool in many freeware and commercial packages (e.g. the WEKA). As time goes by these tools will certainly be improved by the originator in the package migration course. This usually means less chance of error and more accuracy. It has provided a great attraction for my research for now and the future, for most of the concerns for accurate, usable prototype implementation is obviated.

f) *Ease of parallel computation:* Since the same NN construct can be trained and transformed into different named/dedicated modules. This would ease parallel computation, if it is necessary, tremendously. For example, in the following case, $\{NN_i \mid i = 1,2,...,n; IN\}$, the dedicated NN modules identified by the respective i will work on the same input set of parameters *NN*. If the output for every dedicated NN is identified as $\{RI_i \mid i = 1,2,...,n; IN\}$ in one-to-one correspondence (i.e. $RI_i \Rightarrow NN_i$), then the correlation between and two different dedicated NN can be deduced form their RI outputs. This facilitates Type 2 discoveries, which is the second main drive in this research.

The effort in this chapter has successfully established the fact that the NN approach is indeed suitable for aiding Type 2 herbal discovery.

7.11 Conclusion and Connective Statement

The investigation has concluded that the NN (backpropagation) approach indeed can provide significant advantages such as follows: i) reduction of programming efforts and errors; ii) ease of parallel computation; iii) ease of real-time NN pruning to reduce execution time; iv) ease of weighing relative parameter significance; v) rich user experience; and vi) ready-to-use NN API is commonplace. The next logical step forward is to investigate how to apply the NN approach successfully in adding the Type 2 herbal discoveries.

7.12 Key References

[Bishop95] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995

[Callan03] R. Callan, Artificial Intelligence, Macmillan, May 2003

[Cranefield01] S. Cranefield, S. Haustein, and M. Purvis. UML-based Ontology Modelling for Software Agents, Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada, 28 May – 1 June 2001

[Dillon93] T.S. Dillon and P.L. Tan, Object Oriented Conceptual Models, Prentice Hall, 1993

[Finney94] R.L. Finney, G.B. Thomas and M.D. Weir, Calculus, Addison-Wesley, 1994

[Funahashi89] K. Funahashi, On the Approximation Realization of Continuous Mappings by Neural Networks, Neural Networks, Vol. 2, No. 3, 1989, 183-192

[Gallant92] A.R. Gallant and H. White, On Learning the Derivatives of an Unknown Mapping and Its Derivatives Using Multiplayer Feedforward Networks, Neural Networks, Vol. 5, 1992

[Guarino95] N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, 1995, 25-32

[Hagan96] M. Hagan, Neural Network Design, PWS Publishing Company, 1996

[Hornik89] K. Hornik, M. Stinchcombe and H. White, Multilayer Feedforward Networks are Universal Approximators, Neural Networks, Vol. 2, No. 5, 1989, 359-366

[JOONE] http://www.jooneworld.com/

[JWong08c] J.H.K. Wong, W.W.K. Lin and A.K.Y. Wong, Real-Time Enterprise Ontology Evolution to Aid Effective Clinical Telemedicine with Text Mining and Automatic Semantic Aliasing Support, Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, 11-13 November 2008

[Knuth97] D.E. Knuth, The Art of Computer Programming, Addison-Wesley, 1997

[Kogut02] P. Kogut, S. Cranefield, L. Hart, M. Dutra, K. Baclawski, M. Kokar, and J. Smith, UML for Ontology Development, Knowledge Engineering Review Journal Special Issue on Ontologies in Agent Systems, Vol. 17, No. 1, March 2002, 61-64

[Kramer89] A.H. Kramer and A. Sangiovanni-Vincentelli, Efficient Parallel Learning Algorithms for Neural Networks, Advances in Neural Information Processing Systems 1, 1989, 40–48, ISBN 1-558-60015-9

[Lewis96] T. Lewis, The Next 10000 Years: Part 1, IEEE Computer Society, Vol. 29, No. 4, 1996, 64-70

[Lin04] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, HBP: An Optimization Technique to Shorten the Control Cycle Time of the Neural Network Controller (NNC) that Provides Dynamic Buffer Tuning to Eliminate Overflow at the User Level, International Journal of Computer Systems, Science & Engineering, Vol. 19, No. 2, 2004, 75-84

[Marrone07] P. Marrone, The Complete Guide All You Need to Know about Joone, 2007

[NNWJ] http://www.nnwj.de/backpropagation-net.html

[Sarle97] W. Sarle, Neural Networks Frequently Asked Questions, 1997, ftp://ftp.sas.com/pub/neural/FAQ.html

[WEKA] http://www.cs.waikato.ac.nz/ml/weka/

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

[Wong08] A.K.Y. Wong, T.S. Dillon, and W.W.K. Lin, Harnessing the Service Roundtrip Time over the Internet to Support Time-Critical Applications – Concept, Techniques and Cases (invited and contracted by Nova Science Publishers, Incorporated, New York, Feburary 2008

[Yann98] L. Yann, B. Leon, G.B. Orr and K. Muller, Efficient BackProp, Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, Springer, 1998

Chapter 8 Discovery of Individual Herbal Ingredients

8.1 Introduction

In Chapter 7 the NN (backpropagation) was evaluated as a classification technique for discovering TCM elements. Some of the results from this evaluation are presented in Chapter 8. The conclusion is that the NN approach can indeed effectively aid discoveries of useful herbal ingredients because of its four generic advantages (from point *a* to point *d* below), plus a specific one (i.e. point *e*):

a) It ***converges*** naturally to the logical requirement, which is defining the true meaning for logical points such as *a* and *b* in Figure 7.1.1, with respect to the training dataset.
b) The NN is a ***simple***, ***generic***, ***reusable***, ***less error-prone*** API, which can be invoked anytime and anywhere to be trained with the given set of data so that it can ***predict*** the outcome from new data (outside the training set). The NN is trained for specific application domain, and therefore the outcome depends on the training (e.g. in one case, it is for illness application and for another, it is herbal ingredient discovery).
c) The API approach enhances the **software reliability** because it reduces the amount of possible errors, introduced inadvertently in the traditional programming or software engineering process and/or model (e.g. the Waterfall model) [AI06, DoD03, Goolsby04].

d) The NN performance, in light of speed, can be improved by ***automatic real-time logical pruning***, in which the computation excludes those unimportant NN arcs (e.g. by the Hessian method [Lin04]).

e) The weights of the arcs converge to the "***current knowledge***" in the training dataset. From the angle of an enterprise, any entity, which has never existed (Type 1) before or been applied in such a way (Type 2), is a discovery. Since the NN module is trained only with the prescribed dataset, which is obtained from the enterprise, in an exhaustive manner, training is ***considered completed*** as long as the NN has learned all the "***current knowledge***" buried in the current training dataset. In a practical sense, this ***differs*** from "*training is considered completed only if the NN output (O) value satisfies the condition of* $|O-R| \leq \Delta$ *in a consistently manner*", where R is the reference or teacher signal and Δ is the tolerated percentage error.

The following definitions are noted here to help make the rest of the information presented in this chapter clearer:

a) ***Knowledge base***: This is a representation of specific organized information (e.g. ontology, database, etc.). The information consists of relevant elements such as concepts (*C*), entities/items (*I*), arguments (*A*) and the associations (*As*) among them. It may be represented by $K(C, I, A, As)$ symbolically, which says that the knowledge (*K*) embodiment is the population of the following elements subsets: *C*, *I*, *A*, and *As*. A typical example is the ***ontology*** of a domain or discipline (e.g.

medicine, engineering and hotel management).

b) ***Neural network (NN) learning/training***: This transforms $K(C, I, A, As)$ into the neural network representation defined by $NN(w_{i,j})$, where $w_{i,j}$ is the generic representation of the weight of the j^{th} arc at the i^{th} level of the NN hierarchy. The transformation ($\rightarrow$) process is governed by the adopted function (e.g. Sigmoid); then, it means $K(C, I, A, As) \underrightarrow{sigmoid} NN(w_{i,j})$.

c) ***Knowledge classification***: Logically, this is "$< CL, S > \in K; < S \mid CL >; S \in CL$". It says that the sets *CL* and *S* both are in the knowledge domain K. The $S \mid CL$ expression says that the set S on the left of "|" must satisfy some or all of the specified requirements (SR) for CL on the right; so, if *S* belongs to one of the pre-defined classes of well-defined SR. Logically, $SR \mid CL$ is valid.

d) ***Knowledge discovery***: In this thesis, two types of discoveries are dealt with:

   i) ***Type 1***: Set P is not part of K (i.e. $P \notin K$) but it does possess all the attributes of one of the predefined classes in the set of pre-defined classes *CL*. Then, logically it is "$< P \in CL > \wedge < P \notin K >$", where $\wedge$ is the logical "AND". This means that ***P is a new occurrence*** with respect to the extant K. This phenomenon is common in open web-based text mining; and it is the way to enliven the otherwise closed ontology for real-time evolution, supported by the technique

of *automatic semantic aliasing* (Chapter 6 – Living Ontology, Semantic Aliasing and Relevance Index). Then, P in the context of this thesis is a new occurrence or ***discovery***. The newly discovered P, however, is not physically incorporated into the extant ontology immediately, for it can only be included later via the process of consensus certification. Actually, all newly discovered items are temporarily appended in the reserved buffer areas in the system, and their associations (or similarity) with the pre-defined "host" (e.g. illness, herbs, etc.) in the extant ontology are clearly defined by the relevance indices (to the "host"). If we assume that the two sets P and Q are defined by the attributes s_i from the $\Omega = \{s_i\}$ population for $i = 1,2,..., j,..,n$ (i.e. $P(s_i)$ and $Q(s_i)$) but $P(s_i) \neq Q(s_i)$, logically, it means $P \cap Q \neq \Theta$, where $\cap$ for the intersection of the two sets and Θ for a null set. If $P_r(P \cup Q) = P_r(P) + P_r(Q) - P_r(P \cap Q)$ holds, it means $P_r(PQ) = P_r(P \cap Q) = P_r(P) + P_r(Q) - P_r(P \cup Q)$, where (P_r indicates the probability). Conceptually the *relevance index* (RI) is equal to $P_r(PQ)$ or $P_r(P \cap Q)$. Yet the RI should be computed with respect to the angle of the "host" reference (Chapter 6).

ii) ***Type 2***: For $\Omega = \{s_i\}$ and $i = 1,2,..., j,..,k,...,n$ the following are true logically: $CL_j = (s_j)$; $CL_k = (s_k)$; $CL_j \neq CL_k$; and

$< CL_j, CL_k > \in K$. The expression $CL_j = (s_j)$ says that the class CL_j is defined by the set of attributes represented by s, where the subscript j (i.e. s_j) marks the particular of s. Normally, $P_r(CL_j CL_k) = \Theta$ indicates that the two classes CL_j and CL_k are ***logically independent***. If both the $< X \in CL_j > \wedge < X \in CL_k >$ and $X \in K$ conditions are logically ($\wedge$ for logical AND) satisfied by the set X, then X is a discovery provided that either/both of the $< X \in CL_j >$ or/and $< X \in CL_k >$ association were not previously listed in K. This discovery is not ***Type 1***, for $X \in K$, but is a hidden association in K only. $P_r(UV) = P_r(U \cap V) = P_r(U) + P_r(V) - P_r(U \cup V)$ holds for $U = CL_j = (s_j)$ and $V = CL_k = (s_k)$. Again, conceptually $P_r(UV)$ or $P_r(U \cap V)$ is the relevance index (RI) or the degree of similarity. The actual RI value, however, depends on which is the referential host (details in Chapter 6). Classification of information in general can be achieved in many ways. For example, it is common in the TCM area to divide the different symptoms of an illness (e.g. Flu (感冒)) into different classes or subclasses (證型) to facilitate treatment. In classic TCM, Flu can have three classes of symptoms: ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證) and ***Summer-Heat Dampness***

Syndrome (暑濕證). These classes or syndromes require very different herbal ingredients for treatment. Sometimes a herbal ingredient may treat two classes with different degrees of efficacy (e.g. **Gancao** (甘草)). Then, logically speaking $< X \in CL_j > \wedge < X \in CL_k >$ holds for X = **Gancao**, CL_j = "***Wind Heat Syndrome (風熱證)***", and CL_k = "***Wind Cold Syndrome (風寒證)***" . The RI computed when CL_j is chosen as the referential host (RH) differs from that for CL_k as the RH instead. For example, if only CL_j = "***Wind Heat Syndrome (風熱證)***" were encoded in the extant TCM ontology but not CL_k = "***Wind Cold Syndrome (風寒證)***", then the revelation of the latter is a ***Type 2*** discovery in the context of this PhD research.

This discussion in this chapter is focused mainly on ***Type 2*** discovery. The aim is to verify that the proposed NN (backpropagation) approach indeed has no problem of discovering new associations that are not encoded in the extant TCM ontology chosen for the experiments; this ontology was created by the process of consensus certification, which was commissioned by the proprietor (i.e. Nong's).

When the prototype of a NN model is first generated, it consists of a bare-bones *untrained* backpropagation neural network. By feeding the raw clinical data (obtained from the original clinical Nong's D/P system) into the

NN model, all necessary explicit axiomatic definitions of the logical points in a knowledge subsumption hierarchy (e.g. Figure 8.1.1) will converge automatically through training. The NN is able to learn and acquire the knowledge embedded in the set of training data and converge automatically to the equitable axiomatic definitions (e.g. points *a* and *b* in Figure 8.1.1). The NN approach itself is naturally generic. Its true meaning is determined by the domain of application at the time, and this true meaning is represented by the given training dataset. Intrinsically the NN does not differentiate the origin of the knowledge source. For example, the source of the training data in telemedicine research may come from one of the following types of knowledge sources or ontology setups (e.g. [Fayyad96, Berners98, Berners01]):

a) *Global* – This contains the total formal knowledge of a domain. For example, the global TCM ontology is made up of all the classics, treatises and case histories accumulated over a few thousand years.
b) *Enterprise* – This is a subset of the global ontology isolated to suit the purpose of a company, enterprise and/or establishment. The current Nong's proprietary TCM ontology, which supports ***clinical practice***, is a "local" (to Nong's) standard/vocabulary. Its aim is to effectively support correct communications and interoperability within the company and its collaborating partners.
c) *Local of the local* – Nong's creates/customizes different D/P variants with respect to the specifications of individual clients. Every customized target system would have only a subset of the Nong's enterprise ontology for the designated *in-situ* operation. The system variants are,

however, interoperable to a varying degree, depending on the contents of their local/customized TCM ontology.

The herbal discoveries in this research are mainly based on clinical data, and that is why the verification experiments made use of the Nong's TCM enterprise ontology (with permission). Figure 8.1.1 is actually an example of the subsumption hierarchy for a subset of the Nong's enterprise ontology. The points *a* and *b* are axiomatic points that can assume any logical meanings (e.g. AND, OR, or EXCLUSIVE OR). With respect to the set of training data, the actual logical meanings would converge with respect to the given dataset only. This is the form of machine learning/training argument put forth for this thesis.

The experimental results of chapter 7 strongly support the argument that the NN (backpropagation) is a suitable approach for knowledge classification, which would in its turn provides an effective basis for knowledge discovery. In this chapter, the focus is on investigating how to make use of NN for Type 2 discovery.

Before discovery can begin, the NN must be trained with the given set of data. As a result the discoveries are limited within the ambit of the knowledge contained in training data set. In principle, a discovery is made provided that the given parameter (P) possesses a set of attributes that match that predefined for a "class (C)". Yet, P is not an element in the knowledge (K) base from which different classes are derived. If $CS < C_1, C_2, ...C_n >$ represent the set of classes for K, the ($P \in C \mid CS$), where $\in$ means "belong to" and "|"

means the left-hand side satisfies the condition(s) stated in the right-hand side. In this light $P \in C \mid CS$ says that P belongs to the class C, which is one of those defined for CS [Fayyad96].

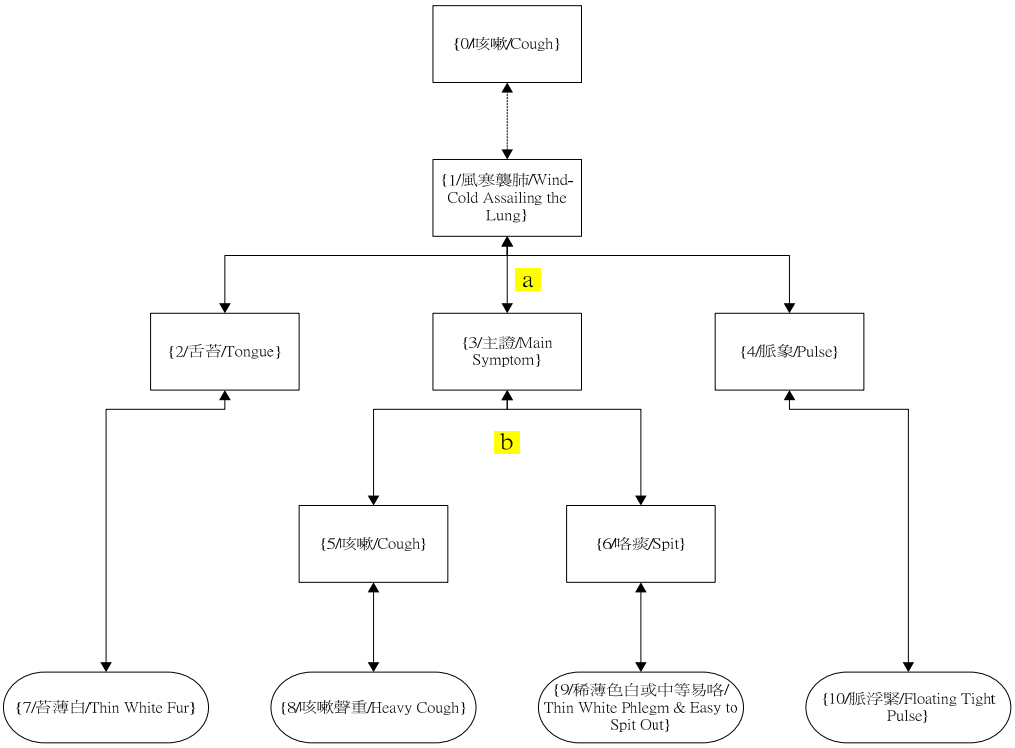


Figure 8.1.1 Organization of raw clinical data (illnesses) in the original Nong's D/P system ("0" and "5" are logically different - "0" is an illness name and "5" a symptom)

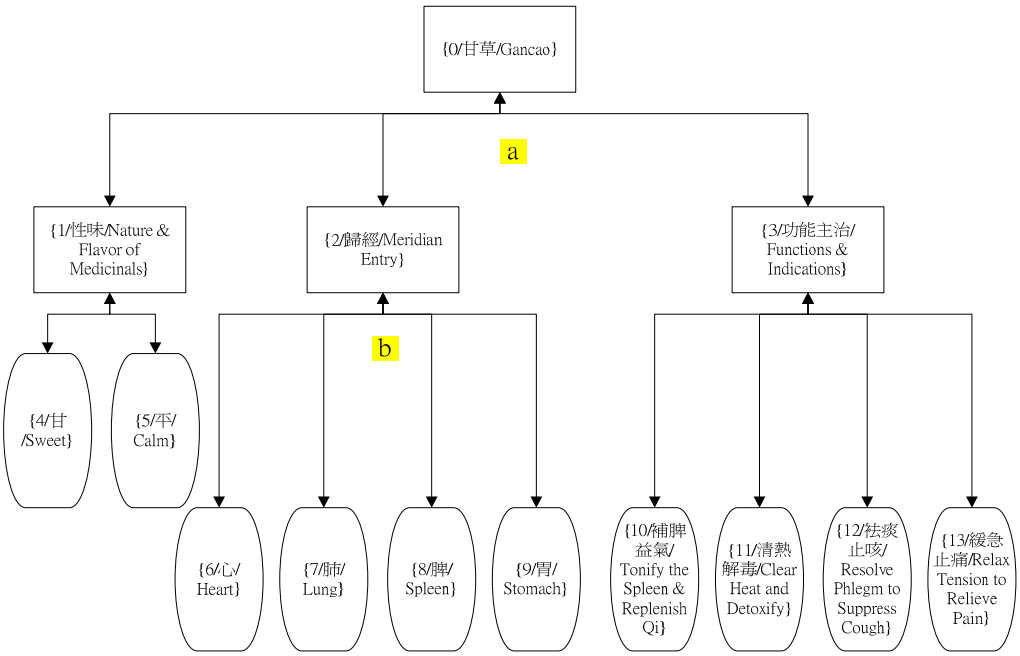


Figure 8.1.2 Logical accentuation from the physical organization of raw clinical data (herbal items) in the Nong's TCM enterprise ontology

Before discoveries of herbal ingredients by NN can proceed, the herbal ingredients should be actuated (as tree roots). This accentuation is purely logical and does not physically affect the organization of the ontology (or database) underneath. The organization is basically a subsumption hierarchy of the concepts, entities (e.g. herbs and symptoms), and the axiomatic relations/associations among them. Figure 8.1.2 is an example of such logical accentuations from the ontology.

In this example (Figure 8.1.2), there are fourteen clinical entities uniquely identified. In the set {0/甘草/Gancao} the unique symbols, "0" , "甘草" and "Gancao" have the same connotation – "0" as the identifier. The arcs/interconnections/paths are bi-directional because their traversals should be logically transitive. The entities can be placed anywhere within the ontology

(realized in a form of a database), but their retrievals depends on the "*retrieval algorithm (RA)*" implemented, as indicated in Chapter 7 (sections 7.1 and 7.2).

8.2 Preparation for Experiments

Many experiments were conducted to verify that if the backpropagation NN approach can indeed be applied to herbal ingredient discoveries. The setup for the experiments will be discussed in the next section. These experiments also make use of the Nong's proprietary enterprise TCM ontology, but concentrate on only three illnesses' sub-ontologies (also sizeable), namely, Influenza, Inability to Sleep (Insomnia) and Constipation. From the same Nong's TCM ontology, sixty herbal items sub-ontologies were isolated so that their degrees of relevance (i.e. RI values) to the three illnesses can be reasoned by the NN artificial intelligence.

8.2.1 The Setup for NN Verification Experiments

Figure 8.2.1.1 depicts the environment for the NN verification experiments. The details in the box "***Herbal Items File (Global)***" are classical TCM information, which forms the global TCM ontology.

The global TCM ontology provides the standard TCM vocabulary reference anytime, anywhere (e.g. [WHO07]). From the global ontology, the proprietary the Nong's or PuraPharm enterprise TCM ontology was created. From this enterprise TCM ontology different diagnosis/prescription (D/P)

system versions (variants) for mobile-clinic operation have been customized. Compared to the global TCM ontology, the Nong's enterprise TCM ontology is considered a local (master) version. The TCM ontology customized for the different system variants are at the "*local of the local*" level. Within the box "Neural Network (Enterprise)" the NN (backpropagation) is ***trained*** with the Nong's dataset of patients' cases. This data set represents the "***current clinical knowledge*** (CCK)" of the enterprise. Any entity that does not exist in the CCK is a discovery from Nong's perspective. It is a ***Type 1*** discovery if the entity is an herbal ingredient and a ***Type 2*** if it is new usage for an entity already present in CCK. Since the requirement of the NN training is to ensure that the weights of the NN arcs would represent the CCK, the learned NN is called the ***well-trained NN (enterprise)*** in this thesis.

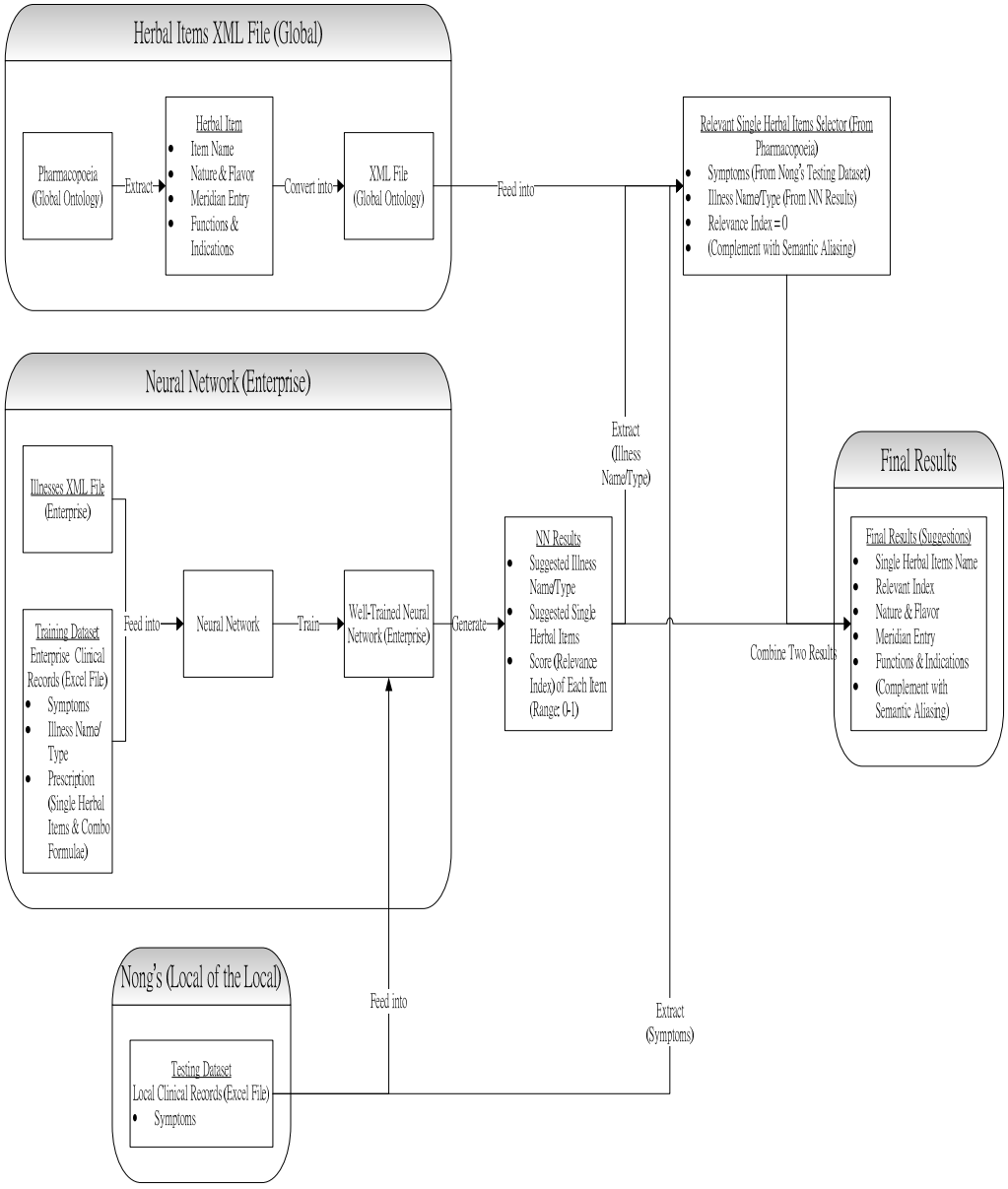


Figure 8.2.1.1 Setup for the NN verification experiments - herbal discoveries (Type 2)

The box "***Nong's (local of the local)***" says that any "local of the local" patients' cases can be used as input for the NN to reason and discover. The results produced by the reasoning process would be put into the temporary "***NN Results***" file. Since every NN module is now dedicated to specific purpose (e.g. an herb, an illness, or an illness type) through training, the contents for the "***NN***

Results" depends on the NN module. If the contents represent a discovery, then it will be combined with relevant information extracted from the global ontology as well as the clinical records (local of the local) before the "***Final Results***" record is created. This record will be stored temporarily and may eventually be incorporated into the enterprise TCM ontology via the rigorous process of consensus certification.

8.2.1.1 NN Training

Training means the NN module learns to carry out a specific function. For example, for the 60 herbs chosen for the NN experiments (Table 9.4.2.1) there should be 60 NN modules, one for every herb (e.g. **Gancao** (甘草) or Radix Glycyrrhizae). Similarly we can also train the NN module to reason for a particular illness type (e.g. ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證), or ***Summer-Heat Dampness Syndrome*** (暑濕證)). *To help the reader to quickly feel the complexity of the experiments, Table 9.4.2.1, which contains both pinyin and Latin indications/translations, is shown in the appendix II. In the subsequent experiments, either pinyin, Chinese characters is used in poltting graphs for analytic puroposes*.

Training in the context of this thesis is defined as $< Cases(s_i) \mid Subject >$ logically; a patient's case (i.e. *Case*) must satisfy (i.e. "|") the definition of *Subject*. The parameter s_i for $i = 1,2,...j,...,k,....,n$ is the set of symptoms for the NN to reason with respect to the indicated *Subject*. The

meaning of *Subject* is user-defined, and if we train a NN module for the herb Tian Hua Fen (天花粉) or Radix Trichosanthis (i.e.天花粉 NN module), then it means $Subject = $天花粉. In the training process all the cases that include Radix Trichosanthis (天花粉) for treatment, independent of their illness type (i.e. 證型), will be used to train the target Radix Trichosanthis (天花粉) NN module. The training is considered complete when all the relevant cases in the training dataset are exhausted. Then, the Radix Trichosanthis (天花粉) NN module has acquired the "***current clinical knowledge (CCK)***" with respect to the training dataset and is ready for use. The arc weights in the NN have settled down to the respective values for CCK. In a similar manner the NN module dedicated to ***Wind Cold Syndrome*** (風寒證) can be trained with cases that satisfied the criterion: $< Cases \mid Subect = WindColdSyndrome >$.

8.2.1.2 NN Application

The application of a well-trained NN module is straightforward, and the aim is to satisfy $< Cases(s_i) \mid Subject(p_j) >$ logically, for $j = 1,2,...,k,...,l,....,n$. Basically patients' clinical cases defined by specific sets of symptoms s_i and, at the same time, have satisfied the indicated set of parameters p_j will be accepted as input by NN for meaningful reasoning/inference to produce the respective RI values. For example, if $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) is the case, then the Radix Trichosanthis (天花粉) NN module will reason only with those clinical records/cases that have

included the two entities: ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證). In this research there is ***convention*** pre-defined for the NN application; the first entry in the Subject set invokes the corresponding named NN model. For example, $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) invokes the 天花粉 NN module to reason and produce the RI value as the outcome. The semantics of this RI outcome is "the collective relevance of the syndromes extracted from the patients' cases to the named herbal ingredient Radix Trichosanthis (天花粉)." The second parameter (e.g. ***Wind Cold Syndrome*** (風寒證)) has no contribution to the RI calculation; it is simply is an ***inert parameter***. It helps the plotting of the 3-dimensional RI view for visualization purposes (e.g. the Figures 9.4.2.1.2.14a and 9.4.2.1.2.14b in Chapter 9).

8.2.1.3 Type 2 Discovery

In the following discussion we postulate what a Type 2 discovery means. Firstly, let us to reiterate the following facts:

a) ***Convention***: The $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) syntax means that the 1st parameter specified in Subject invokes the ***named/dedicated*** NN module (i.e. the 天花粉 module) to produce the corresponding RI value. The second parameter is inert but helps to provide a 3-dimensional view (e.g. the example of 60 herbs versus 10 illness types in Figure 9.4.2.1.2.14a in

Chapter 9).

b) ***"Well-trained" named NN module***: Training is represented by the following syntax, where $Cases(s_i)$ represent those clinical cases that are each defined by a unique set of symptoms (i.e. syndrome) namely s_i: $< Cases(s_i) \mid Subject >$. For $< Cases(s_i) \mid Subject =$ ***Radix Trichosanthis /*** (天花粉)>, only the named ***Radix Trichosanthis (***天花粉) NN module will be invoked for training. When training is completed, the well-trained NN module has acquired all the "***current clinical knowledge*** (CCK)" embedded in the training dataset TD). If a subset of cases is randomly extracted from TD, namely TD_{sub}, the application invocation by the syntax $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) would generate all the RI value of "syndromes versus the main subject ***Radix Trichosanthis*** (天花粉)" but concentrated only on the illness type, Wind Cold Syndrome (風寒證). This can be illustrated by the simple artificial Table 8.2.1.3.1.

(Herbal names in Table 9.4.2.1 in the Appendix II)	天花粉 (Tian Hua Fen)	酸棗仁 (Suan Zao Ren)	合歡皮 (He Huan Pi)	薏苡仁 (Yi Yi Ren)
風寒證 (Wind Cold Syndrome)	RI=0.00	RI=0.31	RI=0.00	RI=0.02
風熱證 (Wind Heat Syndrome)	RI=0.47	RI=0.44	RI=0.00	RI=0.37
暑濕證(Summer-Heat Dampness Syndrome)	RI=0.77	RI=0.92	RI=0.81	RI=1.00

Table 8.2.1.3.1 Artificial table to mimic the RI table of the application, a well-trained NN module (TD_{sub} assumed as the input)

c) ***Actual NN application***: The TD_{sub} input in the above artificial example postulates the following: i) *Radix Trichosanthis* (天花粉) is not used in

Wind Cold Syndrome (風寒證) (RI=0); ii) *Cortex Albiziae* (He Huan Pi / 合歡皮) is not used in both Wind Cold Syndrome (風寒證) and Summer-Heat Dampness Syndrome (暑濕證) (both RI values equal to 0) as far as the current CCK is concerned. Now, for example, the syntax is invoked: $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (合歡皮); ***Wind Cold Syndrome*** (風寒證)), and the set of input cases (e.g. AP_{set}) is new and contains the "local of the local" type of cases (e.g. from one of the D/P system variants in the filed). Table 8.2.1.3.2 would be the kind of RI values produced by the different NN modules (e.g. *Cortex Albiziae* / 合歡皮). Table 8.2.1.3.3 indicates that there is a Type 2 discovery.

(Herbal names in Table 9.2.5.2.1 in the Appendix II)	天花粉 (Tian Hua Fen)	酸棗仁 (Suan Zao Ren)	合歡皮 (He Huan Pi)	薏苡仁 (Yi Yi Ren)
風寒證 (Wind Cold Syndrome)	RI=0.00	RI=0.31	***RI=0.63***	RI=0.02
風熱證 (Wind Heat Syndrome)	RI=0.47	RI=0.44	***RI=0.74***	RI=0.37
暑濕證(Summer-Heat Dampness Syndrome)	RI=0.77	RI=0.92	RI=0.81	RI=1.00

Table 8.2.1.3.2 Artificial table to mimic the RI table of the application, a well-trained NN module (TD_{sub} assumed as the input)

	合歡皮 *Cortex Albiziae*
風寒證 Wind Cold Syndrome	***RI=0.63***
風熱證 Wind Heat Syndrome	***RI=0.74***
暑濕證 Summer-Heat Dampness Syndrome	RI=0.81

Table 8.2.1.3.3 Type 2 discovery

In the CCK knowledge, *Cortex Albiziae* (合歡皮) was never used to treat Wind Cold Syndrome (風寒證) and Wind Heat Syndrome (風熱證), but the AP_{set} cases of the "local of the local" type indicate the

possibilities clinically. Therefore, from the enterprise's perspective ***"****Cortex Albiziae(合歡皮)* ***can treat Wind Cold Syndrome (風寒證) and Wind Heat Syndrome (風熱證)"*** is a Type 2 discovery. This new clinical knowledge may be incorporated into the Nong's enterprise (master/local) TCM ontology in the next round of consensus certification by a sufficient number of medical experts.

8.2.2 Experimental Results and Data Analysis

Many experiments were conducted with the philosophy encoded in Figure 8.2.1.1. The experimental results unanimously concluded that the NN (backpropagation) approach is the right direction to attain Type 2 discovery. For two independent classes CL_j and CL_k, if both the $< X \in CL_j > \wedge < X \in CL_k >$ and $X \in K$ conditions are logically satisfied, then X is a Type 2 discovery provided that either/both of the $< X \in CL_j >$ or/and $< X \in CL_k >$ association were not previously listed in K; K is the knowledge base or ontology.

For the verification experiments, 60 herbal ingredients were selected against 10 illness types, which belong to 3 illnesses. For example, Flu (感冒) has three illness types (證型) that are classified according to three sets of symptoms (syndromes): ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證) and ***Summer-Heat Dampness Syndrome*** (暑濕證). The three illness types require very different herbal ingredients for effective treatment.

The syntax, $< Cases(s_i) \mid Subject(p_j) >$, for $j = 1,2,...,k,...,l,....,n$ is porposed for invoking the named NN module. If $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) is the case, then the ***Radix Trichosanthis (***天花粉) NN module will be invoked to reason only with those clinical records/cases that have included the two entities: ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證). The proposed ***convention*** for the invocation is:

i) The first entry in the Subject set invokes the corresponding named NN model; $Subject(p_j) = Subject($ ***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) invokes the ***Radix Trichosanthis (***天花粉) NN module to reason and produce the RI value as the outcome. The semantics of this RI outcome is "*the collective relevance of the syndromes extracted from the patients*' cases to the named herbal ingredient 天花粉 (*Radix Trichosanthis*).

ii) The second entry/parameter (e.g. ***Wind Cold Syndrome*** (風寒證)) has no contribution to the RI calculation, but this ***inert parameter*** helps the plotting of the 3-dimensional RI view for visualization and interpretation purposes (e.g. Chapter 9, section 9.4.2).

Therefore, there are 60 named "herbal" NN modules and 10 "illness types" NN modules all together (i.e. 70 total) for the verification experiments; for example, the "天花粉" (*Radix Trichosanthis*) NN module. The focus of

these experiments is to explore how to establish the foundation whereby Type 2 discoveries can be achieved. The steps for every experiment are:

i) ***NN training:*** The named NN module is trained with a few hundred real patient medical records/cases. The aim is to let the NN learn and acquire the "***current clinical knowledge***" (CCK). All the 70 named NN modules are trained with the same dataset purely in the form of "symptoms versus the referential herbal ingredient that the NN module is named after", independent of the illness types. As a result, every NN has absorbed the same CCK as the basic knowledge, stored as arc-weights in the "*well-trained*" or *learned* NN. In this thesis, we call this learning outcome: "***convergence to the CCK***".

ii) ***RI computation:*** Every NN computes the collective RI of "symptoms versus the referential herb - its own name". These RI values reflect the relevance of the herbal ingredient against the set of symptoms embedded in the CCK. Then, a 3-dimensional plot (e.g. the plot in Chapter 9, section 9.4.2) is produced as the basis for exposing possible Type 2 discoveries. This 3-dimensional plot is called the ***basic referential plot*** (BRP), and its size MxN is scalable; M is the number of herbal ingredients and N is the number of illness types or classes.

iii) ***Type 2 discoveries:*** A set of "***raw***" field medical records of the "*local of the local*" nature are selected, and the symptoms that contain within this set act as the input to excite the NN module. Normally, these field records are not part of the training dataset. The NN would first filter out those records that do not contain the herb that the NN module is named

after (e.g. 天花粉 (*Radix Trichosanthis*)). Then, the collective RI value is computed to reflect the relevance between the raw input set of target symptoms and the referential herbal ingredient - name of the NN module. For this raw data set the "***mini referential plot***" (MRP) is produced and scrutinized. If there is RI discrepancy between the same herb between the BRP and MRP, a Type 2 discovery is a possibility.

In order to evaluate the execution time of the named NN modules, timing analysis was conducted. The execution time is measured in clock cycles because the number of clock cycles would remain the same even a faster CPU is involved. Yet, the physical time would change with respect to the physical clock speed. The formula for computing the physical execution time is $ExecutionTime = \frac{NumberOfExecutionClockCycle}{ClockRate}$. This is an important concept because the number of clock cycles needed to execute a named NN modules does not depend on the speed of the clock rate of the computer platform. In effect, the speed of the NN execution scales proportionally with the clock rate. In the section 9.4.2 of Chapter 9 a set of selected experimental timing analyses are presented.

In fact, one of the main reasons for selecting the NN approach for Type 2 herbal discoveries is that the NN speed can be tuned in a real-time manner [Lin04]. To demonstrate this point some experimental results are also presented in section 9.4.2 of Chapter 9. Real-time logical NN pruning will eliminate unimportant arc-weight computations logically and therefore reduce the

execution time. Logical pruning means that the process would not actually cut the NN arcs but rather ignore them it for the next round of NN computations. In this light, every NN module works in a twin mode: one Chief NN modules and one Learner NN module. Theoretically, after the learner has finished training, it swaps position with the Chief, which becomes the learner to acquire new knowledge. Pruning occurs immediately after the learner has completed its training session but before assuming the position of the Chief. That is, the new Chief would compute faster due to its logically pruned skeletal structure.

8.3 Recap

This chapter presents how the NN (backpropagation) approach, which was established in Chapter 7 as a classification technique, can be applied for the Type 2 discoveries of individual herbal ingredients. Since the NN module is trained only with the prescribed dataset, training is ***considered completed*** in the context of Type 2 discovery as long as the NN has learned all the "***current knowledge***" buried in the current training dataset. This ***differs*** from the usually adopted criteria for supervised learning: "*training is considered completed only if the NN output (O) value satisfies the condition of* $|O-R| \leq \Delta$ *in a consistently manner*", where R is the reference or teacher signal and Δ is the tolerated percentage error. There is ***convention*** pre-defined for the named NN applications. The $Subject(p_j) = Subject($***Radix Trichosanthis*** (天花粉); ***Wind Cold Syndrome*** (風寒證)) syntax invokes the ***Radix Trichosanthis (***天花粉) NN module to reason and produce the RI value as the outcome. That is, the first entry in the *Subject* set invokes the corresponding named NN model. The

semantics of the RI outcome is: "*the collective relevance of the syndromes extracted from the patients' cases to the named herbal ingredient* ***Radix Trichosanthis (***天花粉)". The second parameter (e.g. ***Wind Cold Syndrome*** (風寒證)) has no contribution to the RI calculation but is simply an ***inert parameter***, which helps the plotting of the 3-dimensional RI view for visualization purposes (e.g. the Figures 9.4.2.1.2.14a and 9.4.2.1.2.14b).

8.4 Conclusion and Connective Statement

This thesis deals with two types of discoveries:

i) ***Type 1***: Set P is not part of K (i.e. $P \notin K$) but it does possess all the attributes of one of the pre-defined classes in the set of pre-defined classes *CL*. Then, logically it is "$< P \in CL > \wedge < P \notin K >$", where $\wedge$ is the logical "AND". This means that ***P is a new occurrence*** or discovery with respect to the extant K. If we assume that the two sets P and Q are defined by the attributes s_i from the $\Omega = \{s_i\}$ population for $i = 1,2,..., j,.., n$ (i.e. $P(s_i)$ and $Q(s_i)$) but $P(s_i) \neq Q(s_i)$, logically, it means $P \cap Q \neq \Theta$, where $\cap$ for the intersection of the two sets and Θ for a null set. If the following logical relationship $P_r(P \cup Q) = P_r(P) + P_r(Q) - P_r(P \cap Q)$ prevails, it means $P_r(PQ) = P_r(P \cap Q) = P_r(P) + P_r(Q) - P_r(P \cup Q)$, where P_r denotes the probability. Conceptually the *relevance index* (RI) from the angle of the referential host (Chapter 5) is equal to $P_r(PQ)$ or $P_r(P \cap Q)$.

ii) ***Type 2***: For $\Omega = \{s_i\}$ and $i = 1,2,..., j,..,k,...,n$ the following are logically true: $CL_j = (s_j)$; $CL_k = (s_k)$; $CL_j \neq CL_k$; and $< CL_j, CL_k > \in K$. $CL_j = (s_j)$ says that the class CL_j is defined by the set of attributes represented by s, where the subscript j (i.e. s_j) marks the particular of s. If both the $< X \in CL_j > \wedge < X \in CL_k >$ and $X \in K$ conditions are logically ($\wedge$ for logical AND) satisfied by the set X, X is a discovery if either/both of the $< X \in CL_j >$ or/and $< X \in CL_k >$ association were not previously listed in K. This discovery is not ***Type 2***, for $X \in K$.

This chapter (Chapter 8) is dedicated to the investigation of Type 2 discovery by using the NN (backpropagation) approach. The experimental results so far have confirmed that this approach is the right direction to follow. The next logical step is to include some of the verification results in light of discoveries of individual herbal ingredients in the next Chapter (Chapter 9).

8.5 Key References

[AI06] Enterprise Ontology, AIAI, Artificial Intelligence Application Institute, April 2006, http://www.aiai.ed.ac.uk/project/enterprise/enterprise/omtology.html

[Berners98] T. Berners-Lee, Semantic Web Road Map, 1998, http://www.w3c.org/DesignIssues/Semantic.html

[Berners01] T. Berners-Lee, H. James and L. Ora, The Semantic Web, Scientific American Magazine, 17 May 2001

[DoD03] Criscimagna, NH2003, Interoperability, Vol. 10, Reliability Analysis Center (US DoD Information Analysis Center), 1-16

[Fayyad96] U.M. Fayyad, S.G. Djorgovski and N. Weir, Automating the Analysis and Cataloging of Sky Surveys, in Advances in Knowledge Discovery and Data Mining, eds. Y.M. Fayyad, AAAI/MIT Press, 1996

[Goolsby04] K. Goolsby and F.K. Whitlow, What Causes Outsourcing Failures? Outsourcing Journal, 2004, http://www.outsourcing-journal.com/aug2004-failure.html

[Lin04] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, HBP: An Optimization Technique to Shorten the Control Cycle Time of the Neural Network Controller (NNC) that Provides Dynamic Buffer Tuning to Eliminate Overflow at the User Level, International Journal of Computer Systems, Science & Engineering, Vol. 19, No. 2, 2004, 75-84

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

Chapter 9 Walkthrough of Selected Experimental Results

9.1 Introduction

As mentioned earlier, the aim of this thesis is to discover useful TCM herbal ingredients and prescriptions. In fact, every prescription contains several herbal ingredients that assume various roles: *principal*, *courtier*, *assistant* and *messenger*. The discoveries are meaningful, however, if and only if the following issues are satisfactorily addressed: i) the TCM knowledge base (in this case is the given TCM onto-core) is axiomatic and unambiguous; ii) the knowledge base embeds correct TCM semantics; iii) data mining starts working with the customized knowledge core, which gradually and automatically evolves (the living ontology concept); iii) the parser must deal with queries correctly in the sense that every query has the corresponding semantic path in the semantic net; and iv) semantic transitivity must exist among the three layers in the WD^2UHI based prototype, namely, the query system (i.e. GUI) in the top layer, semantic net in the middle layer (parser is also part of this layer), and TCM onto-core in the bottom layer.

In order to create a trusted WD^2UHI based prototype with minimum error-prone human intervention, a semantically based automatic software system generator is needed. The goal is to generate the target system directly from the named TCM onto-core for the given user specification. Automatic system generation in the context of this research is complete in the sense that the customized system will also be remotely installed/deployed for validation

and test run immediately over the Internet. The issue of remote system installation involves reliable/trusted communication over the mobile Internet (wireless and wireline). It is, however, not the focal research in this thesis; it was scrutinized only in the preliminary investigation at the start of the thesis to help define the focal research ambit and objectives more clearly.

As a recap, this thesis has the following major objectives:

a) ***First objective:*** Contributed to a solution for the development of a trusted WD^2UHI platform that has a standard, unambiguous, "living" and semantically-transitive TCM onto-core to support local and global interoperability.
b) ***Second objective:*** Proposed solutions to discover herbal ingredients and prescriptions correctly and meaningfully.

The solutions proposed to achieve the two main objectives above are as follows:

a) ***Trusted prototypes***: The ***EOD-ISD approach*** for automatic generation of prototypes from the respective iconic specifications was proposed, verified and used in the research. This approach is the extended version of the original MI concept proposed by Nong's. The EOD-ISD mechanism generates the target system in a semantic manner in the sense that the local TCM onto-core is customized from the named/given "source onto-core". The prototypes generated for experiments in this

research were all customized from the Nong's master/enterprise TCM onto-core for telemedicine clinical practice, with permission. Therefore, all the proposed solutions can be verified in the real TCM clinical environment. The relevant publications that discuss the significance of the contributions by the proposed solutions in this category include: p6, p8, p9 and p13.

b) ***Living ontology and standardization***: The living ontology concept is realized by the OCOE part of the novel OCOE&CID (***O****n-line* ***C****ontinuous* ***O****ntology* ***E****volution and* ***C****linical* ***I****ntelligence* ***D****iscovery*) approach. An open evolvable TCM onto-core is physically the combination of "*a closed customized skeletal or intrinsic TCM onto-core plus the master aliases table (MAT) contents*". The MAT contents are in four special data structures that can be appended or disconnected to or from the skeletal TCM onto-core manually. As a result, the new knowledge updates in the open evolvable onto-core affect only the MAT contents and not the skeletal/intrinsic onto-core information, which was customized from the master enterprise onto-core given. The MAT contents are necessary for supporting the unique CID part of the OCOE&CID mechanism – the semantic aliasing operation, which *standardizes* new information *text-mined* from open sources (e.g. open web and handwritten D/P information). Standardization is the prelude to extensive discoveries of new prescriptions (high-level) and individual herbal ingredients (low-level). The relevant publications that reveal the significance of contributions by the proposed solutions in this category include: p3 and p4,

c) ***Discoveries***: ***Two principles*** were proposed and adopted: i) the "同病異治，異病同治" [WHO07]" in Chinese terminology or SIMILARITY/SAME (i.e. the Chinese "同") principle in English [JWong09a] in the TCM domain; and ii) the logical axiomatic relationship. These two principles were applied (both explicitly and implicitly) in both the high-level and low-level discoveries; high-level discoveries are basically algorithmic and the low-level discoveries are based on the NN approach. Those publications that present the contribution significance by the proposed solutions in this category include: p7 and p10.

9.2 Experimental Results

In order to provide a clear visualization for the reader, pinyin, Latin, Chinese and or English will be used in the presentation of experimental results. The choice should, from our point of view, provide the necessary clarity.

9.2.1 RDF and OWL Verifications

Many experiments were conducted to verify the suitability of RDF and OWL in annotating TCM ontology. These experiments used the classical knowledge in the Nong's master/enterprise TCM onto-core as the basis. The results collectively indicate that both RDF and OWL are as suitable as the XML, which is the metadata for extant Nong's enterprise TCM onto-core for clinical practice. The OWL is more powerful than the RDF because it allows transitive

logical interpretations. The experiments were aided by the STV (Semantic Transitivity Visualizer), which is also a novel contribution from this thesis. The basis for a semantically based D/P system to work properly is that semantic transitivity always exists in its local onto-core.

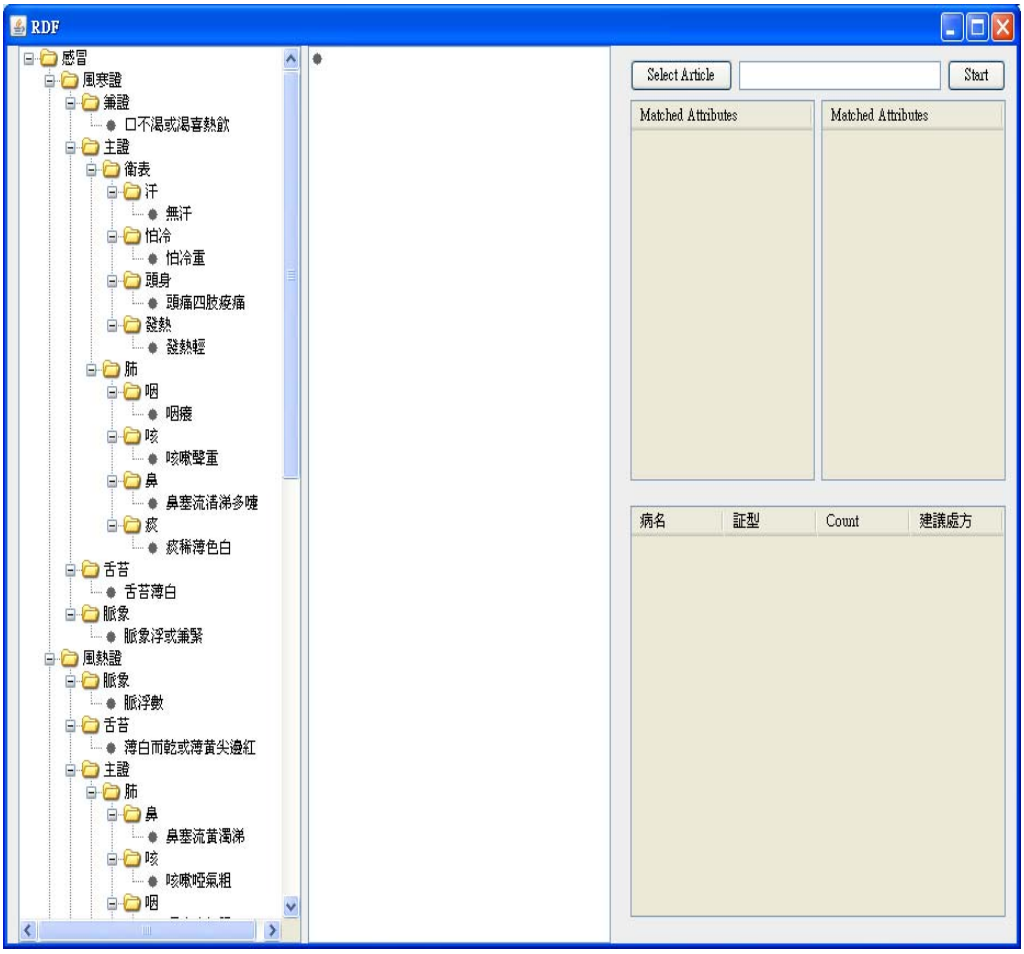


Figure 9.2.1.1 Invoked STV of a parsing operation visualization (RDF)

The screen capture in Figure 9.2.1.1 shows how the STV, which is an original contribution from this thesis, visualizes the partial DOM tree with respect to the query. This STV is needed to debug and verify the EOD-ISD approach proposed by Nong's to automate telemedicine system generation. The program first reads the RDF file for the illnesses information and presents them

in a DOM tree format. Then it extracts all the attributes in the DOM tree which will be used later in the query part. When the user selects a diagnosis case (in plain text format) containing all the symptoms of a patient for an illness case (an example is shown below), then the program will process the input content and start querying for the suggested illness name and type, and prescription.

不渴 (not thirsty), 冷重 (loath cold ambience), 頭痛四肢痠痛 (head and body ache) 鼻塞流黃濁涕 (running nose), 有汗 (sweat) 薄白 (tongue coating is thin & white)

Figure 9.2.1.2 An example of the input attributes/symptoms in a diagnosis case

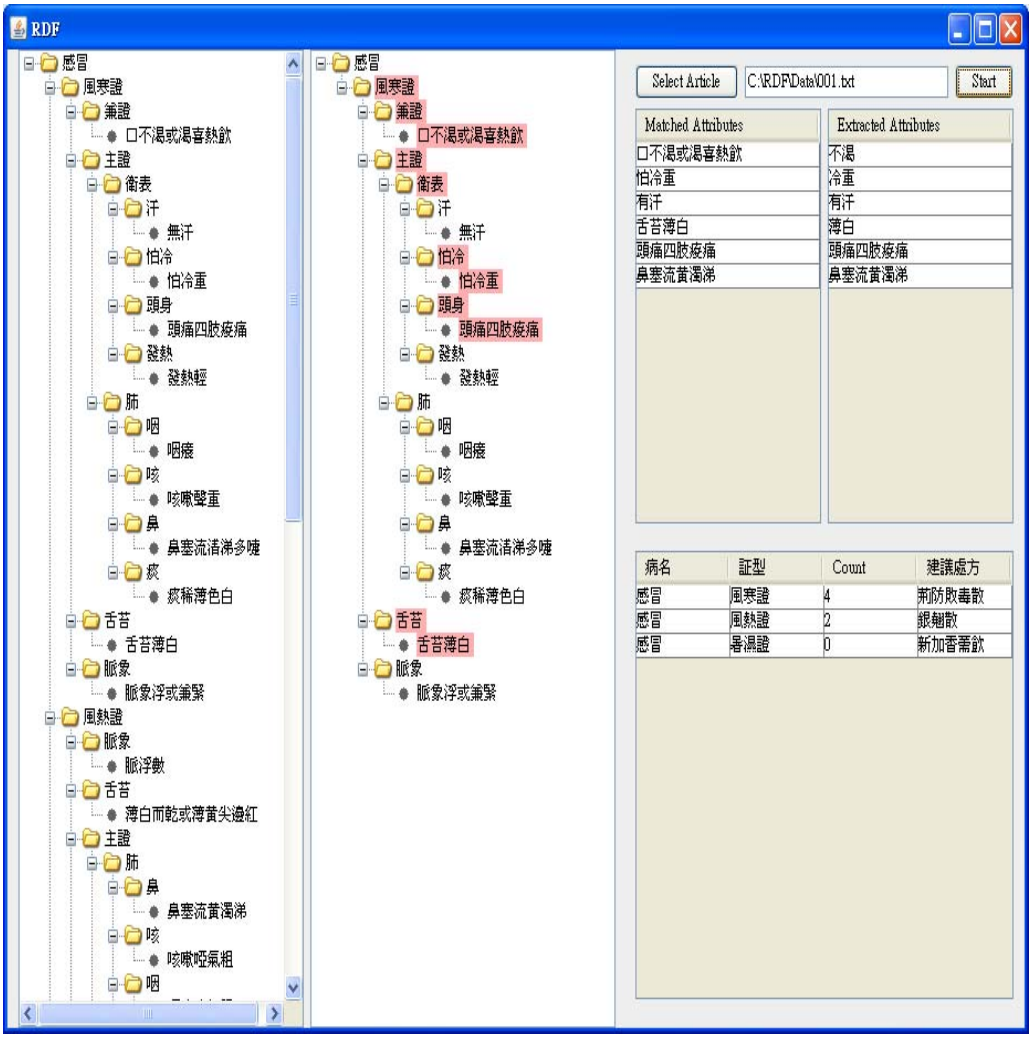


Figure 9.2.1.3 Example of a parsing operation result

After putting the diagnosis case into the program, all the attributes are extracted and displayed in the table of “Extracted Attributes”. The extracted attributes that can be mapped to the attributes in the DOM tree will then be shown in the table “Matched Attributes”. The program then counts the frequency of the matched items and provides a list of suggestions of the illness name and type according to the previous diagnosis case input. It will also provide a suggested prescription to every illness type respectively.

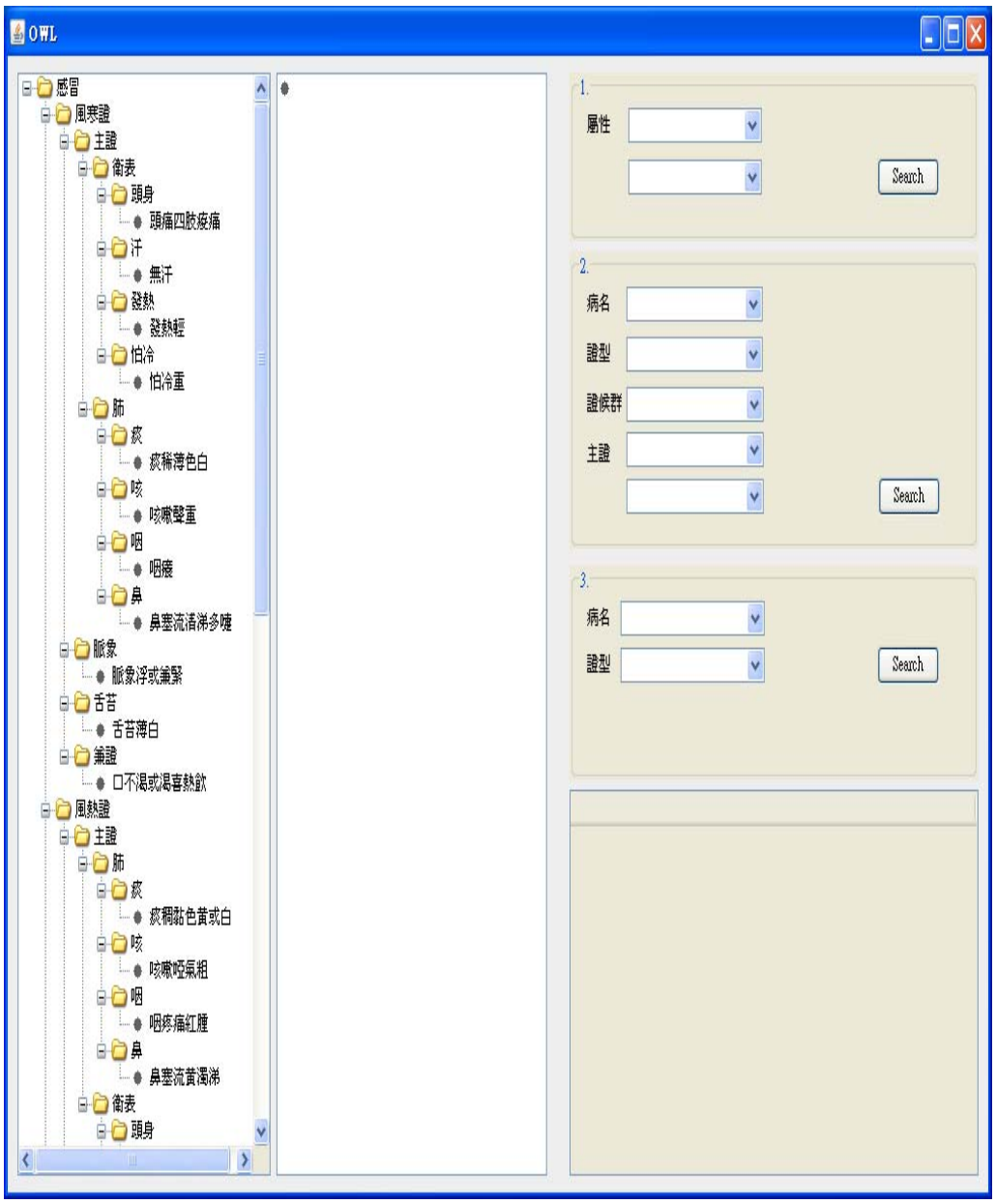


Figure 9.2.1.4 Invoked STV of a parsing operation visualization (OWL)

Figure 9.2.1.4 shows how the STV visualizes the partial DOM tree with respect to the query using OWL. The program first reads the OWL file for the illnesses information and presents them in a DOM tree format. Then it extracts all the attributes in the DOM tree, which will be used later in the query part. The program supports three different types of query. For query type 1, the user can select one of the attributes (symptoms) listed in the DOM tree, either “病

名” (the illness name) or “證型” (the illness type), then the program will return the illness name or type to which the symptom belongs.

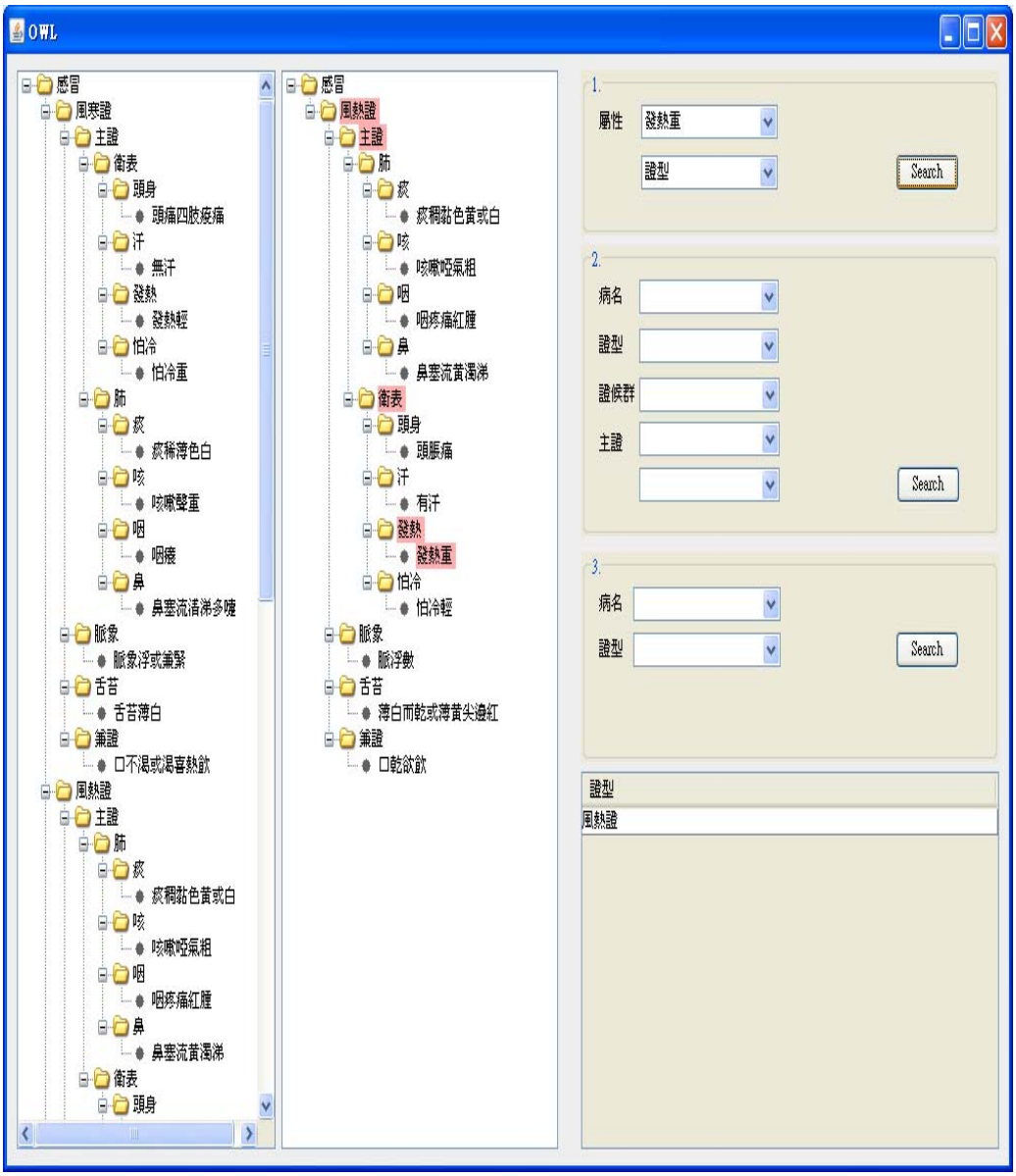


Figure 9.2.1.5 An example of an inference result (query type 1)

For query type 2, the aim is to retrieve all symptoms within the scope provided by the user. The user can select the scope from the highest level “病名” (the illness name) to the lowest level “主證分類” (part of the main

symptom). The program will then return all the symptoms within the suggested scope, which means the diagnostic result, according to these symptoms, may be within the suggested scope provided by the user.

For query type 3, the user can select “病名” (the illness name) and its “證型” (the illness type), and the program will return the suggested prescription.

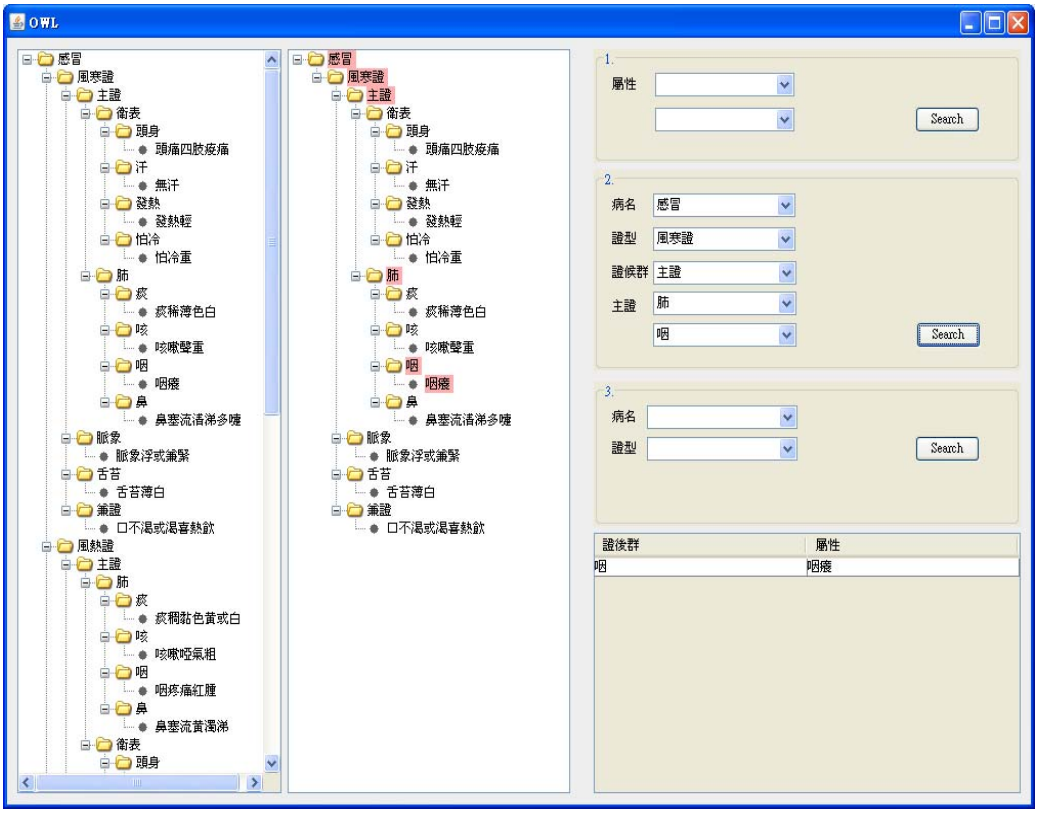


Figure 9.2.1.6 An example of an inference result (query type 2)

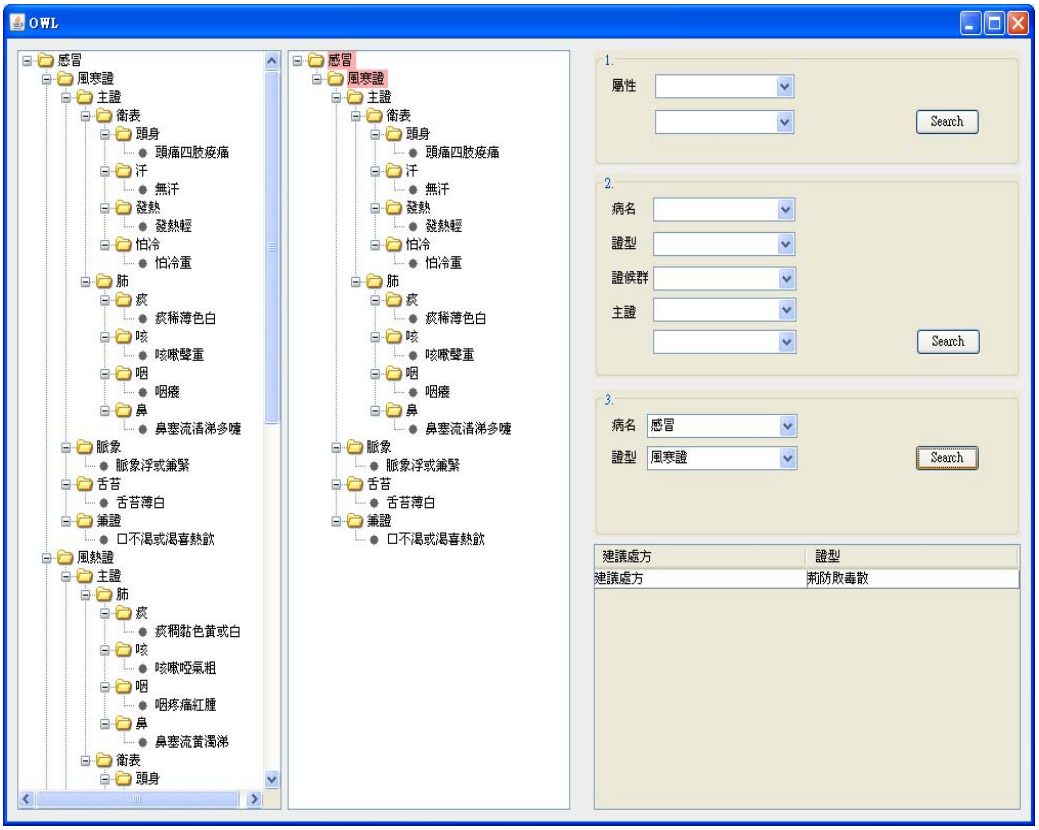


Figure 9.2.1.7 An example of an inference result (query type 3)

9.2.2 Essential Engineering Support by EOD-ISD

In the EOD-ISD approach, the user needs to provide the MI/iconic specification only, and the EOD-ISD generator automatically customizes the target system accurately in one step. Therefore, this kind of quick software engineering support is essential for producing WD^2UHI prototypes, which are D/P variants for testing the algorithms for herbal discoveries. The MI or iconic specification is a collection of icons selected from the icons library. Each icon represents a node in a distinctive explicit semantic path in the given enterprise TCM onto-core. From this MI specification a 3-layer D/P system variant is generated. The GUI of the target system, which is language independent, has

the same appearance as the MI specification. Figure 9.2.2.1 is the GUI automatically generated from the given MI specification of the same appearance.

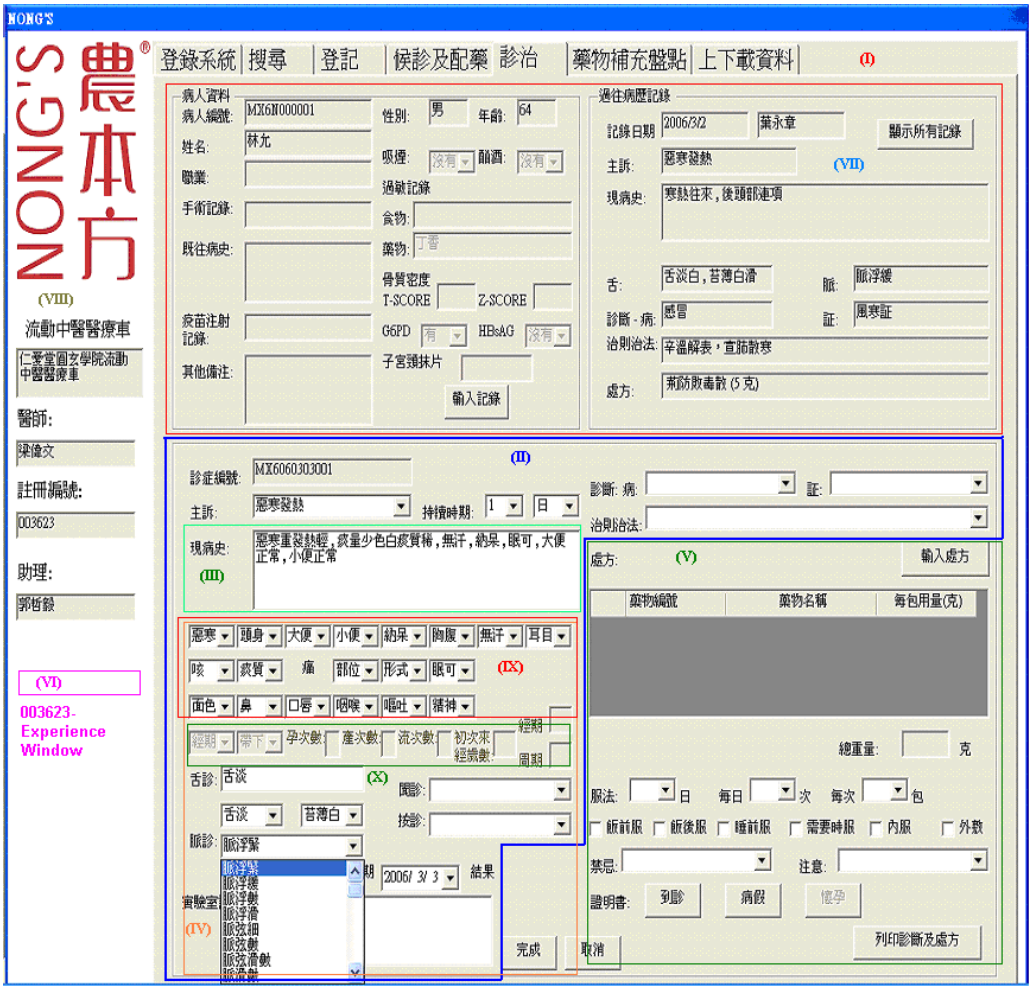


Figure 9.2.2.1 A D/P MI specification

The D/P GUI in Figure 9.2.2.1 is generated from 10 icons; one section for each icon, for example:

i) **Section (I) –** The bar of control icons.

ii) **Section (II) –** Patient registration number (or diagnostic Identifier) (i.e. MX6060303001) waiting for treatment and the important fields to be filled later: i) patient's complaint ("主訴"), ii) diagnosis ("診斷"): illness/type ("病"/"証"), and the treatment principle ("治則治法").

iii) **Section (III) –** Symptoms ("現病史") obtained by a standard TCM diagnostic procedure that has been crystallized from eons of clinical experience.

iv) **Section (IV) –** Pulse diagnosis ("脈診").

v) **Section (V) –** Prescription(s) ("處方") for the diagnosis filled in section (II); printing the final prescription and dispensing it in the MC.

vi) **Section (VI) –** Experience window (repository) entrance of the logon TCM physician with unique official medical practice registration number (e.g. 003623 as shown).

vii) **Section (IX) –** Specific questions (e.g. Do you loathe cold ambience conditions ("惡寒/怕冷")?), and general physical inspection (e.g. complexion ("面色") – pale, red or dark).

viii) Section (X) – Tongue diagnosis ("舌診") (e.g. texture and coating color).

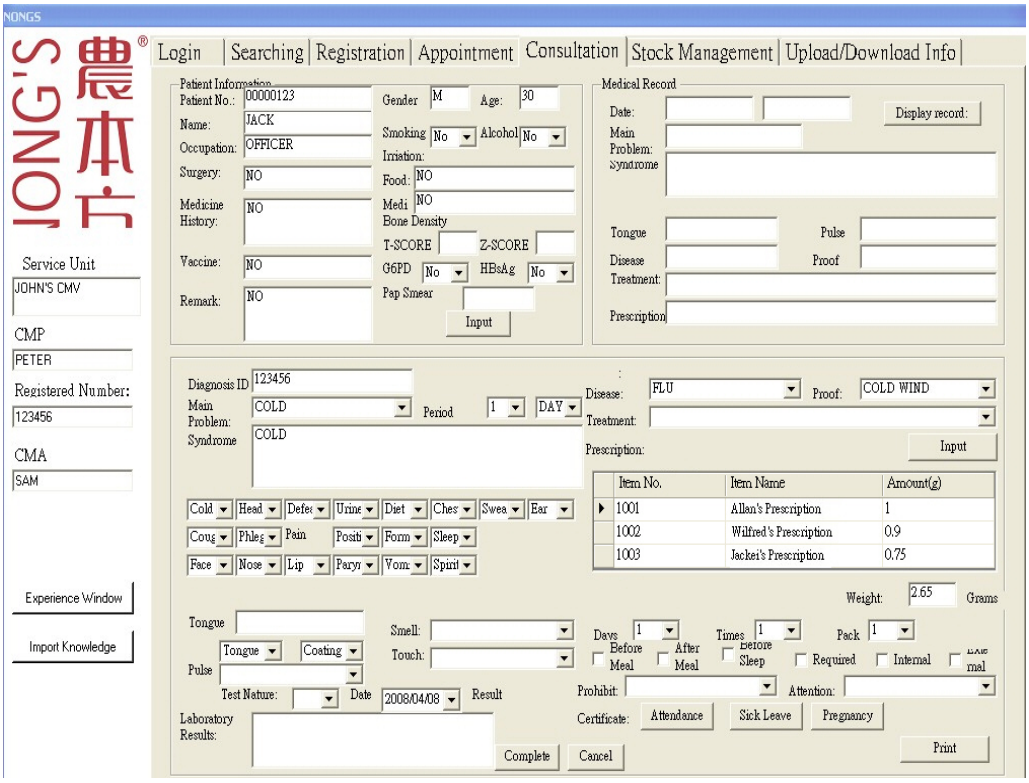


Figure 9.2.2.2 English GUI of Figure 9.2.2.1

Figure 9.2.2.2 is the English version for the GUI for the same MI specification that has generated Figure 9.2.2.1.

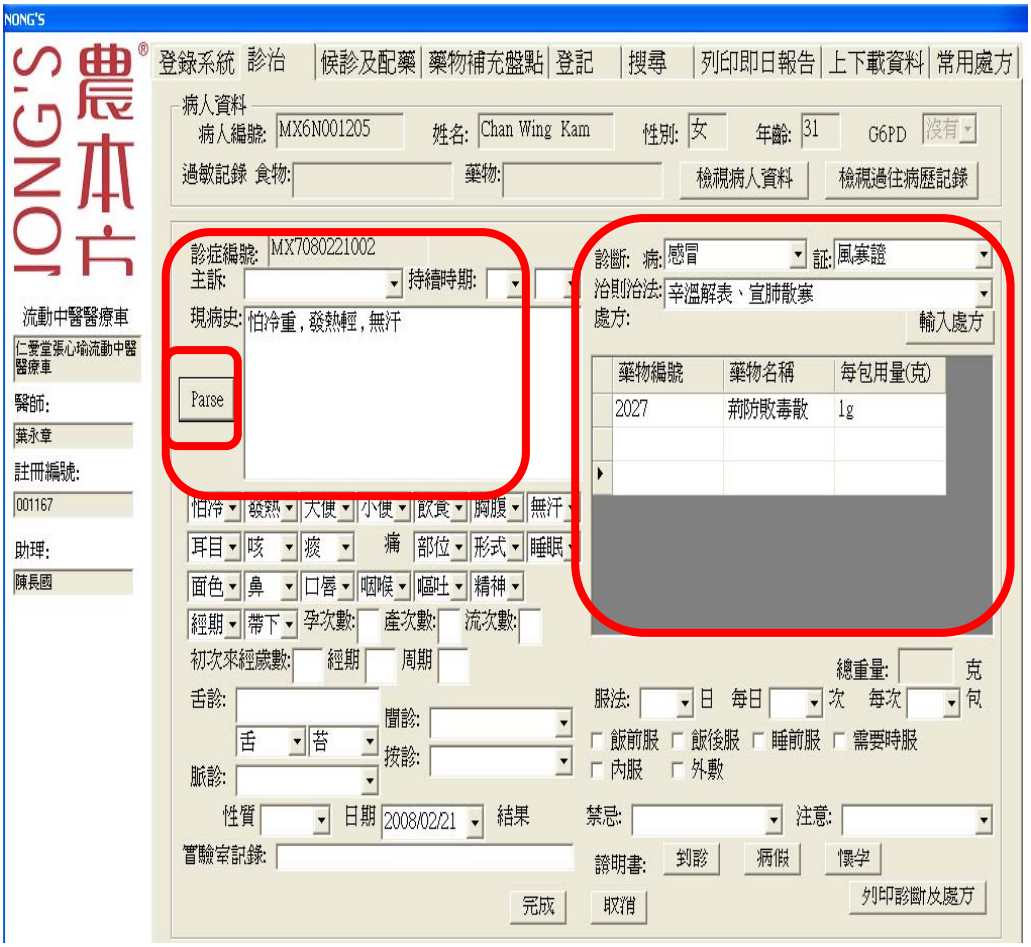


Figure 9.2.2.3 Verification of a customized D/P system indeed works correctly

Look (望)	Listen&Smell (聞)	Question (問)	Pulse-diagnosis (切)	Illness Concluded
pale face	cough, bad breadth	headache, fever, loathe cold ambience conditions (惡寒/怕冷)	taut and fast	Influenza (感冒)

Table 9.2.2.1 A traditional "望, 聞, 問, 切" **diagnosis example (manual conclusion)**

Figure 9.2.2.3 is the result of an experiment that verified that the customized D/P worked correctly. This type of verifications was accomplished by real physicians, who would be invited to use the customized system to treat

patients based on the standard TCM diagnostic procedure of four steps: Look (望), Listen&Smell (聞), Question (問), and Pulse-diagnosis (切). For example, Table 9.2.2.1 shows how this procedure of four steps reaches the diagnostic conclusion from the symptoms – Influenza (感冒).

Computer-aided diagnosis/prescription, via the GUI of the customized D/P system variant, is shown in Figure 9.2.2.3. The physician keys in the symptoms obtained from the patient as the standard procedure. In Figure 9.2.2.3 the "Symptoms (現病史)" window echoes the keyed-in symptoms, which were obtained with the four steps: Look (望), Listen&Smell (聞), Question (問), and Pulse-diagnosis (切). Table 9.2.2.1 shows an example of how these four steps would be applied by the physician to reach a diagnostic conclusion in the traditional and manual way; in this case Influenza (感冒) is concluded. The standard "key-in" D/P operation potentially allows the D/P results be absorbed as immediate feedback to enrich TCM onto-core (provided it is open). In contrast, feedback is not possible for the traditional handwritten D/P because it involves "non-standard" terms, which are either not enshrined in TCM classics or included in the enterprise TCM vocabulary {V}. Yet, these non-standard terms can be standardized by automatic semantic aliasing, which is explained in detail in Chapter 6, to a varying degree as indicated by the computed relevance indices.

The keyed-in symptoms are accepted immediately if they are standard, as those enshrined in the enterprise vocabulary {V}. In the D/P GUI, specific

terms are grouped for designated windows (e.g. the 怕冷 (loath cold ambience) symptom can be selected from one of the windows), (as shown in the Symptoms ("現病史") window in Figure 9.2.2.1) the specific window in section (IX) of the GUI in Figure 9.2.2.1. To each question, the physician keys-in the answer from the patient. For example, if the answer to the question of profuse sweating is "NO", the physician selects 無汗 (no perspiration) and keys it in. The answer will be automatically checked against the enterprise vocabulary {V} before it is echoed (e.g. "Symptoms (現病史)" window). All the echoed answers would become the actual parameters for the implicit query to be constructed by the D/P interface. The implicit query excites the parser, which returns the logical diagnostic conclusion by inference. In Figure 9.2.2.3, the parser concluded for the implicit query $Q\{$怕冷重, 發熱輕, 無汗$\}$ the following: a) diagnosis (診斷) – illness (病) is Flu (感冒) and type (証) is "wind cold" (風寒); b) treatment principle (治則) – heating and sweating (辛温解表); and c) prescription (處方) – "荊防敗毒散" ("Jing Je Infusion"), an old recipe established centuries ago.

9.2.3 Semantic Transitivity Visualization

Semantic visualization is achieved through the Semantic TCM Visualizer (STV) either manually (via the Engineering Interface (EI)) (e.g. Figure 6.2) or by pressing the "Parse" button on the D/P interface (e.g. Figure 9.2.2.3 – the earlier version called the primodial form [Ng08]). The STV interface has two main parts as shown in Figure 9.2.3.1: i) the left window

shows the partial DOM tree corresponding to the input parameters (highlighted); and ii) the right side of small “semantic windows (SW)” through which the user could manually select and key in the parameters. The aim of the SW is to support system debugging. The rationale is that those parameters keyed in via the semantic windows become the actual parameters for the implicit query built by the STV interface. If the parser fails to draw a logical conclusion for the implicit query, there could be errors in the DOM tree, the parser program, or both.

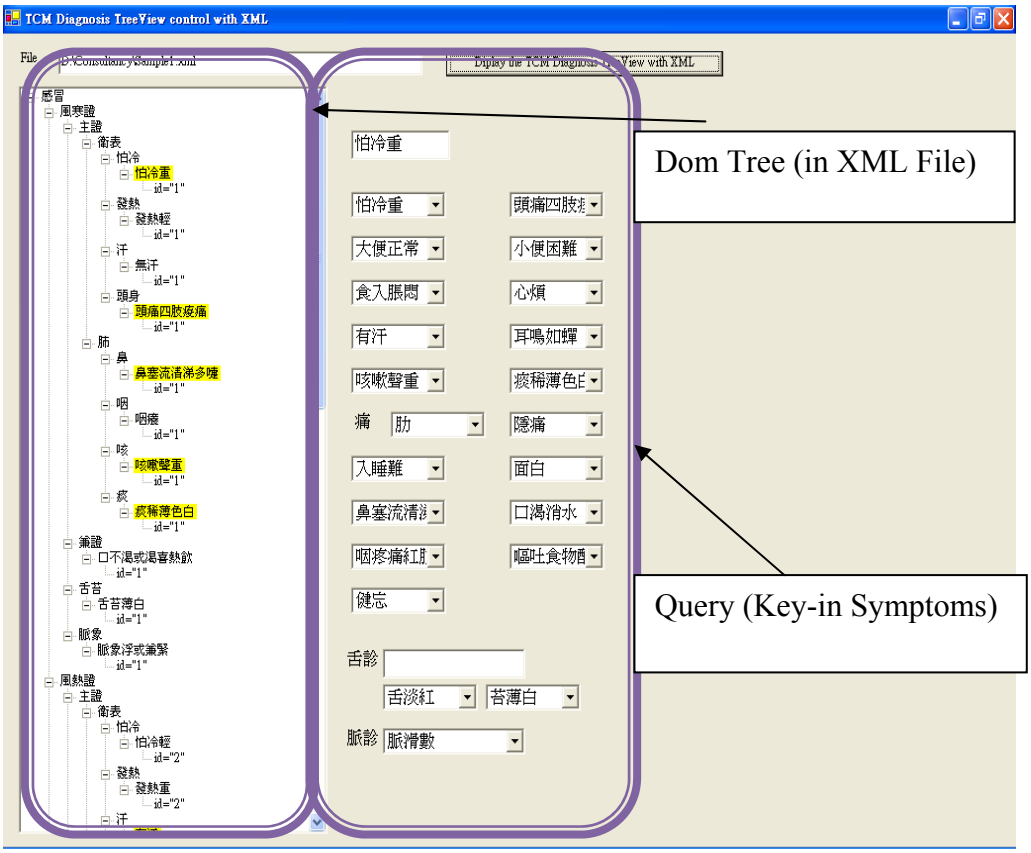


Figure 9.2.3.1 Invoked STV to visualize the parsing operation [Ng08]

The novel STV proposed in this thesis is an improved version of the previous primodial form [Ng08], which I helped performed many verification tests. It serves the following purposes:

a) **Verification and debugging:** Figure 9.2.3.1 is the screen capture of the STV in which the windows on right side allow the user to selectively key in parameters. With these parameters, the STV mechanism forms the implicit query from which the parser draws the logical conclusion from the semantic net by inference. The keyed-in parameters for the implicit query are highlighted in the DOM tree, and the logical conclusion in this case is Influenza (感冒). For debugging purposes the user invokes the STV manually through the Engineering Interface (Figure 6.2).

b) **Clinical visualization:** This is invoked via the Parser button in the D/P GUI (Figure 9.2.2.3 – the primodial form), which allows the physician to visualize how the parser draws its conclusions for an implicit query that was constructed with keyed-in parameters (i.e. symptoms) such as those in window (III) ("Symptoms (現病史)") of Figure 9.2.2.3. The illness and prescription conclusions should be displayed respectively in the windows (II) and (V). In Figure 9.2.3.1, the Parser button, the symptoms and the corresponding logical D/P conclusion are highlighted. Clinical visualization is for clinical decision support in the field rather than verifying/debugging the DOM tree.

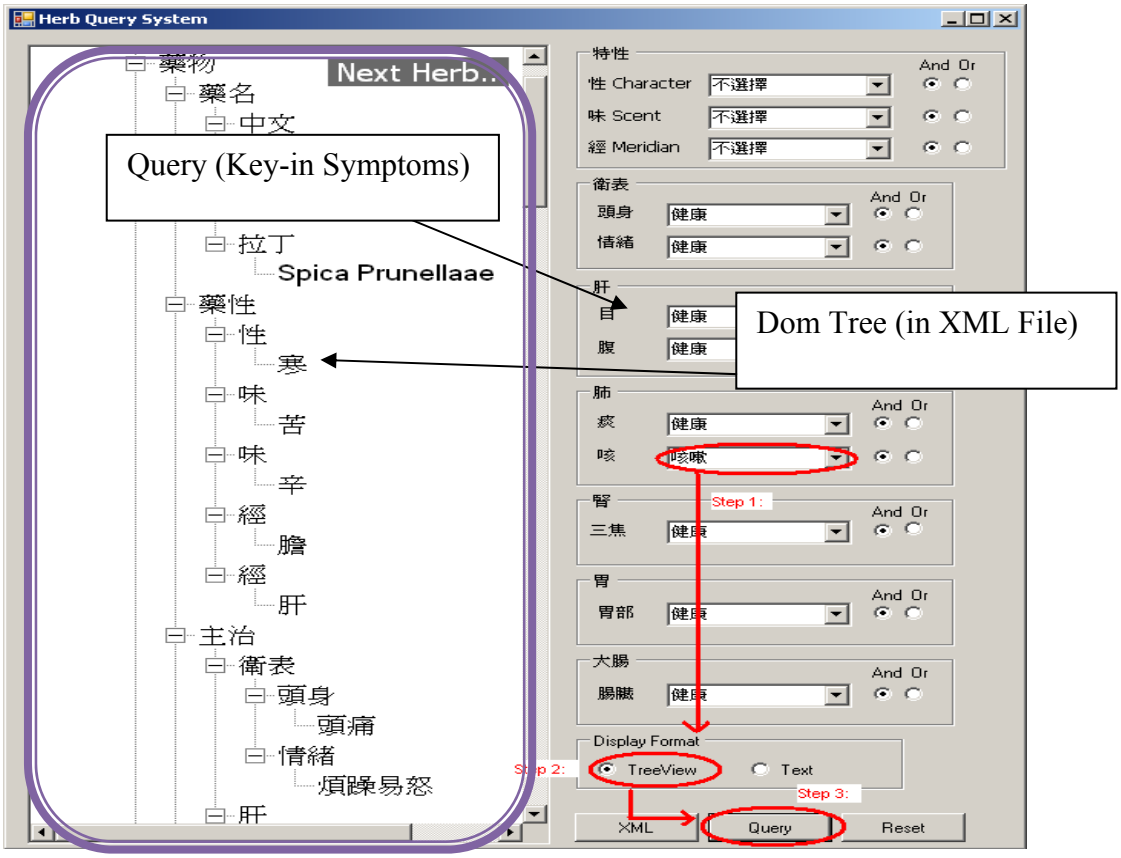


Figure 9.2.3.2 DOM tree view for the symptom "咳嗽 – coughing with sputum" [Ng08]

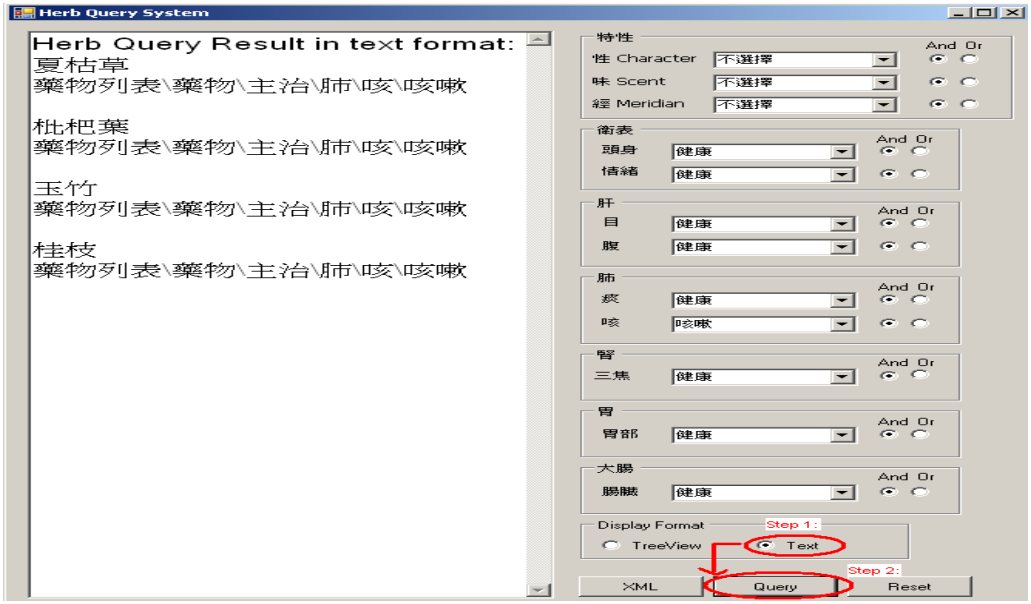


Figure 9.2.3.3 Textual view of the DOM tree view in Figure 9.2.3.1 (cross-referencing) [Ng08]

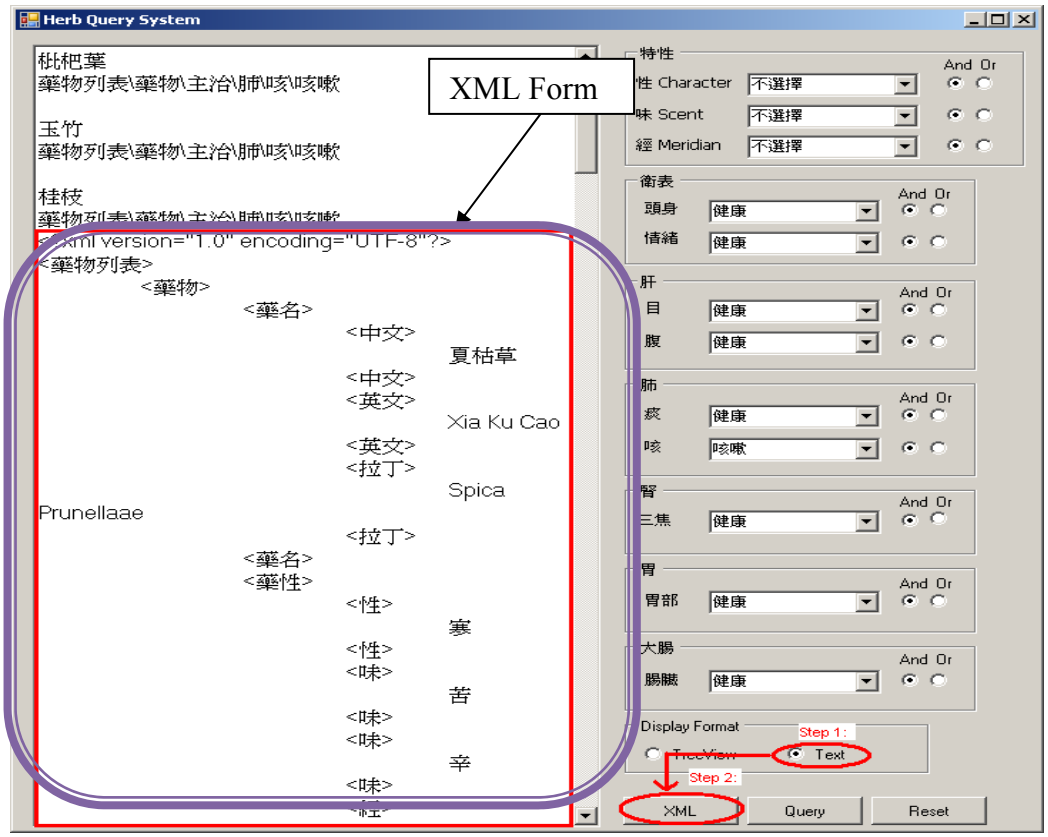


Figure 9.2.3.4 Textual view (Figure 9.2.3.1) versus its XML form (in TCM onto-core) [Ng08]

Figures 9.2.3.2, 9.2.3.3 and 9.2.3.4 are screen captures from one of the many experiments that successfully verified the cross-layer cross-referencing capability of primodial STV [Ng08]. Consistent cross-referencing among different layers in a D/P system confirms the semantic transitivity among them. Figure 9.2.3.2 shows the "DOM tree view" – the middle semantic net layer. Figure 9.2.3.3 shows the "textual view" of the contents in Figure 9.2.3.2. This simple view, which lets the layman understand the nature of herbs, is basically a textual abstract. Figure 9.2.3.4 associates/connects this textual abstract with/to the XML segment excerpt from the bottom ontology layer. The excerpt is shown in the left window of the STV GUI; this window is scrollable and the user is reminded of this possible by the display: "Next Herb…" as shown in Figure 9.2.3.2. The user can switch to cross-examine the different views, which associate with different architectural layers, back and forth. This can be achieved by activating the right GUI buttons, as marked (circulated) in Figures 9.2.3.2, 9.2.3.3 and 9.2.3.4, which together show how correct cross-layer cross-referencing operations are actually supported by congruent semantic transitivity. The philosophy of the primodial STV is the basis for the novel (very much improved) version proposed in this thesis.

9.2.4 The OCOE&CID Approach

The The OCOE&CID (***O**n-line **C**ontinuous **O**ntology **E**volution and **C**linical **I**ntelligence **D**iscovery*) approach, which is within the EOD-ISD conceptual framework, is aimed at achieving the following objectives: i) Real-time continuous and automatic TCM onto-core evolution; ii) Real-time

automatic semantic aliasing; iii) Special elements to support real-time ontology evolution; iv) Text mining and RI computation; and v) intelligence discovery with the "SAME - 同病異治, 異病同治" principle. The experimental results in this section demonstrate how these objectives are achieved. The word "SAME" is *synonymous* of the Chinese character "同".

9.2.4.1 Enterprise Standard Extension by Text Mining and Semantic Aliasing

The initial construction of the enterprise skeletal TCM onto-core, on which all PP/N telemedicine system variants are based, was semi-manual. The automatic part of this process is extraction of relevant TCM terms, concepts, and their associations from TCM classics by text mining [Holzman03, Bloehdorn05, Yu06]. The extracted items were first pruned manually by domain experts as consensus certification. The accepted items were encoded into the XML metadata format. This XML representation provides the skeleton of the intrinsic TCM onto-core, which was repeatedly tested, verified and validated with the aid of the STV. The extant STV then conducted the cross-layer cross-referencing operation, to verify and confirm the existence of the necessary semantic transitivity in the telemedicine system.

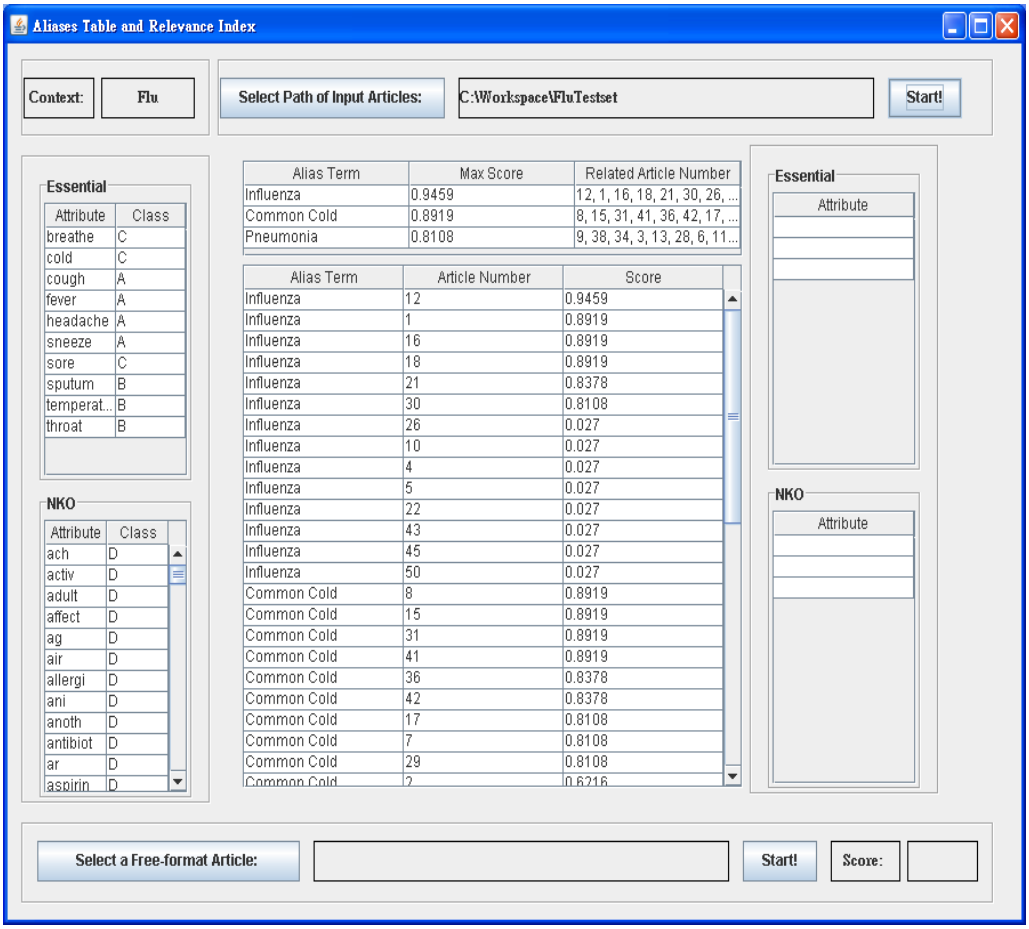


Figure 9.2.4.1.1 Example of enterprise standard extension by data mining and aliasing

Intrinsically the skeletal TCM onto-core does not evolve automatically. In the OCOE&CID framework text mining is applied to make on-line evolution possible. In the OCOE&CID context this is the *enterprise standard extension* (ESE) operation since the enterprise TCM onto-core is the standard for D/P operation. The ESE process involves the following: i) the text miner works on the given set of selected TCM classics; ii) the miner finds attributes (i.e. symptoms) and the illness (or context) from every article in the set; and iii) the miner applies *automatic semantic aliasing* (ASA) to mark the relevance of the newly found items for every referential context (RC) in MAT (the initial RC list

in MAT is the same as the skeletal TCM onto-core – the foundation). In this way the ASA process continuously extends the overall TCM onto-core of the running telemedicine system. Figure 9.2.4.1.1 shows the result obtained from a selected ESE experiment; the C:\Workspace\FluTestset folder contains 50 given TCM classical articles. The screen capture shows: i) the file name of the 50 articles; ii) the RC was Flu when the screen was captured; iii) classes of the essential attributes (i.e. A, B, C, and D; iv) RI score for every article against the target RC at the time (e.g. Flu); and v) articles with the maximum RI (relevance index) scores (i.e. 0.9459, 0.8919, and 0.8108). Via ASA, the essential information about the 50 given articles was absorbed into the MAT contents. Conceptually the closed skeletal TCM onto-core is extended logically by the ESE operation; the enterprise operation standard is continuously extended.

The free-format (FF) semantic aliasing operation is automatic and real-time, while in the off-line ESE operation the set of articles has already been provided. An FF article is arbitrarily found by the text miner over the open web. After automatic semantic aliasing and pruning, some of the findings would be used to enrich the TCM onto-core.

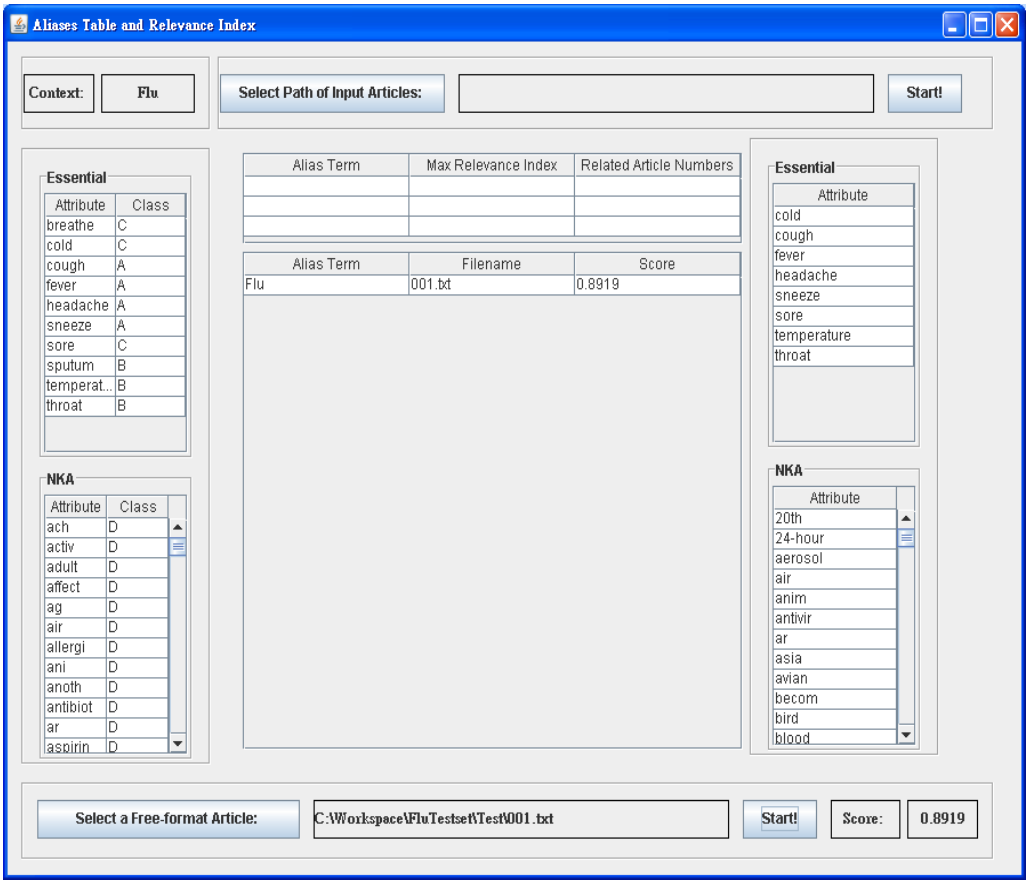


Figure 9.2.4.1.2 On-line continuous TCM onto-core evolutionary process

Figure 9.2.4.1.2 is the result of a real-time FF automatic semantic aliasing experiment. In this case, the article 0.001.text (in the C:\Workspace\FluTestset\Test folder) was the FF article input in the simulation. From that article the text miner first extracted the attributes (both essential and NKA) and the context (i.e. Flu). Then, the ASA process computed the RI for the context with respect to the RC (Flu when the screen was captured). The RI computed for this FF article is 0.8919, based on the concept depicted in Figure 6.3. Interestingly, “Flu” versus “Flu” in Figure 9.2.4.1.2 indicates alias rather than synonym. This outcome came from that fact that the FF Flu article contains fewer essential attributes than the number enshrined in the TCM onto-

core for the Flu RC. This reminds those who would read the FF article that its contents may have only 89.19% relevance to the subject of Flu.

9.2.4.2 Prescription Discovery by the "SAME" Principle

The SAME or "同病異治, 異病同治" principle stated that: "*If the symptoms are the same or similar, different conditions could be treated in the same way medically, independent of whether they come from the same illness or different ones.*" [WHO07] The key to realizing this principle is semantic aliasing and RI computation as shown in Figure 6.3. From the RI scores of the aliases the total set of prescriptions for treating the RC can be concluded. (e.g. $P_{A,a}^{total} = PAa \cup PAb \cup PAc \cup PXx$ from Figure 6.3). Figure 9.2.4.2.1 shows the result from an experiment, which was set up to verify the SAME principle; the TCM onto-core was open and evolvable. The walkthrough of this experimental result is as follows:

a) *Patient's complaint (主訴):* The patient complained of "loathing cold ambience and had fever – 惡寒發熱" (window (II) in Figure 9.2.4.2.1)

b) *Symptoms (現病史):* The eight symptoms provided by the patient were keyed in via the D/P interface (e.g. "no perspiration – 無汗" symptom)
(window (III) of Figure 9.2.4.2.1).

c) *Prescriptions (處方):* Four usable prescriptions were suggested by the D/P system, as a result of ASA; they had different RI scores: 1.0 (directly from the RC), 0.7 (from 1st alias of RC), 0.5 (from 2nd alias of RC), and 0.2 (from 3rd alias of RC). The RI score indicates the relative efficacy of the alias's prescription for treating the RC, as shown in the (IV) section of Figure 9.2.4.2.1. In the original Nong's MC telemedicine medicine system, which does not has the ASA capability, the set of prescriptions for treating a RC is restricted to the one established initially by the consensus certification process (i.e. only the PA_a set for treating the **Illness (A, a)** RC as shown in Figure 6.3).

The union of the aliases' prescription sets and the RC's produces a much bigger usable set, for example, $P_{A,a}^{total} = PAa \cup PAb \cup PAc \cup PXx$. In the OCOE&CID context, this is clinical intelligence discovery because the $PAb \cup PAc \cup PXx$ subset suggested by the D/P system to treat the RC (in addition to PAa) might has never been enshrined in any TCM classics.

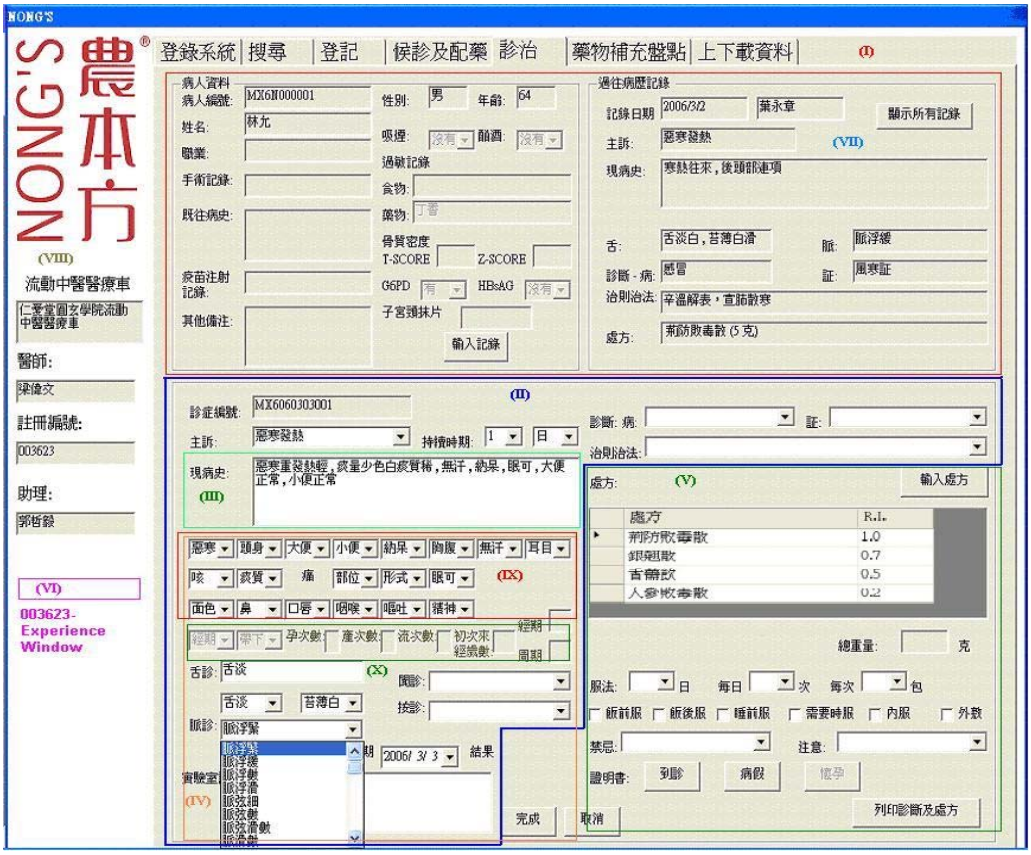


Figure 9.2.4.2.1 ASA has established a larger set of prescriptions for treating the RC

9.2.5 Knowledge Classification – NN Approach

Before fitting the data into the neural network for training, all data has to be pre-processed first. Attributes and classes (the result) are extracted based on the input XML file shown below.

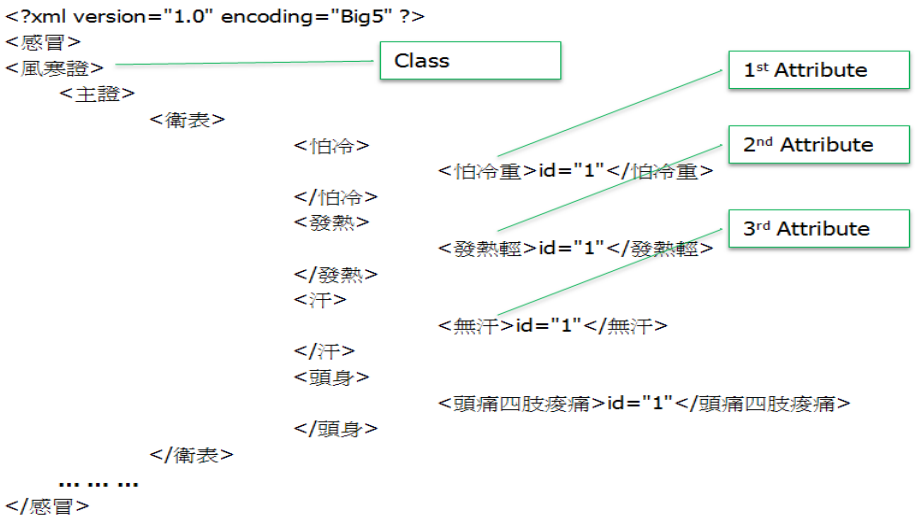


Figure 9.2.5.1 The input XML file

After processing the XML file, attributes are selected that act as the signal of the input node of the neural network. For the training data, real clinical data from Nong's (with permission) was used. An example of the diagnosis record with the prescription number "PR0403000129" is as follows:

i) Statement: "惡寒發熱 頭痛 納差 口淡不渴 頭暈而重 多夢易醒 沉細弱 淡紅苔膩" (Aversion to cold, fever, headache, loss of appetite, bland taste in the mouth, not thirsty, dizziness, profuse dreaming, easily woken up, weak & fine sunken (deep inside) pulse, pale red tongue with slimy fur).

ii) Diagnosis: "感冒 - 風寒束表" (Cold – Wind Cold Syndrome)

9.2.5.1 Experimental Results for Demonstration

The GUI of the neural network model has several sections, as follows:

1. Input file path for the XML file, training dataset and testing dataset.
2. Model selection – Using one network for all illness types, or separate network for each individual illness type.
3. Weighting enabling for different sections of the diagnosis (e.g. 主證 (Main Symptom), 兼證 (Other Symptom), 舌苔 (Tongue) and 脈象 (Pulse)), for example:
   - The weightings of four parameters are 0.5, 0.2, 0.15 and 0.15 respectively, provided that there are four attributes under the category"主證 (Main Symptom)", and one attribute under the remaining three categories. If only two attributes under the category"主證 (Main Symptom)" exist, the calculation of the input value is shown below:
     - $[(1 \bullet \frac{1}{4}) + (1 \bullet \frac{1}{4}) + (0 \bullet \frac{1}{4}) + (0 \bullet \frac{1}{4})] \bullet 0.5 = 0.25$

     The calculation of the input value of "兼證 (Other Symptom)" is (assuming the attribute exists):

     - $1 \bullet 0.2 = 0.2$

     The calculations are the same for the other parameters.
4. Parameters for the neural network:
   - Learning rate: A common parameter in many of the learning algorithms that affects the speed at which the neural network arrives

at the minimum solution. In backpropagation, the learning rate is analogous to the step-size parameter from the gradient-descent algorithm. If the step-size is too high, the system will either oscillate about the true solution, or it will diverge completely. If the step-size is too low, the system will take a long time to converge on the final solution.

- Training episodes: The number of episodes for the neural network to train.
- Momentum: Is used to prevent the system from converging to a local minimum or saddle point. A high momentum parameter can also help to increase the speed of convergence of the system. However, setting the momentum parameter too high can create a risk of overshooting the minimum, which can cause the system to become unstable. A momentum coefficient that is too low cannot reliably avoid local minima, and also can slow the training of the system.
- Decay: Allows connection weights in a network to differentially decay towards zero. This will divide the starting learning rate by the number of episodes to determine what the current learning rate should be. The purpose of using decay in neural network is to help to stop the network from diverging from the target output, and also to improve general performance.

5. XML DOM Tree visualizing section.
6. Result: Shows the number of illness types in XML file, number of attributes of each illness type and the RMSE during the training process

of the neural network).

7. Result – Using the testing dataset to validate the performance of the neural network (using accuracy, precision and recall as the performance measurement).
8. Result export to Excel file (e.g. result and RMSE).

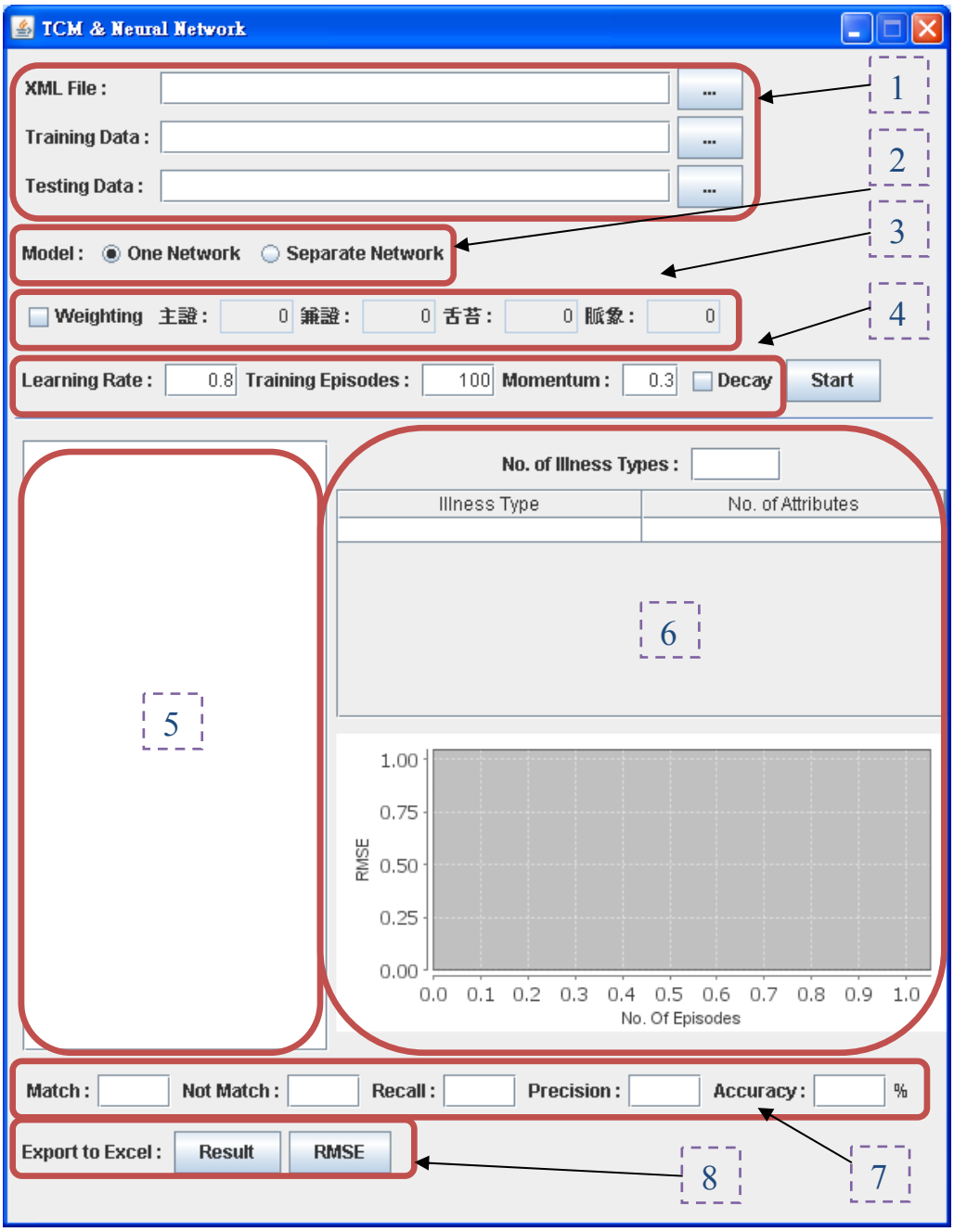


Figure 9.2.5.1.1 The GUI of the TCM neural network model

Weighting of Different Attribute Groups: The weightings of the four attribute groups 主證 (Main Symptom), 兼證 (Other Symptom), 舌苔 (Tongue) and 脈象 (Pulse) represent how important each attribute group is compared to the other attribute groups. The rule is that the sum of the weights of the four attribute groups must be equal to one. The input to the neural network will be processed before feeding into the network. If the attribute exists, the input value is 1; if not, it is 0. If the attribute exists it will be weighted; that is 1 multiplied by the weight value of the relevant attribute group, and then multiplied by 10 (applicable to all) to avoid the decimal value that may decrease the importance of the input.

Prescription No.	Statement	History	Pulse	Tongue	Diagnosis
WC02	流涕三天 (Snivel for 3 days)	流清涕 頭痛 咽癢 (Clear snivel, headache, throat itching)	脈象浮或兼緊 (Floating tight pulse)	舌苔薄白 (Thin white fur)	風寒 (Wind cold syndrome)

Table 9.2.5.1.1 A record in the training dataset

Result of Model 1 (One Network):

Case 1a – 45 training data, 15 testing data, and weighting disabled:

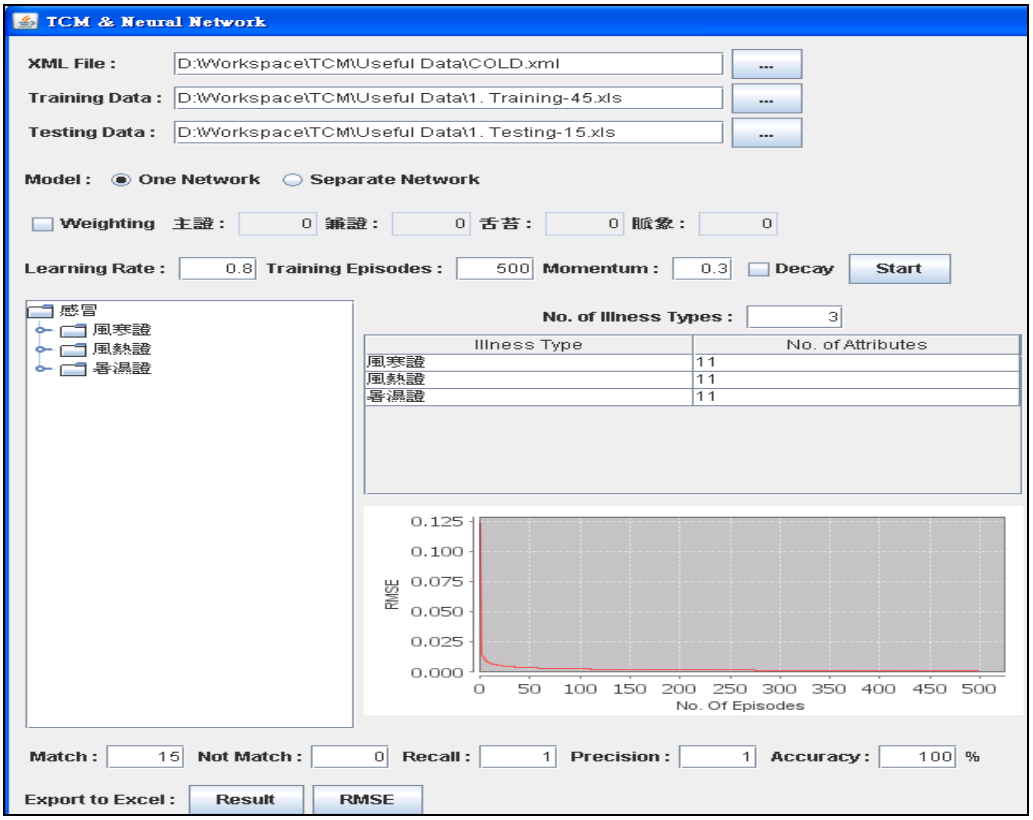


Figure 9.2.5.1.2 Result of model 1 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting and decay disabled)

The result shows 100% accuracy of the prediction result and after 500 training episodes, the RMSE of the neural network converges and the resulting error is less than 0.001.

Case 1b – 45 training data, 15 testing data, and weighting enabled:

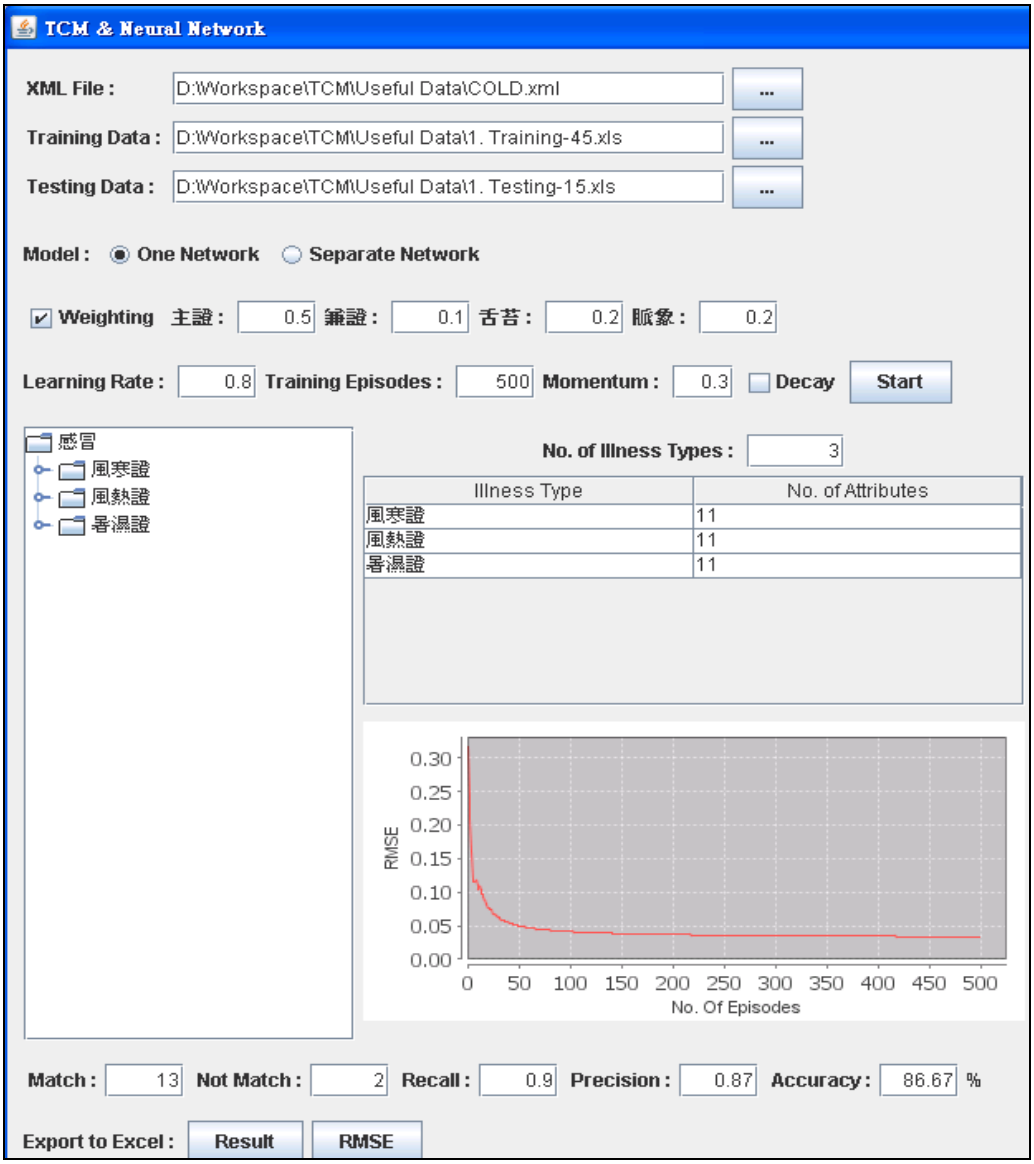


Figure 9.2.5.1.3 Result of model 1 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting enabled and decay disabled)

The result shows 86.67% accuracy of the prediction result and after 500 training episodes, the RMSE of the neural network converges and the resulting error is less than 0.05.

Case 2a – 60 training data, 60 testing data, and weighting disabled:

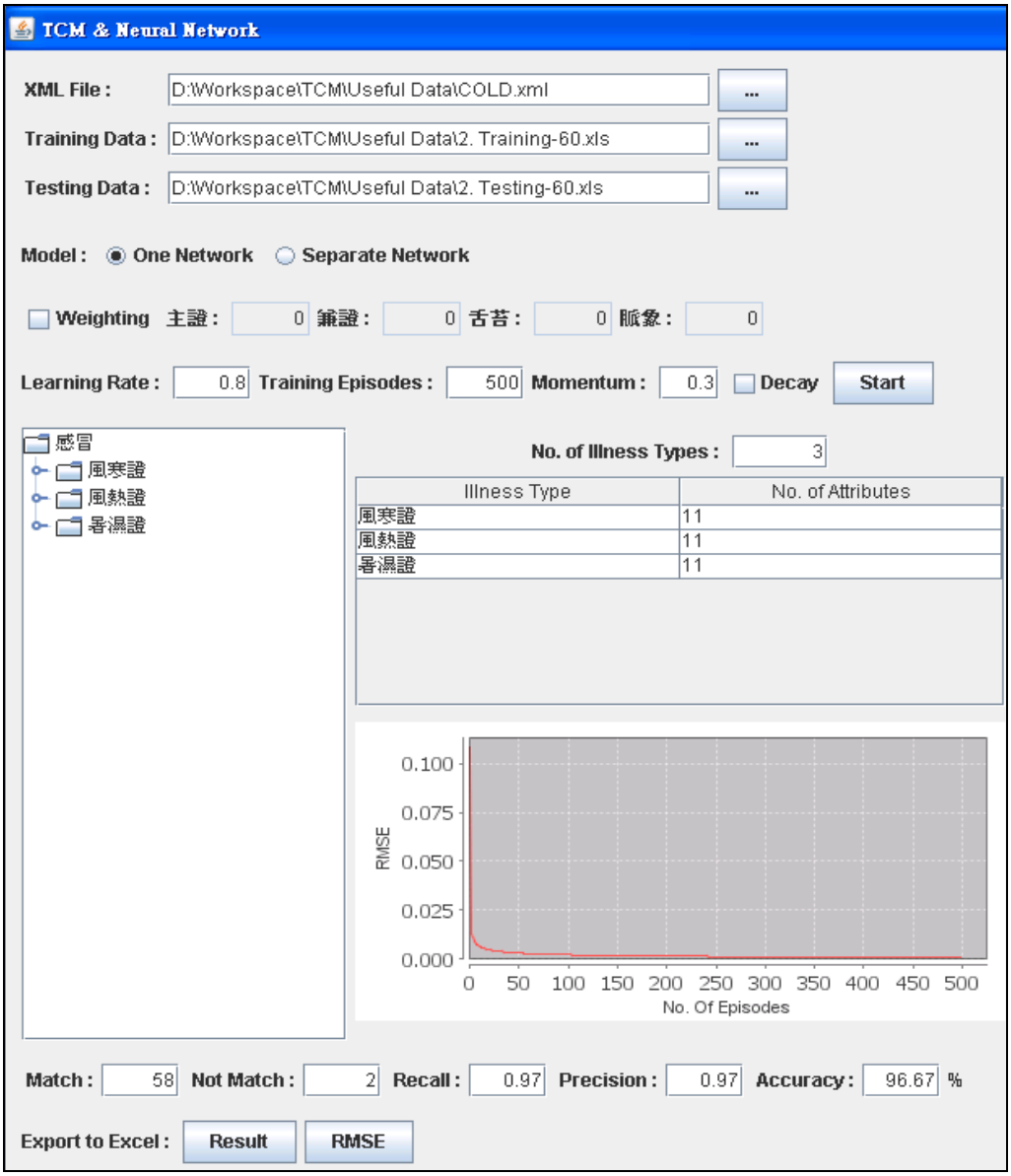


Figure 9.2.5.1.4 Result of model 1 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting and decay disabled)

The result shows 96.67% accuracy of the prediction result and after 500 training episodes, the RMSE of the neural network converges and the resulting error is less than 0.001.

Case 2b – 60 training data, 60 testing data, and weighting enabled:

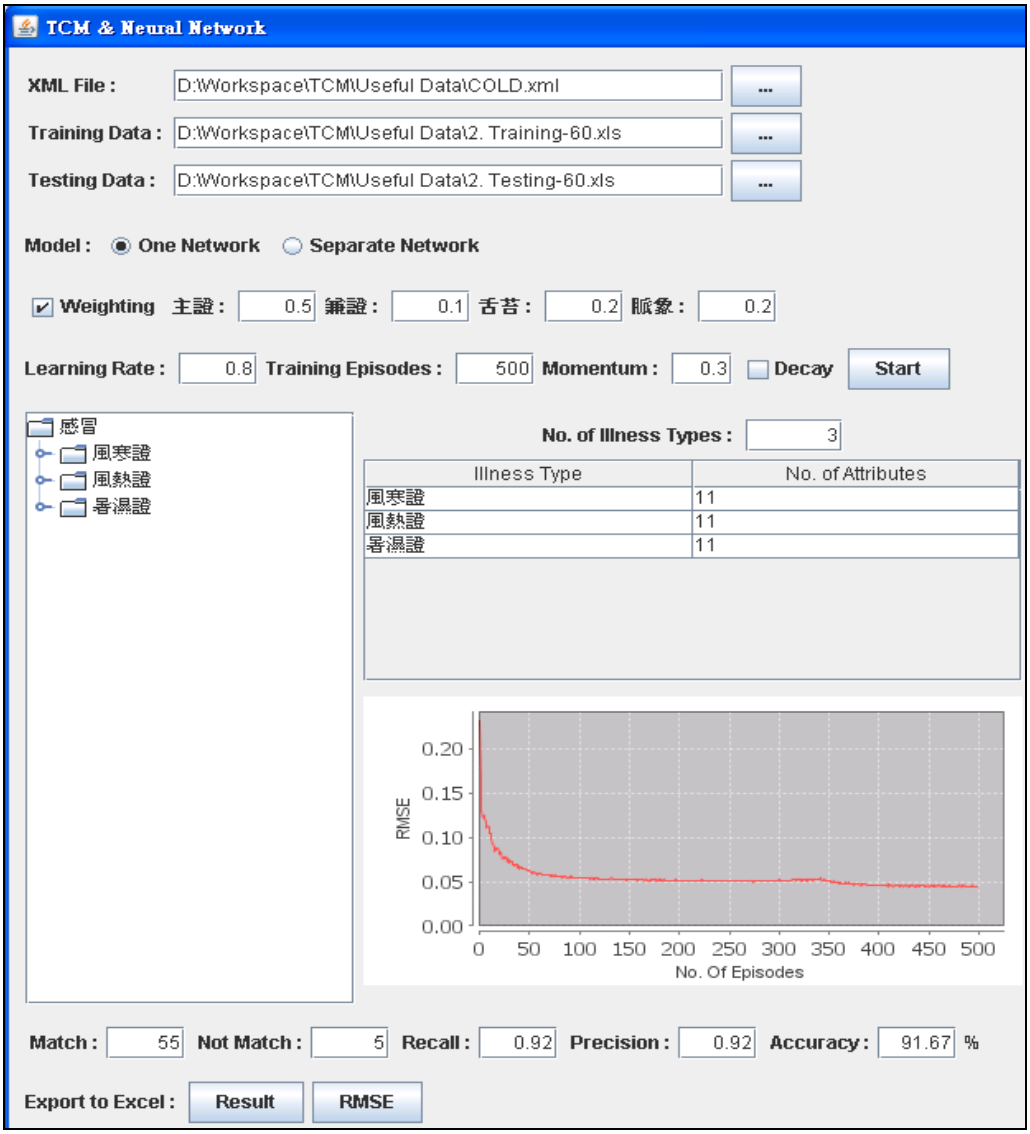


Figure 9.2.5.1.5 Result of model 1 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting enabled and decay disabled)

The result shows 91.67% accuracy of the prediction result and after 500 training episodes, the RMSE of the neural network converges and the resulting error is less than 0.05.

Result of Model 2 (Separate Network):

Case 1a – 45 training data, 15 testing data, and weighting disabled:

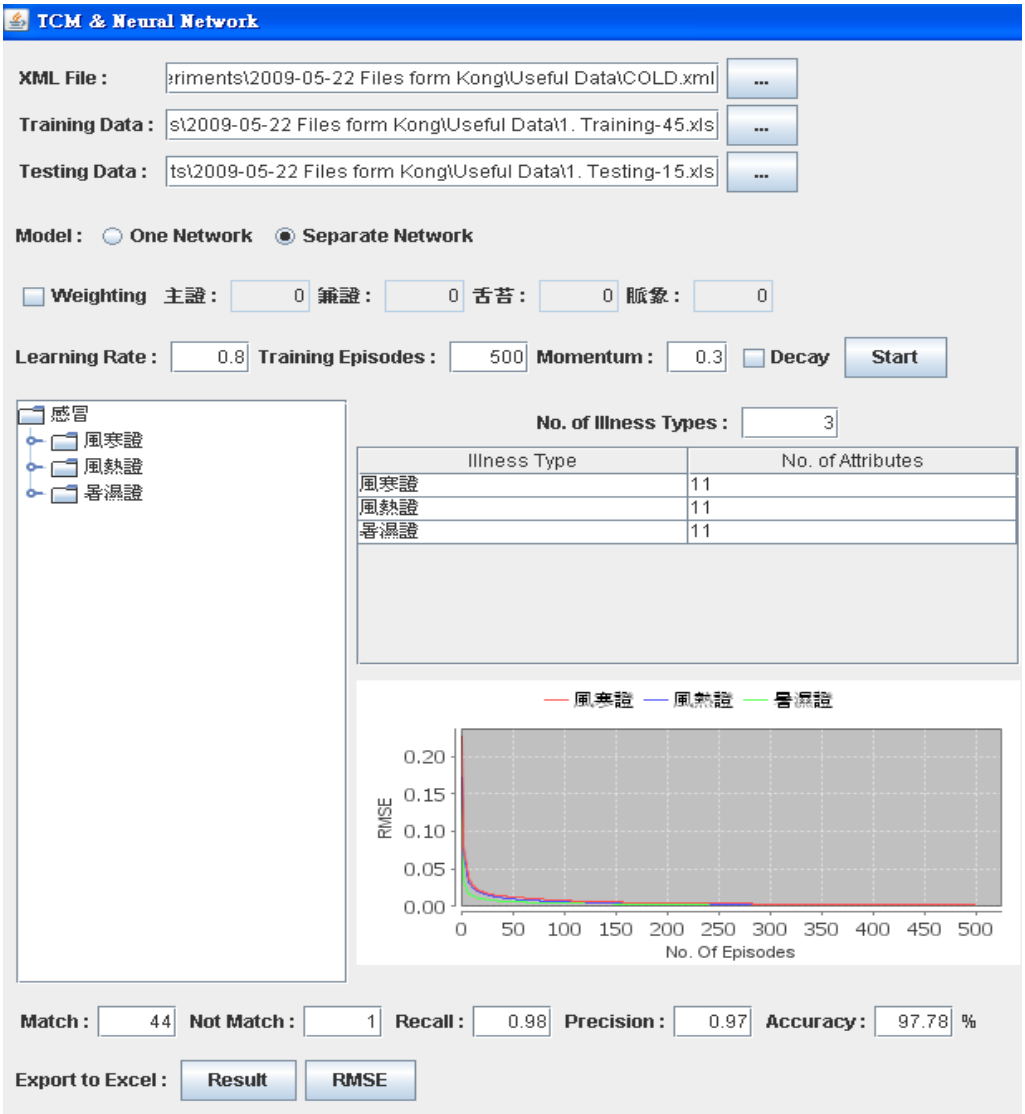


Figure 9.2.5.1.6a Result of model 2 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting and decay disabled)

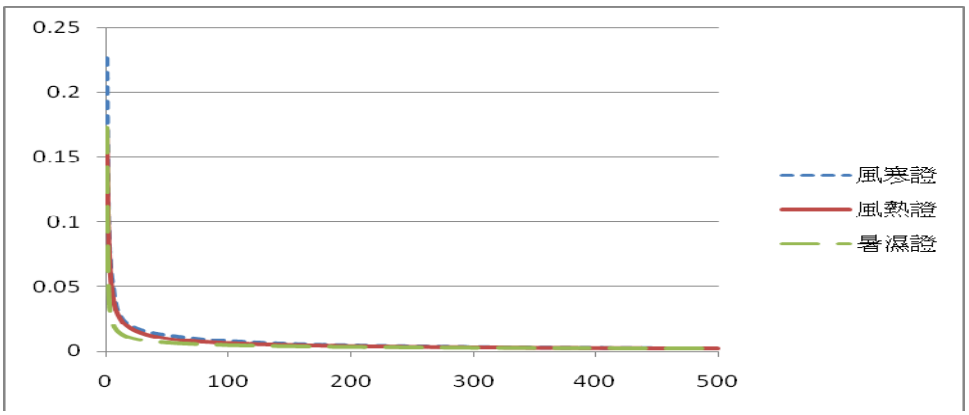


Figure 9.2.5.1.6b Explosion of the plot in Figure 9.2.5.1.6a

The result in Figure 9.2.5.1.6a shows 97.78% accuracy of the prediction result and after 500 training episodes, the NN RMSE converges and the resulting error is less than 0.01. Figure 9.2.5.1.6b is the explosion of the plot in Figure 9.2.5.1.6a.

Case 1b – 45 training data, 15 testing data, and weighting enabled:

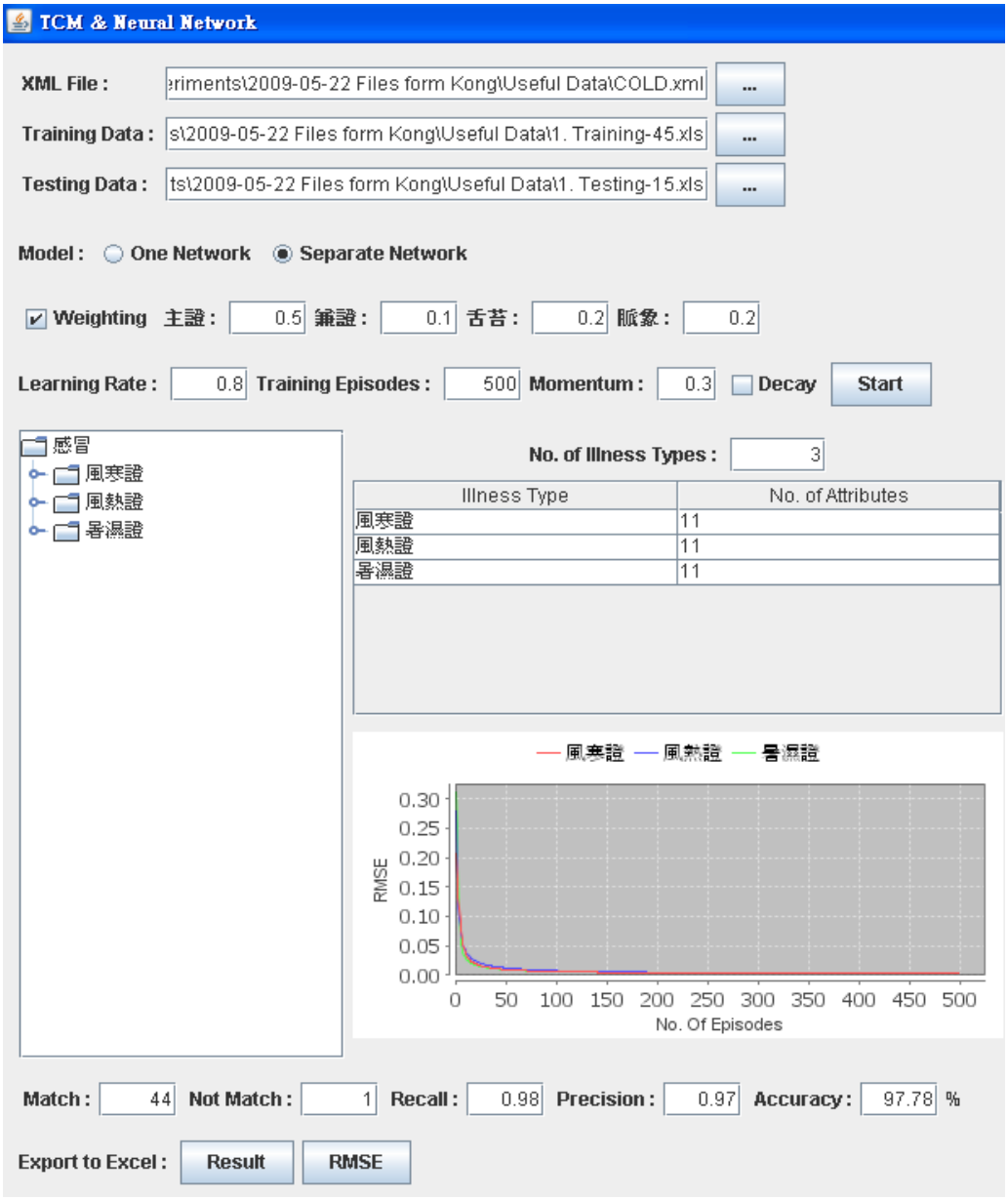


Figure 9.2.5.1.7a Result of model 2 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting enabled and decay disabled)

Figure 9.2.5.1.7a shows 97.78% accuracy of the NN prediction, and after 500 training episodes, the NN RMSE converges and the resulting error is less than 0.01.

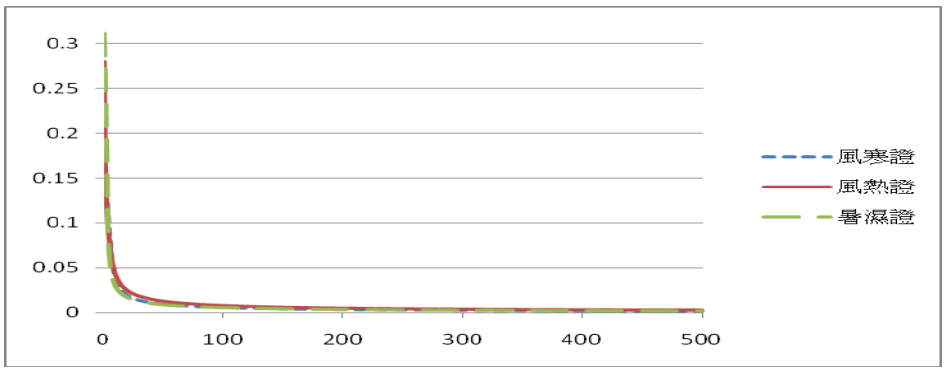


Figure 9.2.5.1.7b Explosion of the plot in Figure 9.2.5.1.7a

Case 2a – 60 training data, 60 testing data, and weighting disabled:

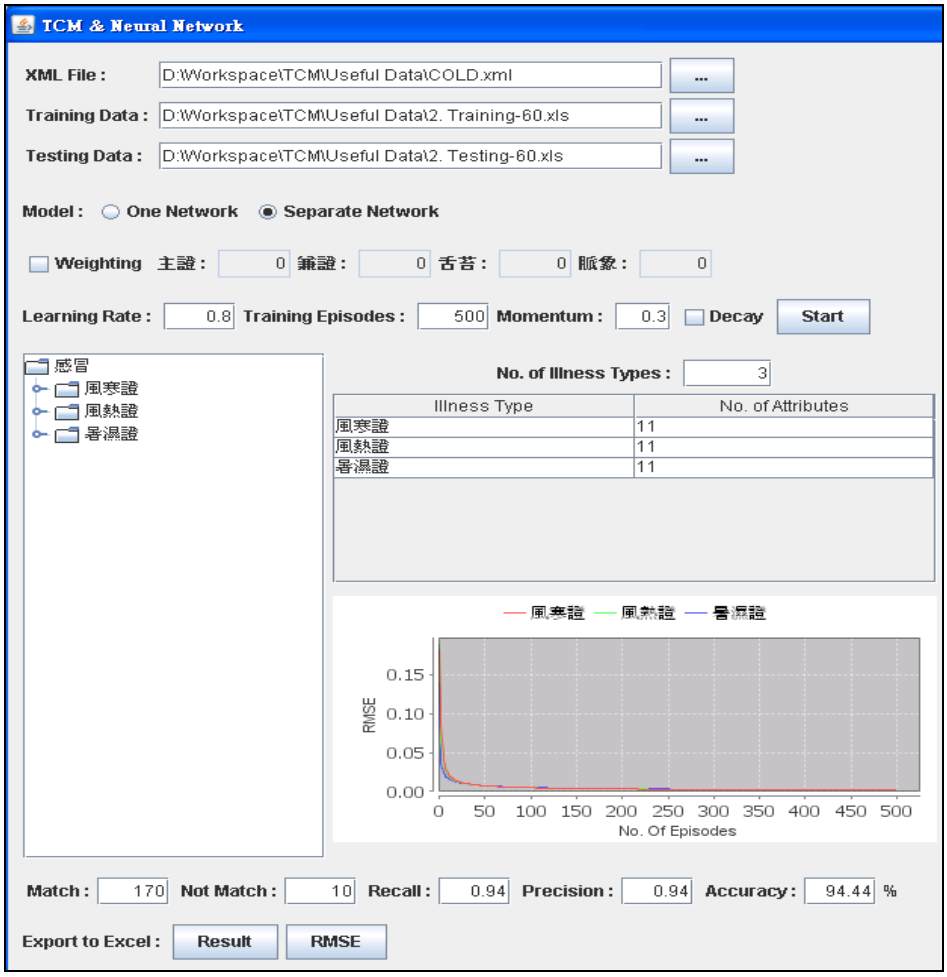


Figure 9.2.5.1.8 Result of model 2 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting and decay disabled)

Figure 9.2.5.1.8 shows 94.44% accuracy of the NN prediction, and after 100 training episodes, the NN RMSE converges and the resulting error is less than 0.01.

Case 2b – 60 training data, 60 testing data, and weighting enabled:

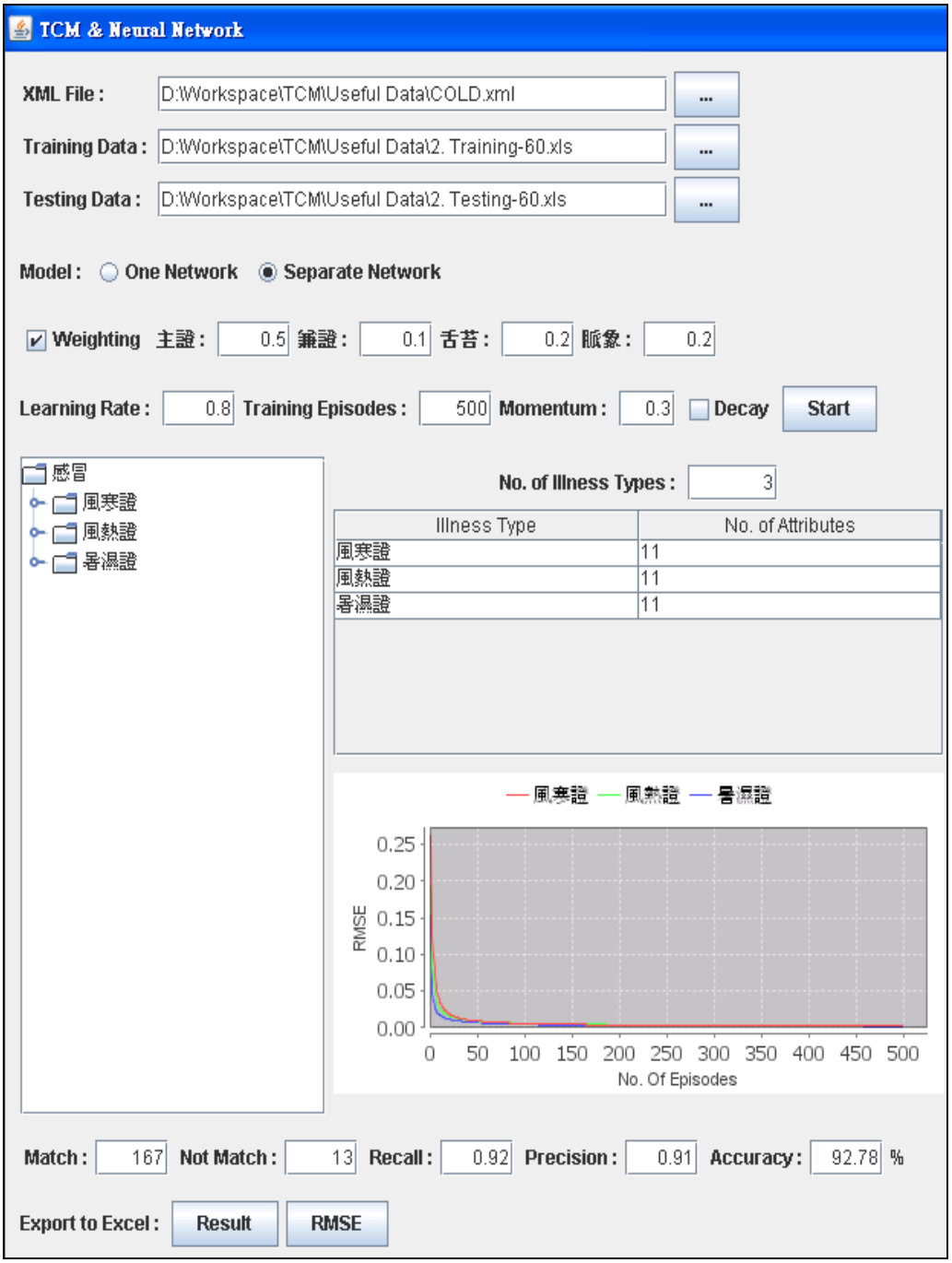


Figure 9.2.5.1.9 Result of model 2 (training episodes = 500, learning rate = 0.8, momentum = 0.3, weighting enabled and decay disabled)

Figure 9.2.5.1.9 shows 92.78% accuracy of the NN prediction, and after 100 training episodes, the NN RMSE converges and the resulting error is less than 0.01.

Result Analysis:

Case	Result
Case 1a – One Network (Weighting Disabled)	Accuracy: 100% Final RMSE: <0.001
Case 1b – One Network (Weighting Enabled)	Accuracy: 86.67% Final RMSE: <0.05
Case 2a – One Network (Weighting Disabled)	Accuracy: 96.67% Final RMSE: <0.001
Case 2b – One Network (Weighting Enabled)	Accuracy: 91.67% Final RMSE: <0.05
Case 1a – Separate Network (Weighting Disabled)	Accuracy: 97.78% Final RMSE: <0.01
Case 1b – Separate Network (Weighting Enabled)	Accuracy: 97.78% Final RMSE: <0.01
Case 2a – Separate Network (Weighting Disabled)	Accuracy: 93.89% Final RMSE: <0.01
Case 2b – Separate Network (Weighting Enabled)	Accuracy: 92.78% Final RMSE: <0.01

Table 9.2.5.1.2 Summary of the experimental results

The result in Table 9.2.5.1.2 shows a very high accuracy with the lowest 86.67% and the highest 100%; all named NN modules converge. There are differences for the RMSE of different illness types because each illness type is represented by one named NN module individually. All the name NN modules involved were trained with the same data set. The arc weights in each named NN modules are randomly initialized before training started. The RMSE changes in the named NN modules will be shown separately.

The user can define a specific tolerance level for a named NN module so that the fininding by this NN for a new inout set can be regard as acceptable. In this way the NN mdoules's performance can be tailored to the particular situations as required by the different users.

Timing Analysis:

Timing analysis estimates the limit of a tool in real-time application. If its execution time is too long, then its accuracy for fast computation is questionable. In this case, timing analysis is achieved by calculating the program execution time (by the java class *System.currentTimeMillis ()*). The CPU clock rate of the machine for this test is "Pentium 4 2.8GHz", and the formula used is: $ExecutionTime = \frac{ExecutionClockCycle}{ClockRate}$. The clock cycle is used as the unit of the program execution as it advances with the current technology. Table 9.2.5.1.3 lists some of the timing analysis results.

Case	No. of Data	Execution Time	Clock Cycle Needed (10^9)
One Network – Case 1a (Weighting Disabled)	Training: 45 Testing: 15	18203ms	50.9684
One Network – Case 1b (Weighting Enabled)	Training: 45 Testing: 15	16750ms	46.9000
One Network – Case 2a (Weighting Disabled)	Training: 60 Testing: 60	22546ms	63.1288
One Network – Case 2b (Weighting Enabled)	Training: 60 Testing: 60	23125ms	64.7500
Separate Network – Case 1a (Weighting Disabled)	Training: 45 Testing: 15	9797ms	27.4316
Separate Network – Case 1b (Weighting Enabled)	Training: 45 Testing: 15	9313ms	26.0764
Separate Network – Case 2a (Weighting Disabled)	Training: 60 Testing: 60	13203ms	36.9684
Separate Network – Case 2b (Weighting Enabled)	Training: 60 Testing: 60	12500ms	35.0000

Table 9.2.5.1.3 Some relevant timing analysis results

9.2.5.2 Discovery of Individual Herbal Ingredients (Low-level)

To facilitate the verification experiments sixty herbal ingredients were selected, as shown in Table 9.2.5.2.1.

(Remarks: lt - Latin; and py - pinyin in Putonghua)

感冒	不寐	便秘
甘草 (lt – Radix Glycyrrhizae) (py - Gan Cao)	知母 (lt – Rhizoma Anemarrhenae) (py - Zhi Mu)	大黃 (lt – Radix et Rhizoma Rhei) (py - Dai Huang)
桔梗 (lt – Radix Platycodi) (py - Jie Geng)	白朮 (lt – Rhizoma Atractylodis) (py - Bai Zhu)	當歸 (lt- Radix Angelicae) (py - Dang Gui)
薄荷 (lt –Herba Menthae) (py - Bo He)	甘草(炙) (lt - Radix Glycyrrhizae (roasted)) (py - Gan Cao (Zhi))	陳皮 (lt – Oericarpium Citri Reticulatae) (py - Chen Pi)
白芷 (lt – Radix Angelicae Dahuricae (py - Bai Zhi)	酸棗仁 (lt – Seman Ziziphi Spinosae) (py - Suan Zao Ren)	黨參 (lt – Radix Codonopsis Pilosulae) (py - Dang Shen)
生薑 (lt – Rhizoma Zingiberis Recens) (py - Sheng Jiang)	合歡皮 (lt – Cortex Albizziae) (py - He Huan Pi)	火麻仁 (lt – Fructus Cannabis) (py - Huo Ma Ren)
防風 (lt – Radix Ledebouriellae) (py - Fang Feng)	川芎 (lt – Rhizoma Ligustici) (py - Chuan Xiong)	麥冬 (lt – Radix Ophiopogonis) (py - Mai Dong)
連翹 (lt – Fructus Forsythiae) (py - Lian Qiao)	白芍 (lt – Radix Paeoniae Alba) (py - Bai Shao)	黃蓍 (lt – Radix Astragali seu Hedysari) (py - Huang Shi/Qi)
牛蒡子 (lt – Fructus Arctii)	茯苓 (lt – Poria)	玄參 (lt – Radix Scrophulariae)

(py - Niu Bang Zi)	(py - Fu Ling)	(py - Xuan Shen)
荆芥 (lt – Herba Schizonepetae) (py - Jing Jie)	地骨皮 (lt – Cortex Lycii Radicis) (py - Di Gu Pi)	黃芩 (lt – Radix Scutellariae) (py - Huang Qin)
葛根 (lt – Radix Puerariae) (py - Ge Gen)	大棗 (lt – Fructus Ziziphi Jujubae) (py - Da Zao)	枳實 (lt – Fructus Aurantii Immaturus) (py - Zhi Shi)
天花粉 (lt – Radix Trichosanthis) (py - Tian Hua Fen)	(牡)丹皮 (lt – Cortex Moutan) (py – (Mu) Dan Pi)	梔子 (lt – Fructus Gardeniae) (py - Zhi Zi)
柴胡 (lt – Radix Bupleuri) (py - Chai Hu)	首烏藤 (lt – Caulis Polygoni Multiflori) (py - Shou Wu Teng)	地黃 (lt – Radix Rehmanniae); raw (py - Di Huang)
金銀花 (lt – Flos Loncierae) (py - Jin Yin Hua)	熟地黃 (lt – Radix Rehmanniae); cooked (py - Shu Di Huang)	澤瀉 (lt – Rhizoma Alismatis) (py - Ze Xie)
辛夷 (lt – Flos Magnoliae) (py - Xin Yi)	遠志 (lt – Radix Polygalae) (py - Yuan Zhi)	厚樸 (lt – Cortex Magnoliae Officinalis) (py - Hou Po)
荆芥穗 (lt – Herba Schizonepetae) (py - Jing Jie Sui)	枸杞子 (lt – Fructus Lycii) (py - Gou Qi Zi)	肉蓯蓉 (lt – Herba Cistanches) (py - Rou Cong Rong)
羌活 (lt – Rhizoma seu Radix Notopterygii) (py - Qiang Huo)	丹參 (lt – Radix Salviae Miltiorrhizae) (py - Dan Shen)	山茱萸 (lt – Fructus Corni) (py - Shan Zhu Yu)
蘆根 (lt – Rhizoma Phragmitis) (py - Lu Gen)	黃柏 (lt - Cortex Phellodendri) (py - Huang Bo)	(幹)山藥 (lt – Rhizoma Dioscoreae) (py – (Gan) Shan Yao)
浙貝母 (lt – Bulbus Fritillariae Thunbergii) (py - Zhe Bei Mu)	人參 (lt – Radix Ginseng) (py - Ren Shen)	薏苡仁 (lt – Cemen Coicis) (py - Yi Yi Ren)
淡竹葉 (lt – Herba Lophatheri) (py - Dan Zhu Ye)	山楂(lt – Fructus Crataegi) (py - Shan Zha)	砂仁 (lt – Fructus Amomi) (py - Sha Ren)
蒼耳子 (lt – Fructus Xanthii) (py - Cang Er Zi)	石菖蒲 (lt – Rhizoma Acori Graninei) (py - Shi Chang Pu)	苦杏仁 (lt – Semen Armeniacae Amarum) (py - Ku Xing Ren)

Table 9.2.5.2.1 Sixty different herbal items selected for the experiments

Three different illnesses' sub-ontologies, which are isolated from the Nong's enterprise TCM ontology for the verification purpose here, are annotated by the XML metadata system (Appendix VII, VIII and IX). For the sake of demonstration, the <甘草> (<Gancao>) herbal sub-ontology is also isolated from the Nong's enterprise TCM ontology, as shown in Figure 9.2.5.2.1.

<甘草> (<Gancao>)

<性味> (<Nature & Flavor of Medicinals>)

<甘>id="1"</甘> (<Sweet>id="1"</Sweet>)

<平>id="1"</平> (<Calm>id="1"</Calm>)

</性味> (</Nature & Flavor of Medicinals>)

<歸經> (<Meridian Entry>)

<心>id=”1”</心> (<Heart>id=”1”</Heart>)

<肺>id=”1”</肺> (<Lung>id=”1”</Lung>)

<脾>id=”1”</脾> (<Spleen>id=”1”</Spleen>)

<胃>id=”1”</胃> (<Stomach>=”1”</Stomach>)

</歸經> (</Meridian Entry>)

<功能主治>

<補脾益氣>id=”1”</補脾益氣>

(<Tonify the Spleen & Replenish Qi>id=”1”</Tonify the Spleen & Replenish Qi>)

<清熱解毒>id=”1”</清熱解毒>

(<Clear Heat and Detoxify>id=”1”</Clear Heat and Detoxify>)

<祛痰止咳>id=”1”</祛痰止咳>

(<Resolve Phlegm to Suppress Cough>id=”1”</Resolve Phlegm to Suppress Cough>)

<緩急止痛>id=”1”</緩急止痛>

(<Relax Tension to Relieve Pain>id=”1”</Relax Tension to Relieve Pain>)

</功能主治>

</甘草> (</Gancao>)

Figure 9.2.5.2.1 Gancao sub-ontology in XML

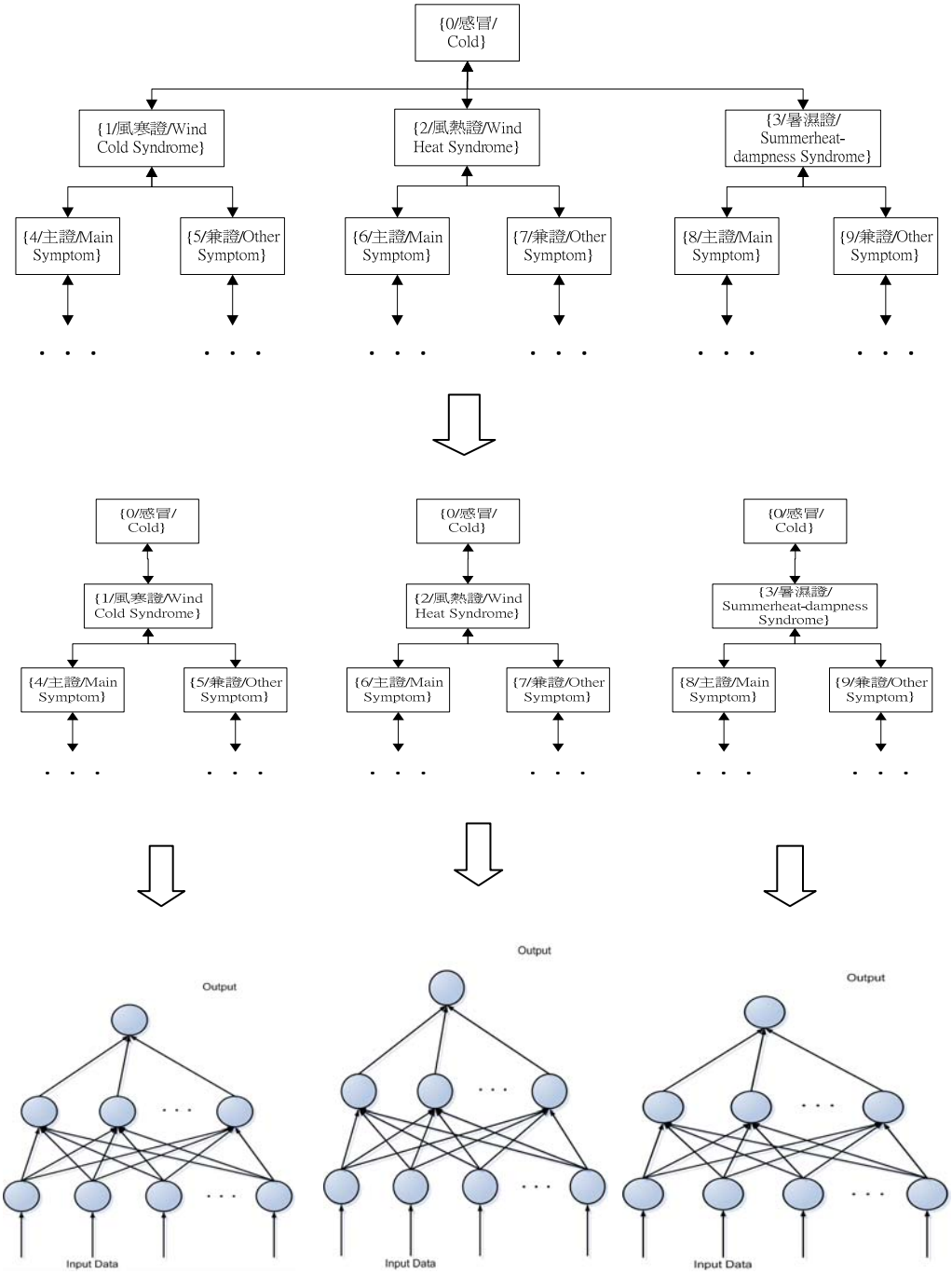


Figure 9.2.5.2.2 Partitioning the NN

In this section, we will show how a DOM (Document Object Model) tree (or semantic net as in section 7.8) can be partitioned into several smaller DOM tree modules of the following characteristics: i) each tree module represents only one illness type (or alternatively one herbal item); ii) the

number of NN modules is equal to the number of illness types (or herbal items); iii) the number of input parameters to each NN is equal to the number of symptoms pre-defined for each illness type (or each herbal item) (i.e. 1 $\Rightarrow$ exist; 0 $\Rightarrow$ absent; $\Rightarrow$ means implying); iv) the number of input neurons is equal to the number of pre-defined symptoms for each illness type; v) the number of hidden neurons in the NN is twice the input neurons; and vi) the number of output neurons is one (i.e. for the relevance index (RI)). Figure 9.2.5.2.2 shows the partitioning process.

9.2.5.2.1 Result and Analysis

Similar to the knowledge classification mechanism introduced in Chapter 7, input data has to be preprocessed before it enters NN. Attributes/symptoms and relevant classes are identified and extracted from the input XML file. Then, attributes (extracted from a patient's record) are selected as the input signal. The NN prototypes are trained with selected clinical data from Nong's (with permission). For example, a Nong's clinical record (i.e. prescription number "PR0401000037") is useful:

i) Statement/syndrome: "流清涕 怕冷重 發熱 頭痛 口乾 頭暈 脈象浮細 淡紅苔薄" (Sniffles, aversion to cold, fever, headache, dry mouth, dizziness, fine floating pulse, pale red tongue of thin fur)
ii) Diagnosis: "感冒 - 風寒束表" (Cold – Wind cold syndrome)

Figure 9.2.5.2.1.1 shows the different sections in the GUI through which experiments can be conducted.

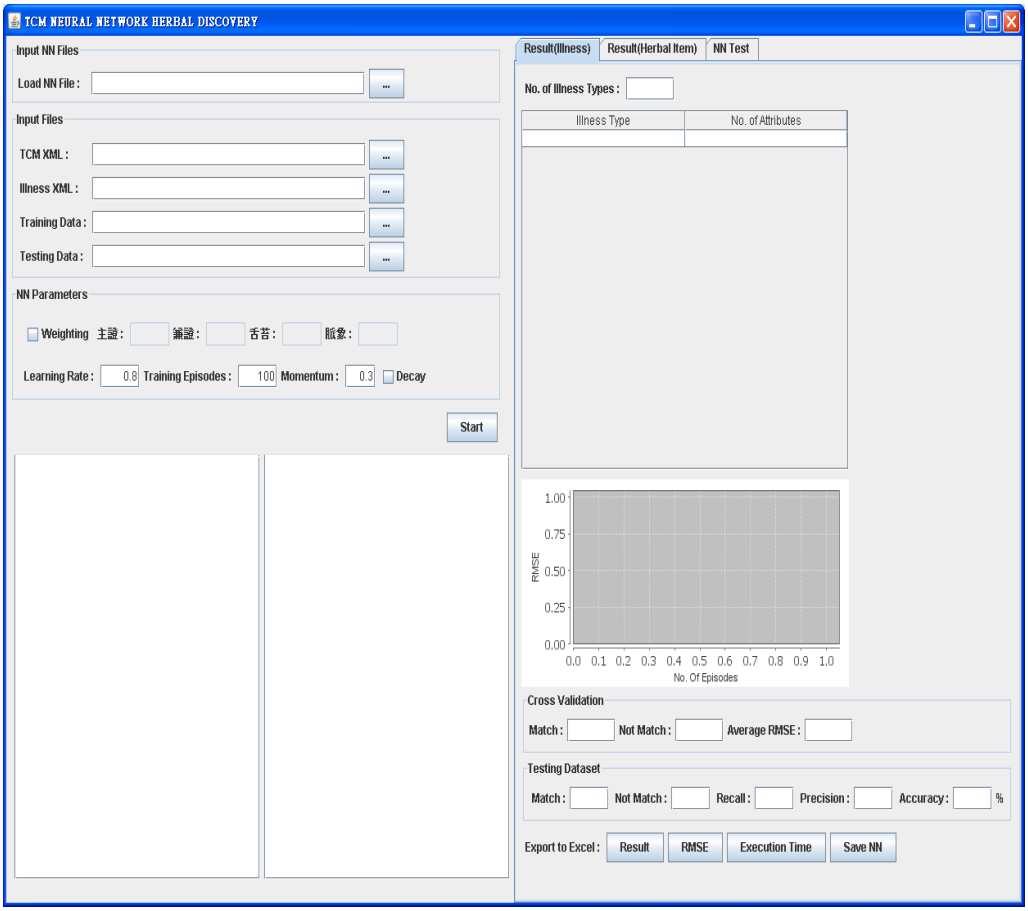


Figure 9.2.5.2.1.1 The GUI of the herbal discovery model

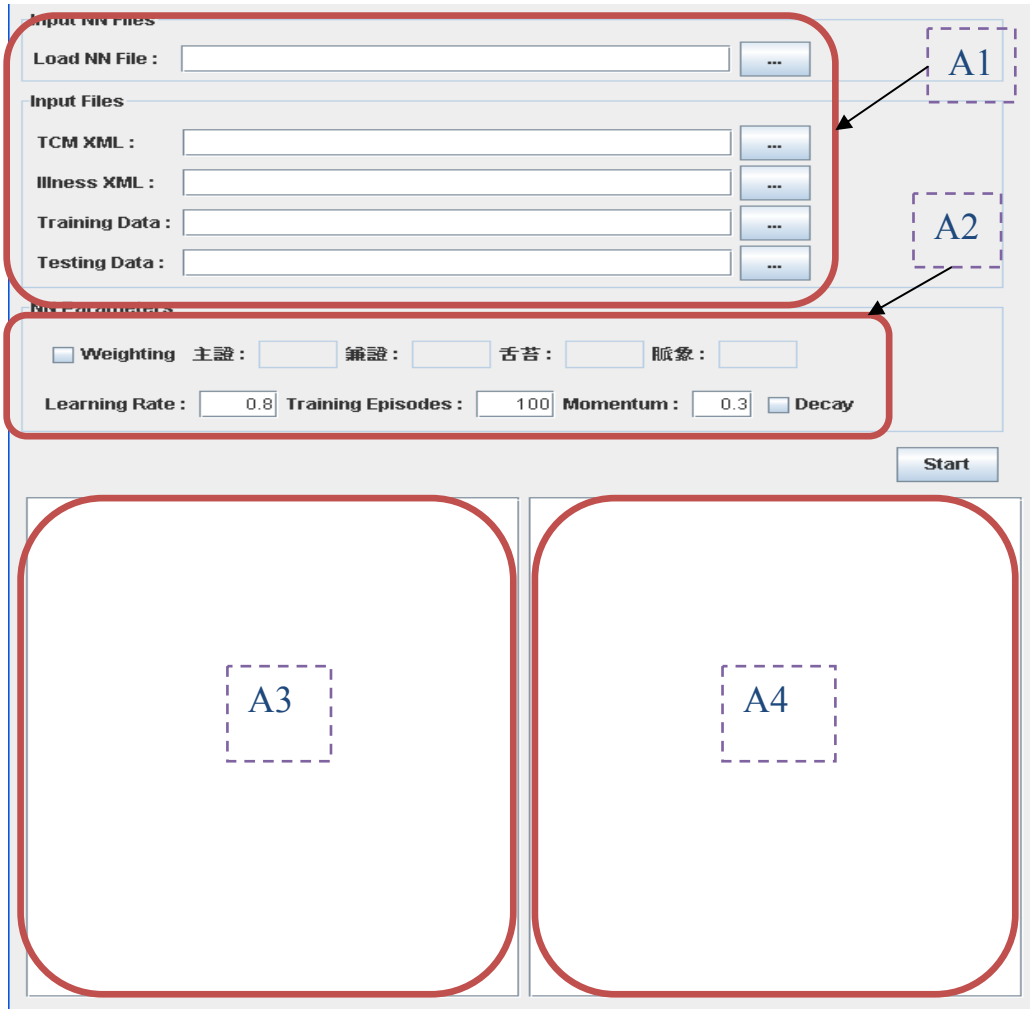


Figure 9.2.5.2.1.2 Experiment environment setting (section A)

The functions of the four different GUI sections (Figure 9.2.5.2.1.2) are:

Section A (Figure 9.2.5.2.1.2):

a) A1 – Input file path for the herbal (TCM) and illness XML file, training dataset and testing dataset. To recall an existing network, it can be selected in the "Load NN File" path.

b) A2 – Weight setting of 主證 (Main Symptom), 兼證 (Other Symptom),

舌苔 (Tongue) and 脈象 (Pulse), and parameters for the neural network, including: learning rate, training episode, momentum and decay). Section 9.2 contains the relevant details.

c) A3 – XML DOM Tree visualizing section (for illnesses).

d) A4 – XML DOM Tree visualizing section (for herbal items).

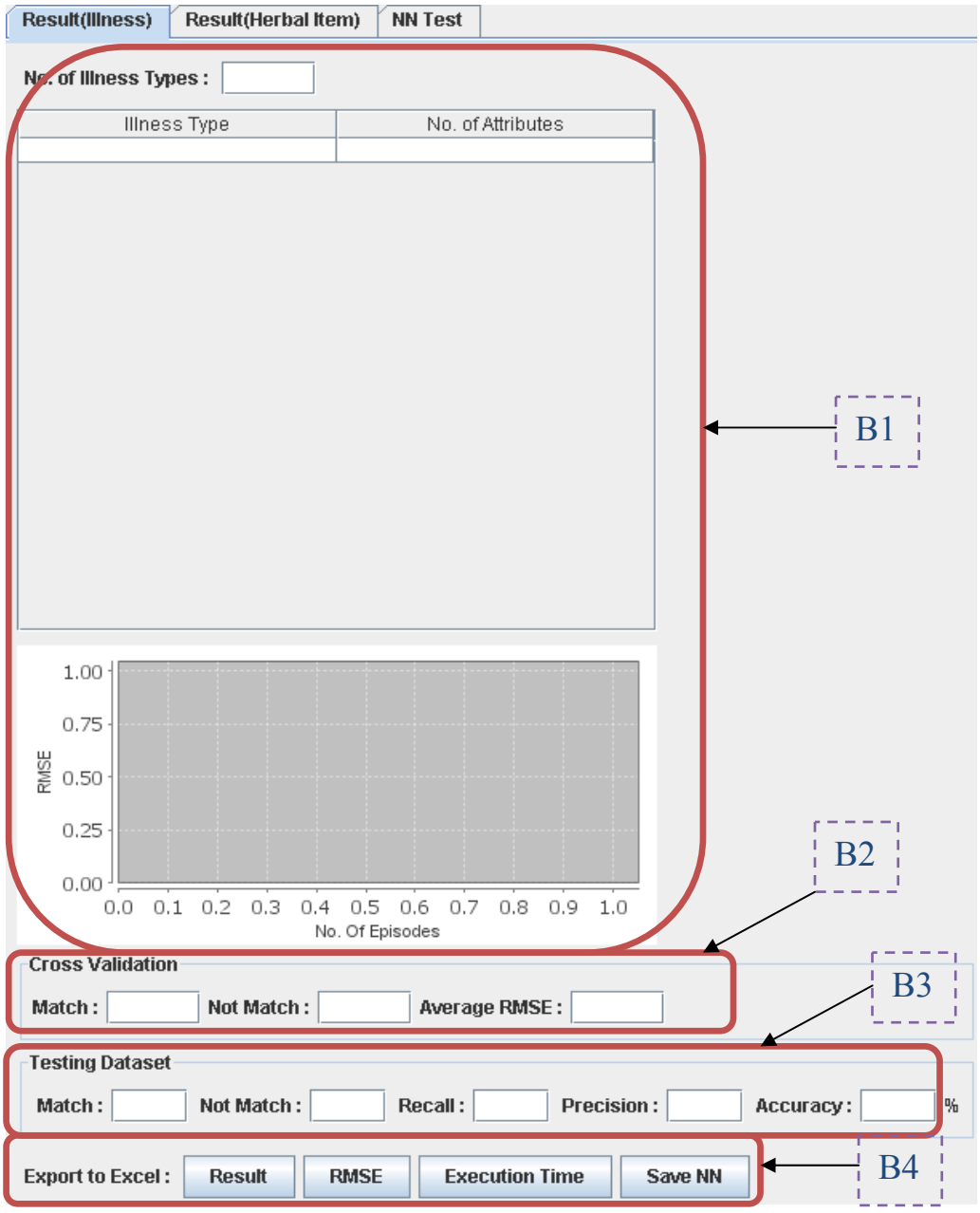


Figure 9.2.5.2.1.3 Function of tab 1 – result (illness) (section B)

Section B (Figure 9.2.5.2.1.3):

a) B1 – Showing the number of illness types in XML file, number of attributes of each illness type and the RMSE during the training process of the neural network.
b) B2 – Using the training dataset to perform cross-validation (using RMSE as the performance measurement).
c) B3 – Using the testing dataset to validate the performance of the neural network (using accuracy, precision and recall as the performance measurement).
d) B4 - Result export to Excel file (e.g. result, RMSE, execution time, and save the NN into a file for future recall).

Section C (Figure 9.2.5.2.1.4):

a) C1 – Showing the suggested herbal items used for every relevant illness type, with the item name, RI score, nature and flavor of medicinal, meridian entry and functions & indications shown.
b) C2 – Result export to Excel file (e.g. result for herbal item and RI score of all items with respect to each illness type).

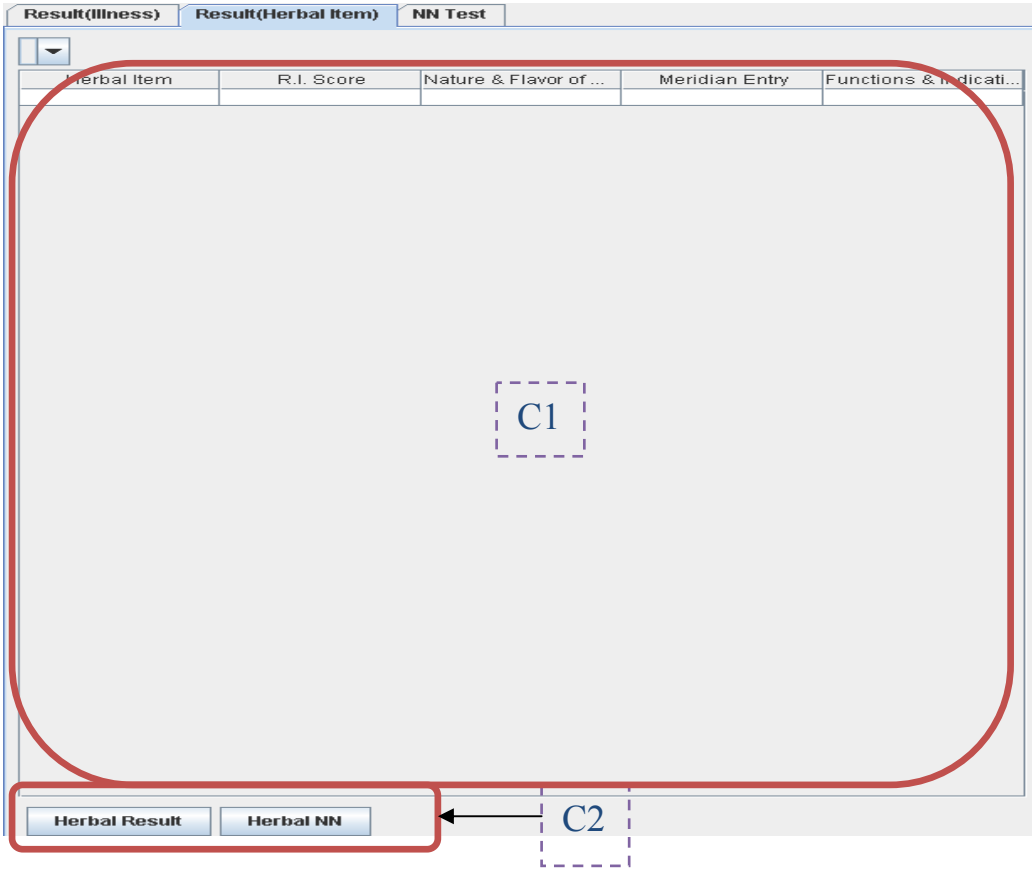


Figure 9.2.5.2.1.4 Function of tab 2 – result (herbal item) (section C)

Section D (Figure 9.2.5.2.1.5):

a) D1 – Symptoms selection: The selected symptoms will be echoed to the text field below as the input of the NN model.

b) D2 – (Same as C1): – Showing the suggested herbal items used for every relevant illness types, with the item name, RI score, nature & flavor of medicinal, meridian entry and functions & indications shown.

c) D3 – (Same as C2): Using the training dataset to perform cross-validation (using RMSE as the performance measurement).

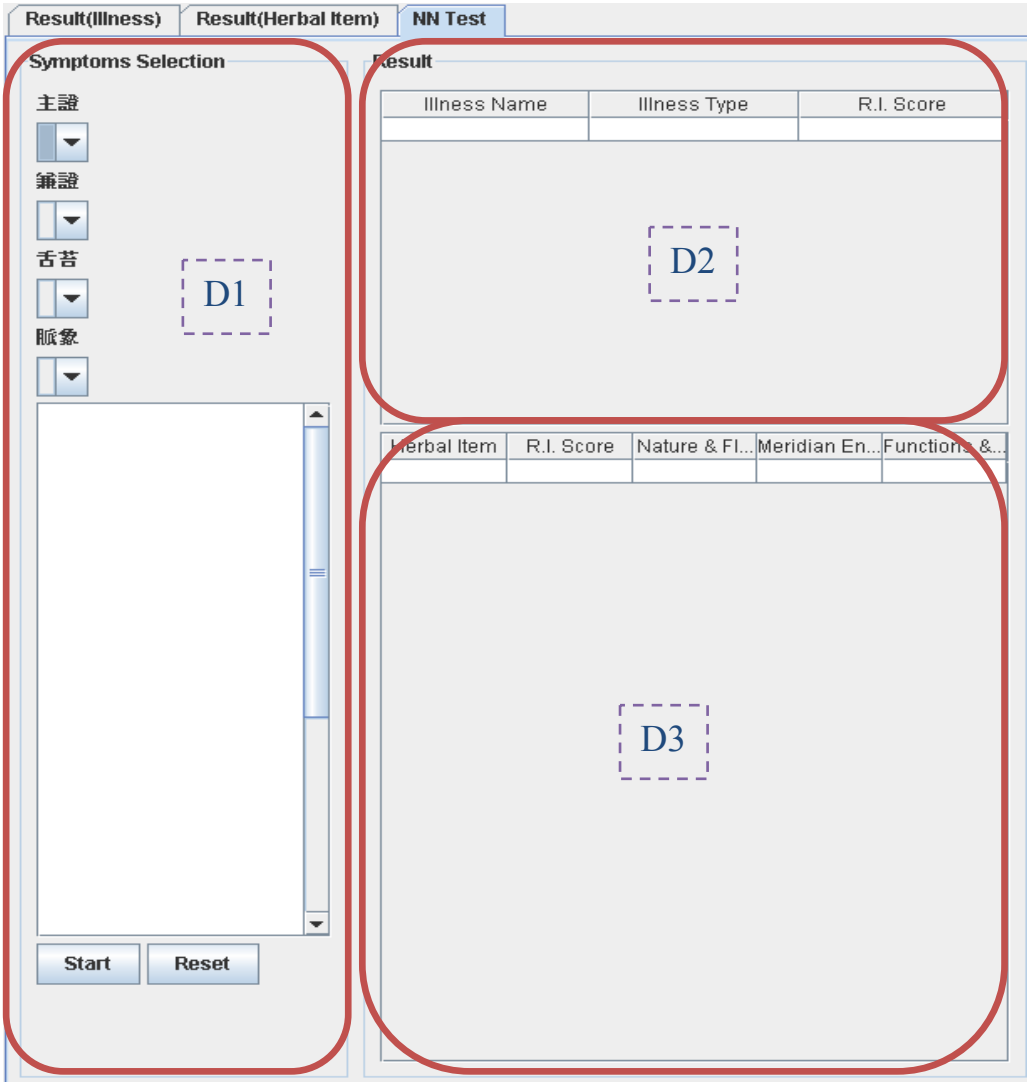


Figure 9.2.5.2.1.5 Function of tab 3 – NN test (section D)

The XML files that could be used to excite the named NN modules are shown in the Figures 9.2.5.2.1, 9.2.5.2.2, and 9.2.5.2.3. Tables 9.2.5.2.1.1 and 9.2.5.2.1.2 show the relevant medical records in the training dataset.

In Table 9.2.5.2.1.3, besides 九味羌活湯 (Jiu Wei Qiang Huo Tang), all other items are single herbal items. 九味羌活湯 (Jiu Wei Qiang Huo Tang) is a *combo* (複方) or "*ancient*" formula. As the aim of this experiment is to discover single herbal items, with Nong's permission, I can make use of the detailed

information of the combo formula from the Nong's product. In the experiments, only the herbal items are considered, but not the quantity (dosage) of use. For the RI calculation in the sections C and D (Tab 2 and Tab 3), the scores are generated by the NN model in the <0,1> range.

Prescription No.	Statement/ syndrome	History	Pulse	Tongue	Diagnosis
PR0401000037	流涕三天 (Sniffles for 3 days)	流清涕怕冷重 發熱 頭痛 口乾 頭暈 (Sniffles, aversion to cold, fever, headache, dry mouth, dizziness)	脈象浮細 (Fine and floating pulse)	淡紅苔薄 (Pale red tongue with thin fur)	風寒形 (Wind cold syndrome)

Table 9.2.5.2.1.1 A record in the training dataset (illness)

Prescription No.	Name of items in the prescription (PR no.)	Quantity	Unit
PR0401000037	九味羌活湯 (Jiu Wei Qiang Huo Tang)	10	Gram
PR0401000037	白朮 (Baizhu)	2	Gram
PR0401000037	炒六神曲 (Shenqu Cha)	3	Gram
PR0401000037	陳皮 (Chenpi)	2	Gram
PR0401000037	法半夏 (Fabanxia)	2	Gram
PR0401000037	生姜 (Shengjiang)	2	Gram
PR0401000037	大棗 (Dazao)	3	Gram

Table 9.2.5.2.1.2 A record in the training dataset (prescription)

Nong's Product Code	Name of Combo (複方) Formula (*Combo means ancient recipe*)	Name of Herbal Items in the combo 2029	Quantity
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	羌活 (Qianghuo)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	防風 (Fangfeng)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	蒼朮 (Cangzhu)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	細辛 (Xixin)	0.1
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	川芎 (Chuanxiong)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	白芷 (Baizhi)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	黃芩 (Huangqin)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	甘草 (Gancao)	1.0
2029	九味羌活湯 (Jiu Wei Qiang Huo Tang)	生地黃 (Dihuang)	1.0

Table 9.2.5.2.1.3 An example of the breakdown of a combo (複方) formula

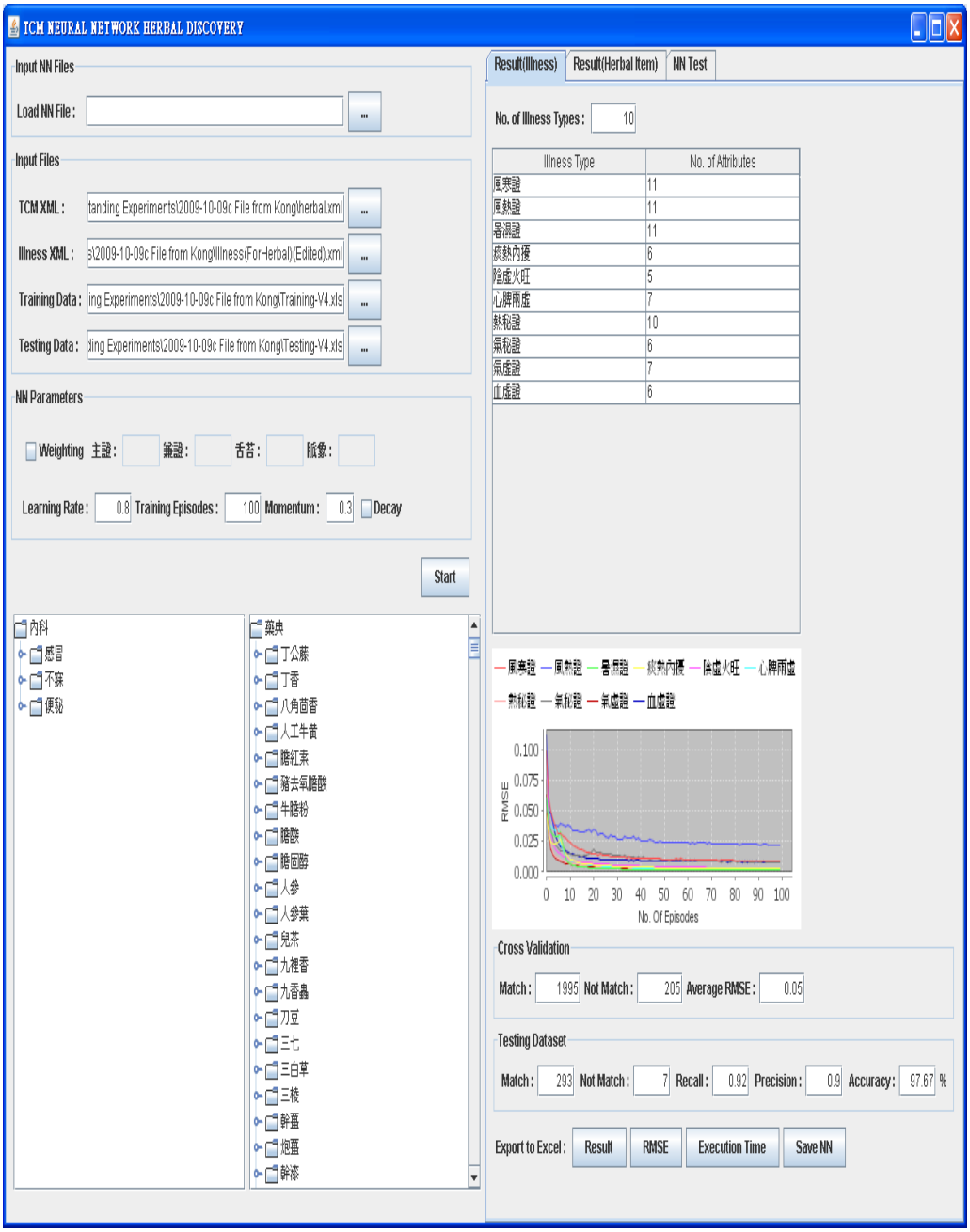


Figure 9.2.5.2.1.6 Training result for a named NN module

Figure 9.2.5.2.1.6 shows how NN training is conducted and accepted for the named NN module. This training only involves three illnesses (i.e. Flu (感冒), Insomnia (不寐) and Constipation (便秘)), which covers ten illness types. For example, Flu (感冒) can be divided into three types: ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證) and ***Summer-Heat Dampness***

Syndrome (暑濕證), as shown in the box on the top right corner of Figure 9.2.5.2.1.6. The number of attributes/symptoms versus the illness types is also shown (e.g. ***Wind Cold Syndrome*** (風寒證) is identified by 11 standard/classical symptoms. On the left bottom corner the herbs that can treat the three illnesses are partially listed. The plot in Figure 9.2.5.2.1.6 relates the RMSE (root mean square error) value with the number of training episodes (rounds) with the same dataset. More rounds mean higher accuracy because of the lower RMSE value. Since the training does not depend on the value of the teacher signal but rather the acceptance level specified/agreed by the user, namely, the "***Accuracy***", which is equal to "$1 - RMSE$", training can stop provided that $Accuracy \geq Specification$. For example, if the specification for this training session is 97%, then no more training episode will be needed because the attained accuracy is 97.67%. That is, the named NN module is ***well-trained*** and ready to be used. The RI values of all the well-trained NN modules will form the 3-dimensional ***basic referential plot*** (BRP) as shown in Figure 8.4.2.15a.

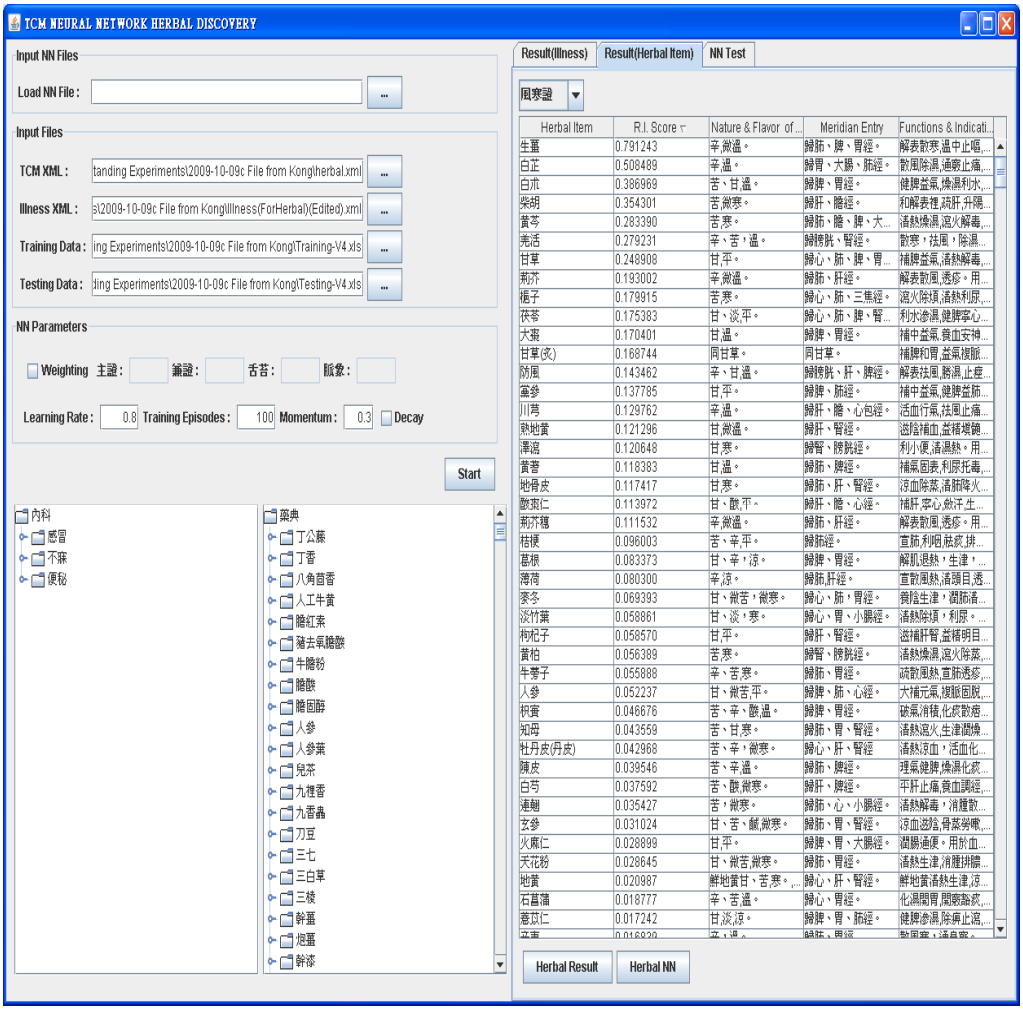


Figure 9.2.5.2.1.7 Experimental result

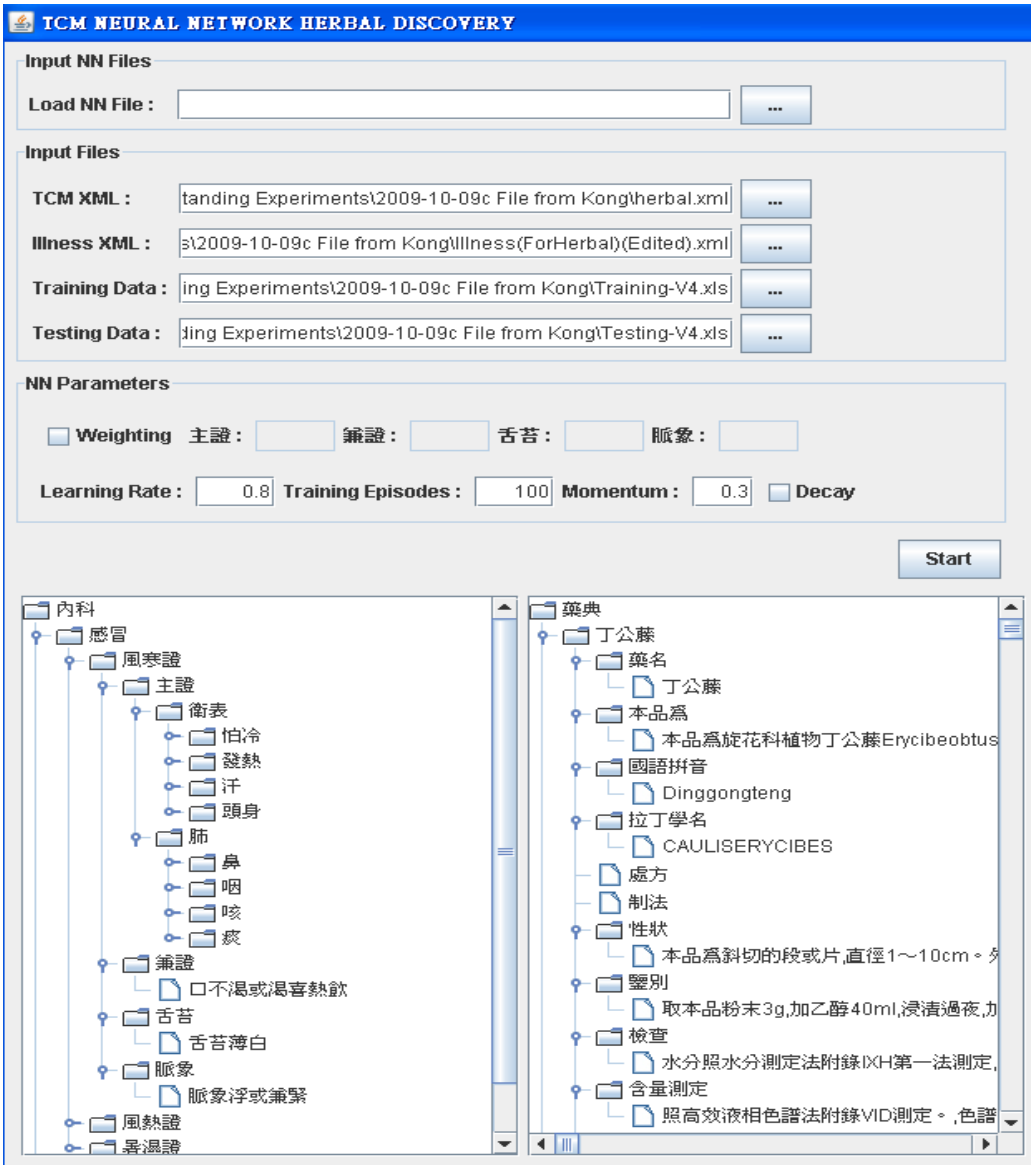


Figure 9.2.5.2.1.8 Experimental result (NN parameters and two XML DOM trees)

Figure 9.2.5.2.1.8 shows part of the RI scores of the ***basic referential plot*** (BRP) as shown in Figure 8.4.2.15a. The functions of every herbal items and its effect on the meridian (one or many) is also shown. Basically, Figure 9.2.5.2.1.8 is an enlargement of the box in the left bottom corner Figure 9.2.5.2.1.7. Figure 8.4.2.15a shows another respective of Figure 9.2.5.2.1.7 and Figure 9.2.5.2.1.8. The two boxes representing the accentuations of two

knowledge categories in the Nong's TCM enterprise ontology upon which the experiments are based. The left box shows that the accentuation is on the illnesses and their respective standard symptoms. The right box accentuates the classical information about the individual herbal items.

Result Analysis (Tab 1- Result (Illness)):

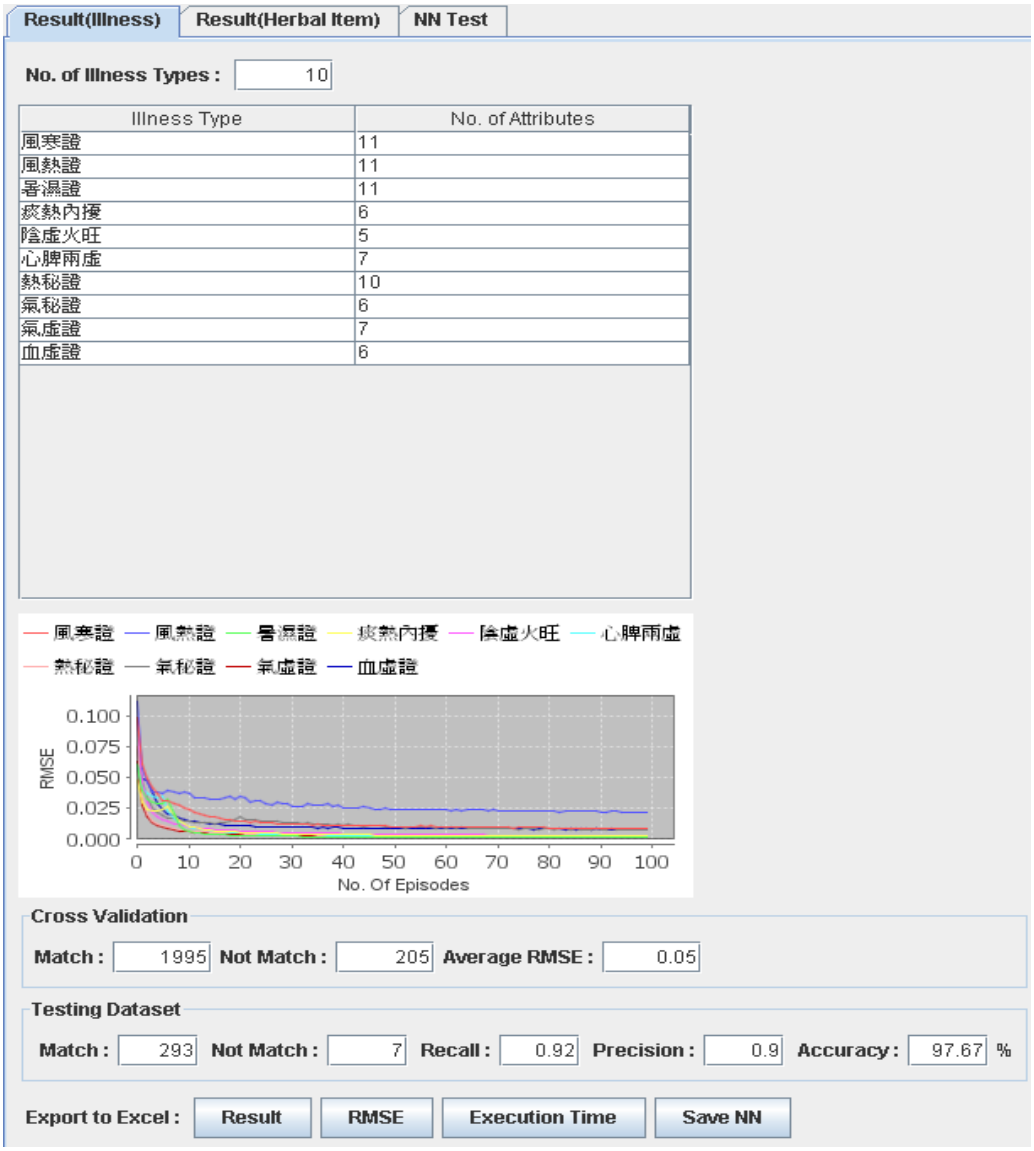


Figure 9.2.5.2.1.9 Experimental result (NN parameters and two XML DOM trees)

Figure 9.2.5.2.1.9 is the explosion of the right-hand side of Figure 9.2.5.2.1.7. The 10 illness types over the plots is to remind the reader that the training of the name NN module is based mainly on its relevance to the syndromes (i.e. a set of symptoms), and the illness types are inert parameters to

help the 3-dimensional visualization only (i.e. the (BRP) as shown in Figure 9.2.5.2.1.14a).

For the cross validation section, the result shows a high accuracy and the resulting average RMSE is 0.05. For the testing dataset, the result shows 97.67% accuracy of the prediction result and after 100 training episodes, the RMSE of the neural networks converges and the resulting error is less than 0.025. The following three figures show the RMSE changes versus episodes in the training process in light of the three illnesses.

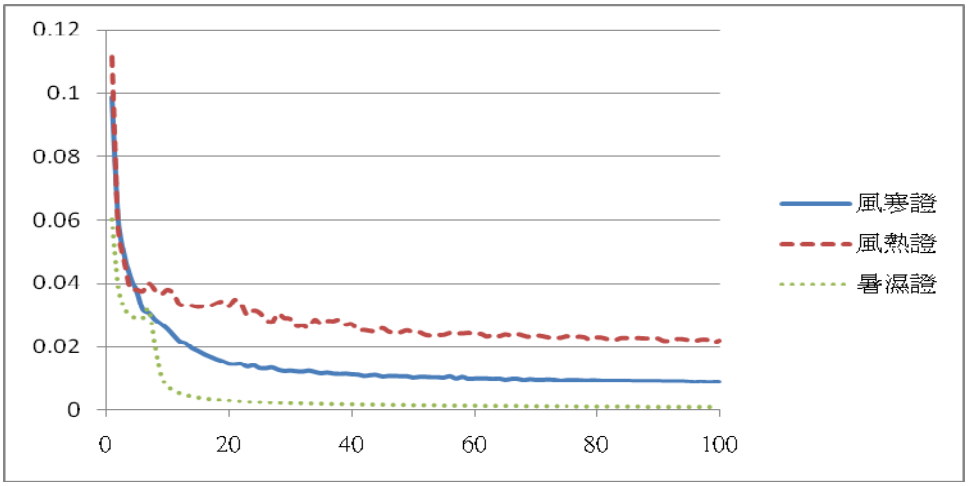


Figure 9.2.5.2.1.10 RMSE changes of the "Flu" NN of module during training

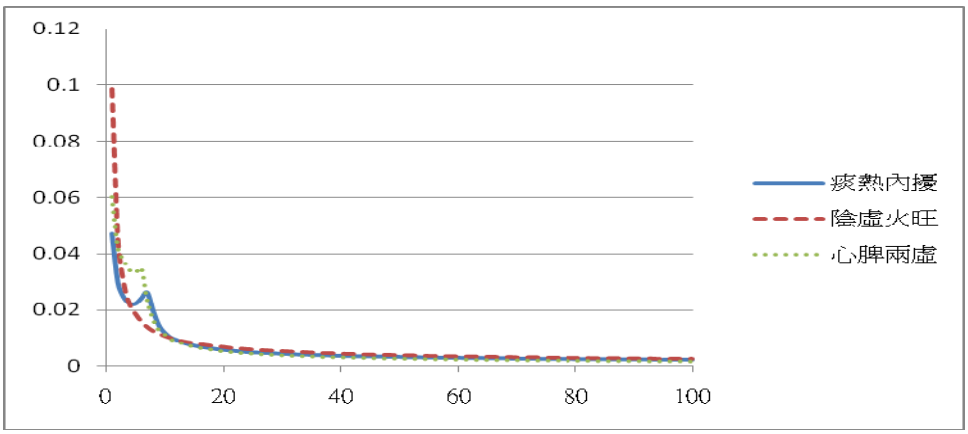


Figure 9.2.5.2.1.11 RMSE changes of the "Insomnia" NN during training

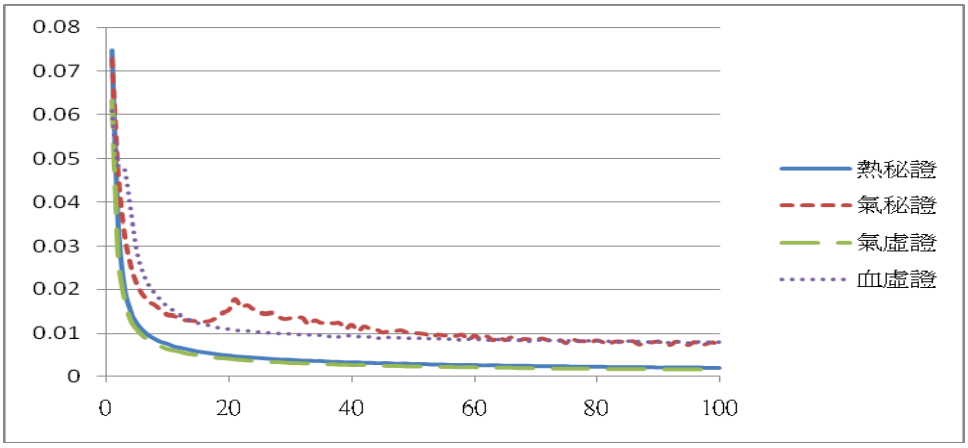


Figure 9.2.5.2.1.12 RMSE of changes of "Constipation" NN during training

Result Analysis (Tab 2 – Result (Herbal Item)):

In Figure 9.2.5.2.1.9, the combo (複方) box at the top contains all 10 illness types in this experiment. In the table, the five columns are: i) herbal item (name); ii) RI score (of each herbal item); iii) nature and flavor; iv) meridian entry; and v) functions and indications. When the Tab 2 function is invoked all the 60 herbs and their corresponding collective RI values computed during

training are listed as shown in Figure 9.2.5.2.1.13. The plotting of these RI values would form the 3-dimensional BRP, shown in Figure 9.2.5.2.1.14a.

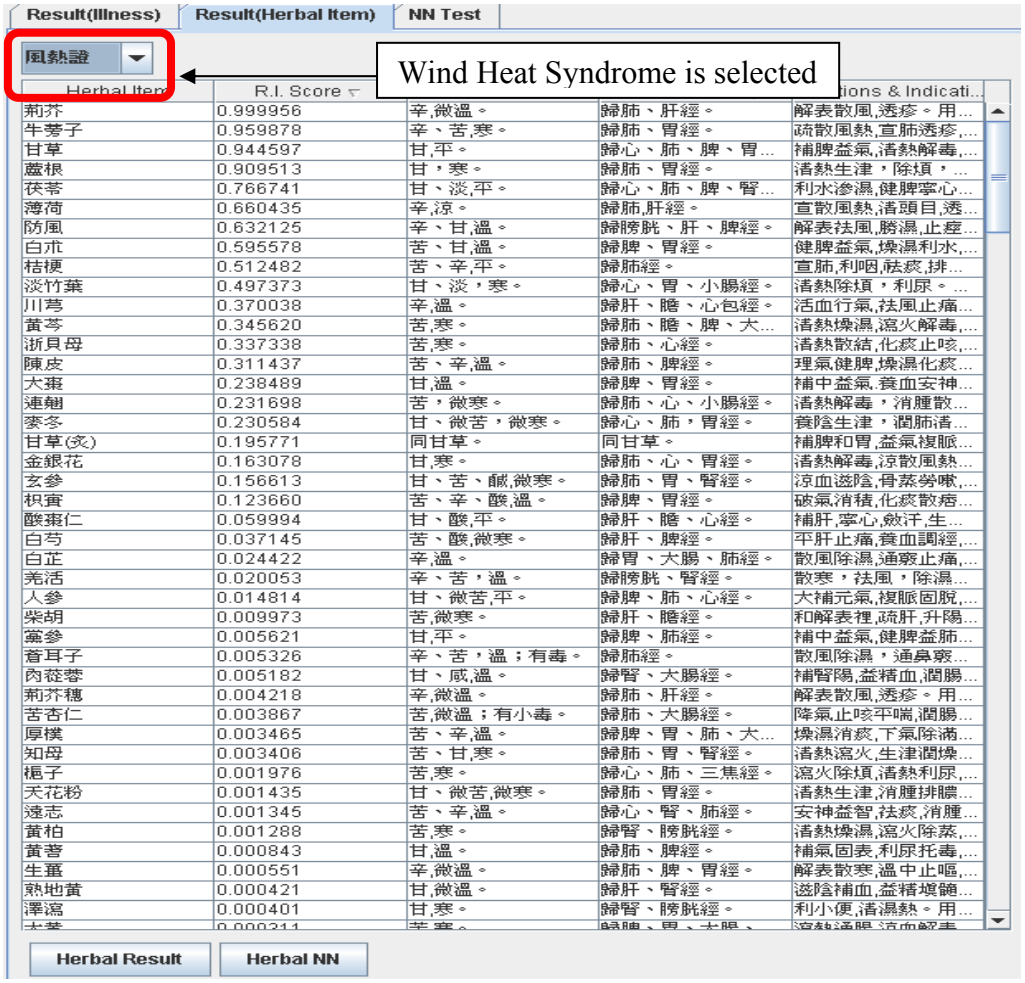


Figure 9.2.5.2.1.13 Herbal items for each illness type (in this case – Wind-Heat)

For example, the ***class (證形)- Wind Heat Syndrome (風熱證)*** is selected, Figure 9.2.5.2.1.13 shows all relevant single herbal items that would treat this illness type. The herbal item of a higher RI score would treat the illness more effectively. The other three columns provide some reference information.

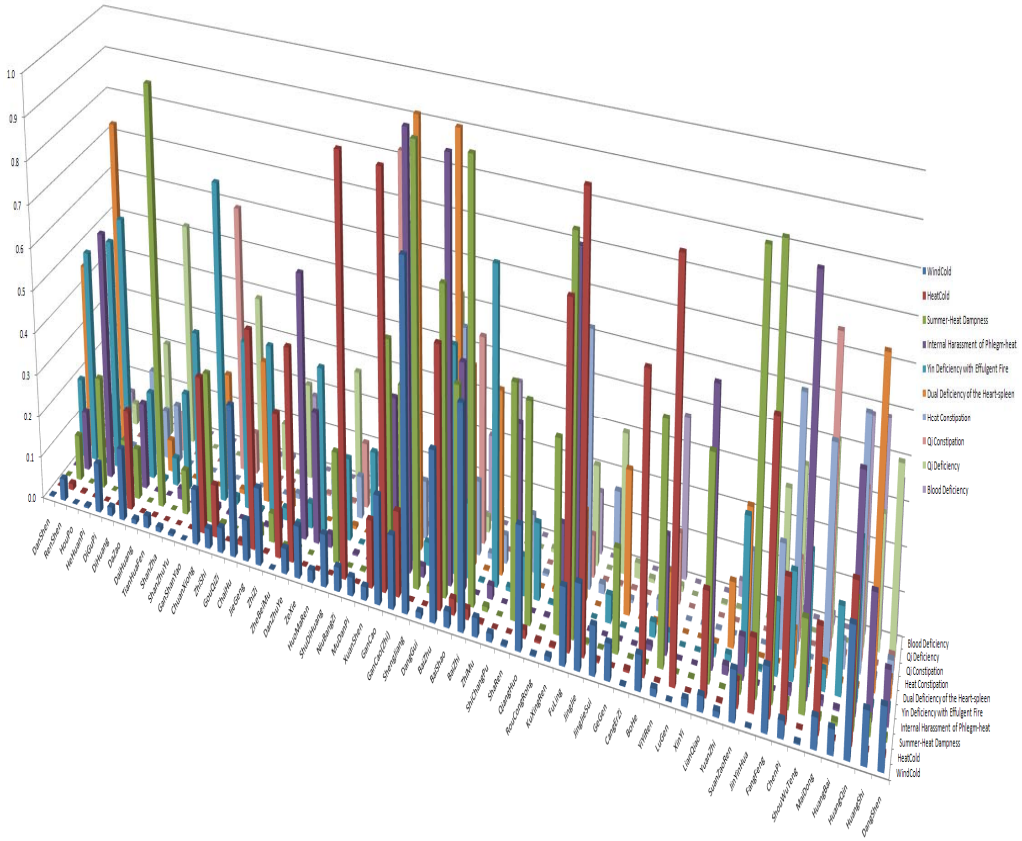


Figure 9.2.5.2.1.14a Result of herbal item suggestions – a 3 dimensional plot *(X axis - 60 pinyin herbs; Y axis - 10 illness classes (證形) in English; and Z axis - RI)*

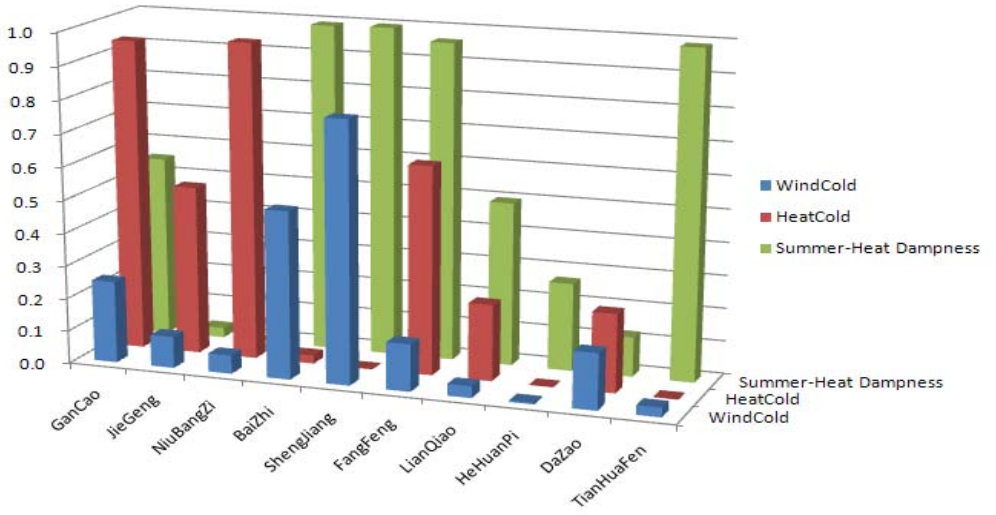


Figure 9.2.5.2.1.14b 3-dimensional plot of a Figure 9.2.5.2.1.14a subset (pinyin for herbs)

In Figure 9.2.5.2.1.14a the ten illness types of classes (證型) on the Y axis belong to three illnesses. For example, Flu (Influenza) alone has three classes (of symptoms: ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證) and ***Summer-Heat Dampness Syndrome*** (暑濕證).) The sixty herbs on the X axis treat the ten illnesses of three types to a varying degree of efficacy. Some herbs may be used to treat more than one class of symptoms (to a varying degree of effectiveness). This possibility is indicated by the vector of relative RI values versus different classes of symptoms or illnesses. This kind of of information provides the basis for Type 2 discoveries.

Figure 9.2.5.2.1.14b is the scrutiny or "explosion of a selected part" of the detailed plot in Figure 9.2.5.2.1.14a. This is the ***mini referential plot (MRP)***, which has two purposes, as follows: i) it allows the user to explode and

scrutinize a part of the ***basic referential plot (BRP)*** and ii) the explosion aids Type 2 discovery. In the latter case, a set of raw data is taken as input by the named NN modules. The referential plot for this raw set is "***mini***" in nature. By comparing the MRP and the BRP, Type 2 discovery may become apparent.

The ploting of 9.2.5.2.1.14a and 9.2.5.2.1.14b make use of the herbal names defined for the verification experiments, namely Table 9.2.5.2.1. To facilitate reader's understanding of the two 3-dimensipnal plots in Figure 9.2.5.2.1.14a and Figure 9.2.5.2.1.14b, Table 9.2.5.2.1.4 indciate the English translations for the illnesses on the Y axis.

Illness Type (Chinese)	Illness Type (English Translation)
風寒證	Wind Cold
風熱證	Wind Heat
暑濕證	Summer-Heat Dampness
痰熱內擾	Internal Harassment of Phlegm-heat
陰虛火旺	Yin Deficiency with Effulgent Fire
心脾兩虛	Dual Deficiency of the Heart-spleen
熱秘證	Heat Constipation
氣秘證	Qi Constipation
氣虛證	Qi Deficiency
血虛證	Blood Deficiency

Table 9.2.5.2.1.4 English translations for 10 illnesses types

In this 3-dimensional BRP ten herbs are selected from the original training dataset. The result is plotted against three types of the Flu (感冒) illness: ***Wind Heat Syndrome*** (風熱證), ***Wind Cold Syndrome*** (風寒證) and ***Summer-Heat Dampness Syndrome*** (暑濕證). The herbal ingredient He Huan Pi (合歡皮) or Cortex Albiziae has different relevance to 風寒證 ($RI \le 0.2$), 風熱證 ($RI \le 0.2$) and 暑濕證 ($0.2 < RI \le 0.3$). The different degrees of relevance are based on the information extracted for patients' cases collected in the past. The

herbal ingredient Tian Hua Fen (天花粉) or Radix Trichosanthis has the following RI values: 風寒證 ($RI \le 0.2$), 風熱證 ($RI \le 0.2$) and 暑濕證 ($0.8 < RI \le 1.0$). For Da Zao (大棗) or Fructus Ziziphi Jujubae the RI values are: 風寒證 ($RI \le 0.2$), 風熱證 ($0.2 < RI \le 0.4$) and 暑濕證 ($RI \le 0.2$). From the angle of ***Summer-Heat Dampness Syndrome*** (暑濕證) and from the medical records, 合歡皮, 天花粉 and 大棗 are usable herbal ingredients for treatment. For example, if 合歡皮 was not present in the set of medical records used for training the named NN and yet the new raw patient cases from the field contain it so that the NN has reasoned and revealed its relevance to 暑濕證 with a RI value, then the applicability of 合歡皮 to treating 暑濕證 is a Type 2 discovery.

Result Analysis (Tab 3 – NN Test):

In this tab, users (or CMPs) can select the symptoms of a patient; the value will then be echoed into the text field. After clicking the start button, the content in the text field will act as the input of the neural network and be processed by the learned neural network. The system would then suggest the illness name and illness type for the input symptoms (with the relevant RI), as well as a list of single herbal items that could treat he problem.

The design of this tab is aimed at fulfilling the following function: i) walk the TCM physician through the prescriptive process (in the diagnosis and prescription phase); ii) as a consultative system whereby the user can learn

anytime, anywhere; and iii) as a training system for medical personnel. The following three steps shown by Figures 9.2.5.2.1.15, 9.2.5.2.1.16 and 9.2.5.2.1.17 illustrate how Tab 3 works.

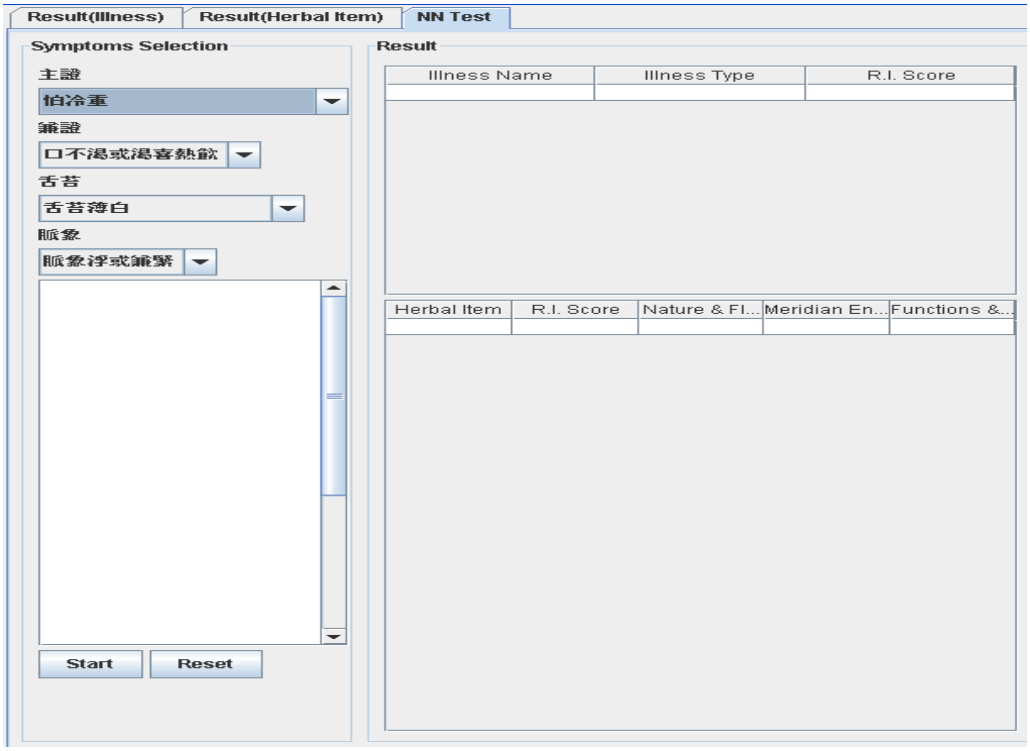


Figure 9.2.5.2.1.15 GUI invoked for user input (step 1)

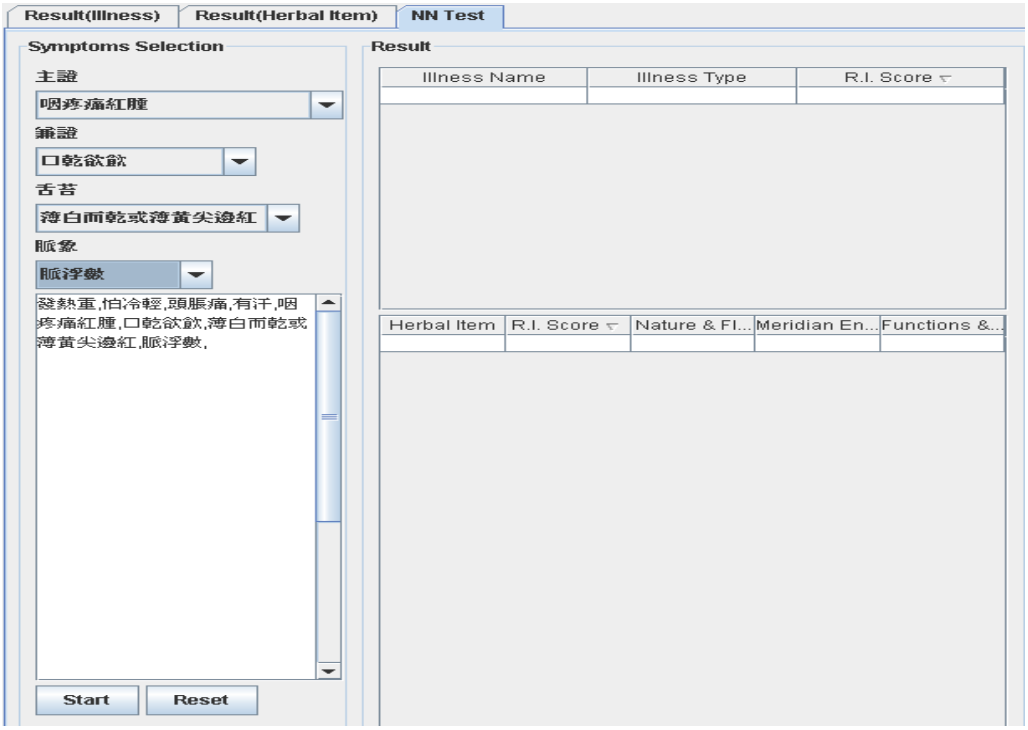


Figure 9.2.5.2.1.16 Symptoms entered by user would be echoed in text field (step 2)

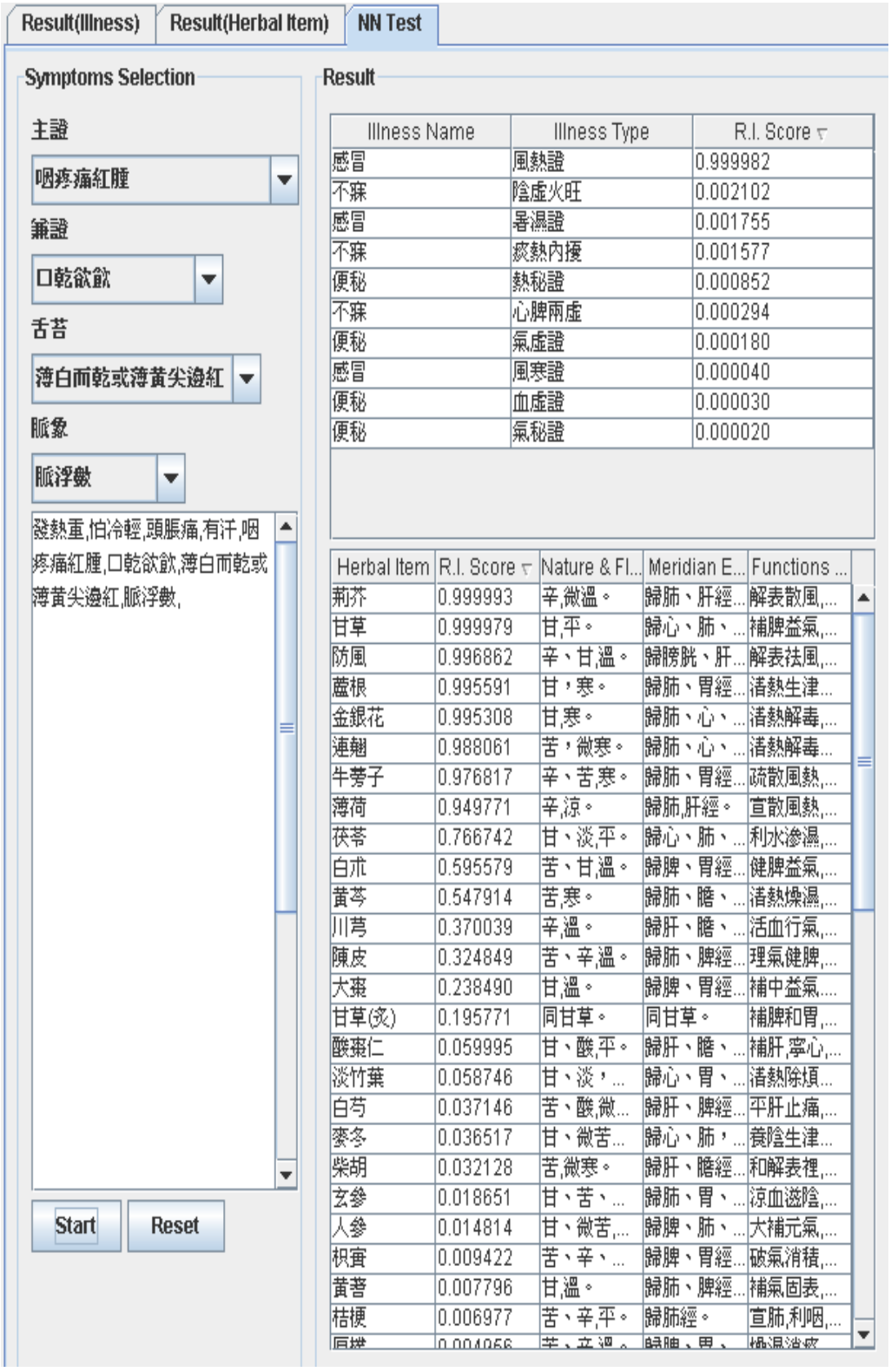


Figure 9.2.5.2.1.17 Illness types versus herbal items found by NN (step 3)

9.2.5.2.2 Timing Analysis:

It is useful to evaluate the execution times for the different named NN modules because the result would shed light on how well the NN modules would support time-critical applications. In the context of this thesis such evaluation is called timing analysis. For time-critical applications some name NN modules may have to be optimized in a real-time manner. This can be achieved by logically pruning the NN structure so that some arc-weight computations can be skipped [Lin04]. The logical pruning does not change the physical NN construct.

The time unit in the timing analysis is the CPU clock cycle because the number of clock cycle would remain the same even a faster CPU is involved. Yet, the physical time would change with respect to the physical clock speed. Table 9.2.5.2.2.1 lists some timing analyses for some named NN modules. The node that produced the results was "Pentium 4 2.8GHz"; $ExecutionTime = \frac{NumberOfExecutionClockCycle}{ClockRate}$.

Case	No. of data sets	Execution time per testing dataset	Clock Cycles (10^9) per dataset
風寒證 (Wind Cold Syndrome)	Training: 220 Testing: 30	9922ms [27.7816 x10^9 ÷ (1/2.8x10^9)]	27.7816 (x10^9 clock cycles)
風熱證 (Wind Heat Syndrome)	Training: 220 Testing: 30	9937ms	27.8236
暑濕證 (Summer-Heat Dampness Syndrome)	Training: 220 Testing: 30	9797ms	27.4316
甘草 (Gancao)	Training: 220 Testing: 30	5656ms	15.8368
桔梗 (Jiegeng)	Training: 220 Testing: 30	5718ms	16.0104
薄荷 (Bohe)	Training: 220 Testing: 30	5656ms	15.8368

Table 9.2.5.2.2.1 Some timing analysis results for demonstration

9.2.5.2.3 Real-time NN Pruning

The named NN (backpropagation) has to be trained it is used. For example, if the NN module is dedicated to the illness Influenza/Flu/Cold inference, then it is named the Flu (感冒) NN module. In fact, the Nong's enterprise TCM onto-core for clinical practice has many illness classes and each of these classes consists of a unique illnesse. Therefore, a NN module can be named after and dedicated to a class or a member of the class. For example, the Flu illness class (refer to *Figure 7.8.2 The isolated Influenza sub-ontology in XML in Chapter 7*) has three subclasses: <感冒/Cold/Flu>={ <風寒證/Wind Cold>, <風熱證/Wind Heat>, <暑濕證/Summerheat-dampness>}

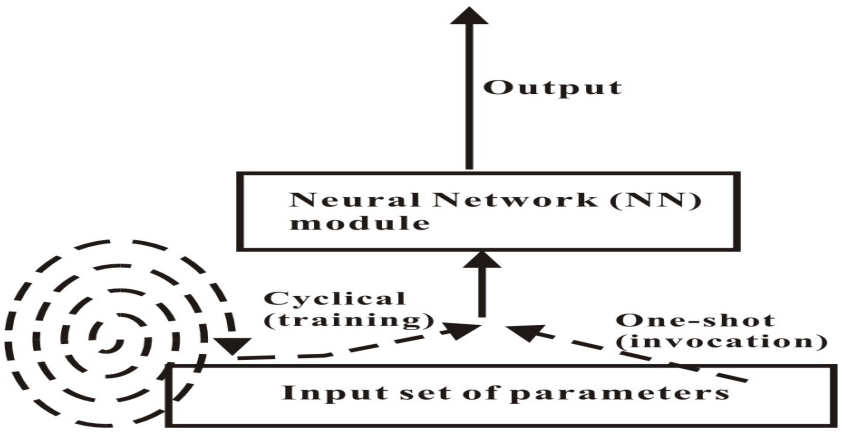


Figure 9.2.5.2.3.1 NN (backpropagation) training and application

The left side of Figure 9.2.5.2.3.1 shows that the training of a ***named NN module*** in the supervised mode but without a teacher signal is cyclical (every cycle is a training *episode*). This differs from "supervised" training with a teacher signal (or reference) rf ; if the difference between the NN output (O)

is consistently within the tolerance/error band of $\pm b$ (i.e. $|O - rf| \leq b$), then training is considered complete. In contrast, "unsupervised" NN training involves no class/name/label; this type of trasining is more suitable for clustering data when the class is not known in advance [Coppin04]. The named NN is trained with the same set of input data for the specified number of cycles/episodes. After the NN has finished training its inference accuracy depends on the RMSE (root mean square error); lower the RMSE the more accurate is the NN prediction. The one-shot application in Figure 9.2.5.2.3.1 means that the NN is invoked with a fresh data input vector to (i.e. the current case) to yield the corresponding logical output.

Real-time NN pruning is logical because the "inert" NN arcs are marked so that they will be excluded or skipped in the inference computation after the NN has learned and invoked. Conceptually, if pruning is part of the training process, then two stages, which form a pruning cycle or *iteration* will be involved: i) 1st stage - the NN should be fully trained; ii) 2nd stage – "inert" arcs in the NN construct are successively pruned. The basic pruning concept can be represented by the pseudo-code in Figure 9.2.5.2.3.2.

$TrainNN()$; /* Train the NN module before the cyclical pruning process begins */

For ($i = 1, n;\ i++$) { /* ***Cycles/iterations*** of the pruning process – loop 1*/

For ($j = 1, m;\ j++$) { /* Pruning the inert arc – loop 2*/

If $ArcWeight \leq WeightThreshold$ then $MarkIT$; /* Mark or prune it */

} /* End mark/prune arcs for exclusion in future computation – loop 2*/

Re*trainNN*(); /* Retrain NN with the original training dataset */

} /* End loop 2 */

Figure 9.2.5.2.3.2 Pseudocode - NN training with real-time pruning of inert arcs

In the pseudocode (Figure 9.2.5.2.3.2) the $ArcWeight \leq WeightThreshold$ condition defines an inert NN to be pruned. The value of the *WeightThreshold* parameter, however, can be adapted on the fly or static. For example, for my NN verification experiments the static $WeightThreshold = 0$ was applied.

The formal adherences of NN pruning are as follows:

a) Hessian matrix – It determines if there is a minimum point for $f(x)$, which can be approximated by the Taylor series. This is, it is continuous and differentiable in the defined *interval of convergence* $(x - x_0)$. If the minimum point exists that this output may be pruned depending on the nature of the function. In light of the NN, $f(x)$ is the activation function.

b) Lagrange remainder – It converge to zero for $f^{(n)}(x)$ (i.e. n^{th} order of differentiation) for a large n, and this happens for the *interval of convergence*. For RMSE = $f(x)$ this convergence is true for the un-pruned NN modules because pruning changes the very nature of $f(x)$ and shorten the *interval of convergence*. Therefore, in practice, it is

necessary to check when the RMSE may rise suddenly and exponentially

9.2.5.2.4 Experimental Results

Many experiments were conducted with similar setup to Figure 9.2.5.2.3.1, and the results unanimously indicate that the Hessian-based pruning technique [Lin04, Hagan96, Finney94, Gallant92] is indeed effective for pruning named NN modules proposed in this thesis. Some selected experimental results are presented in this section to demonstrate this point.

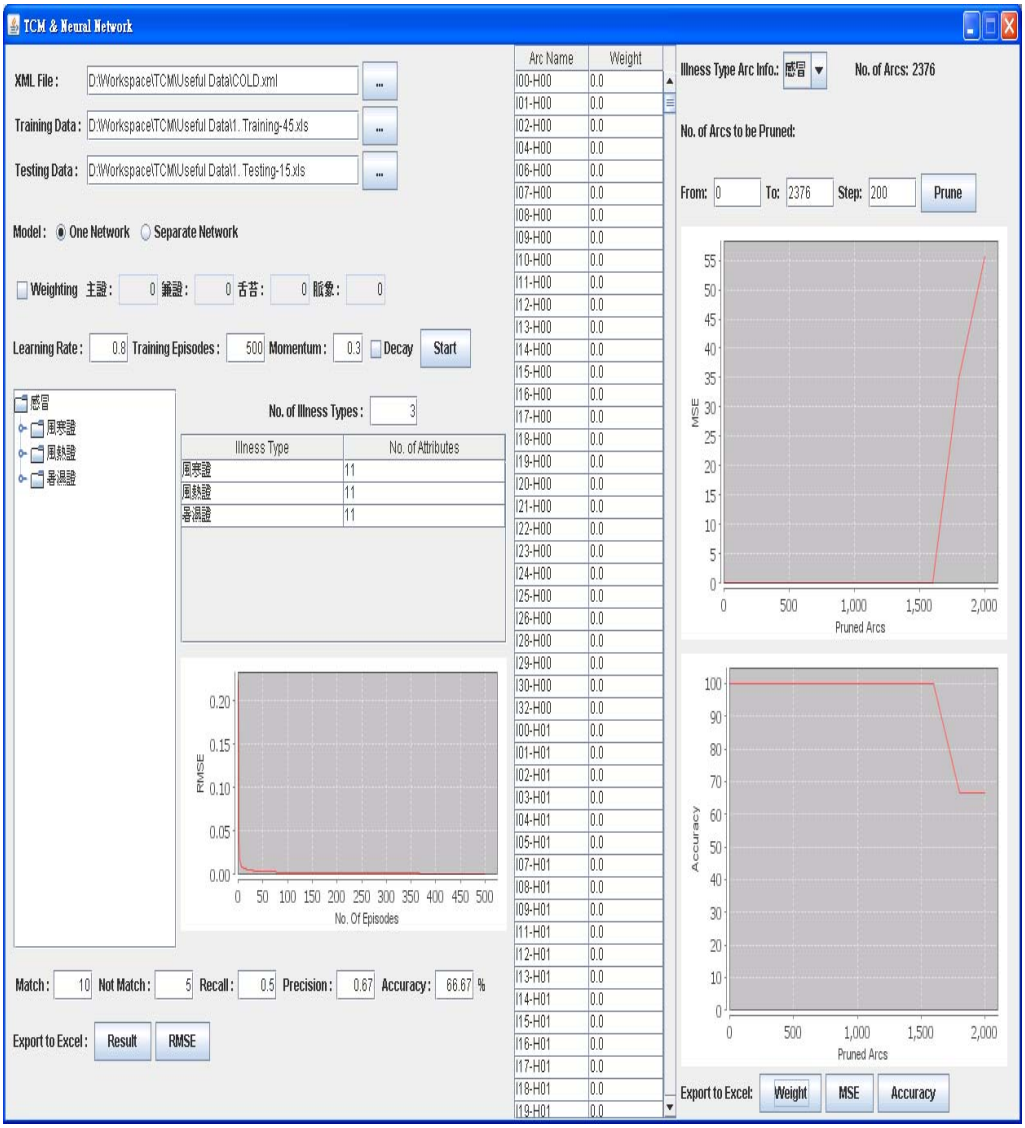


Figure 9.2.5.2.4.1a The GUI of the pruning visualizer

Figure 9.2.5.2.4.1a is the "TCM & Neural Network" system's graphical user interface (GUI) for visualizing the effect of real-time NN pruning. The left side of the GUI is for selecting a named module to be trained, and in this case it is the "Flu/感冒" NN module, which ha three subclasses that each is defined by 11 attributes. The weights of all these attributes are considered equal in this case. The "Training Data" box specifies the chosen set of real patient cases for the training process. After that the user may use a subset of the training data

specified in the “Testing Data” box to test the trained/learned NN module for inference accuracy. The center box in the GUI shows some of the untrained arcs in the Flu NN modules, 2376 arcs in total.

Figure 9.2.5.2.4.1b Explosion of the right side of the GUI in Figure 9.2.5.2.4.1a

The right box shows the pruning results. The pruning process will prune 200 inert arcs at a time (i.e. as an *iteration*) and then compute the MSE (mean square error) value after. MSE can be used because it is a powerful indicator of

error measurement. My observations in different experiments indicate that MSE would follow the same trend as the RMSE, and therefore either can be used. The result shows that if the number of pruned arcs has exceeded a certain number (in this case 1600), then the MSE begins to increase exponentially accompanied by the accuracy decreasing (a reverse manner). Figure 9.2.5.2.4.1b is the explosion of the right side of the Figure 9.2.5.2.4.1a GUI for more clarity.

The improved in computation efficiency by applying real-time NN pruning is shown in Table 9.2.5.2.4.1. In fact, the different results in our many experiments show that assigning different weights to the same set of attributes would yield higher computation efficiency. This complies with real-life clinical practice because for every illness, for example, some symptoms are more confirmative than others. For example, cough always accompanies Flu, Pneumonia and Hay Fever. Peculiar shoulder movement during inhalation is a telltale sign of Pneumonia but not for Flu or Hay Fever. Therefore, “peculiar shoulder movement” should be assigned a much higher attribute weights than others for NN inference.

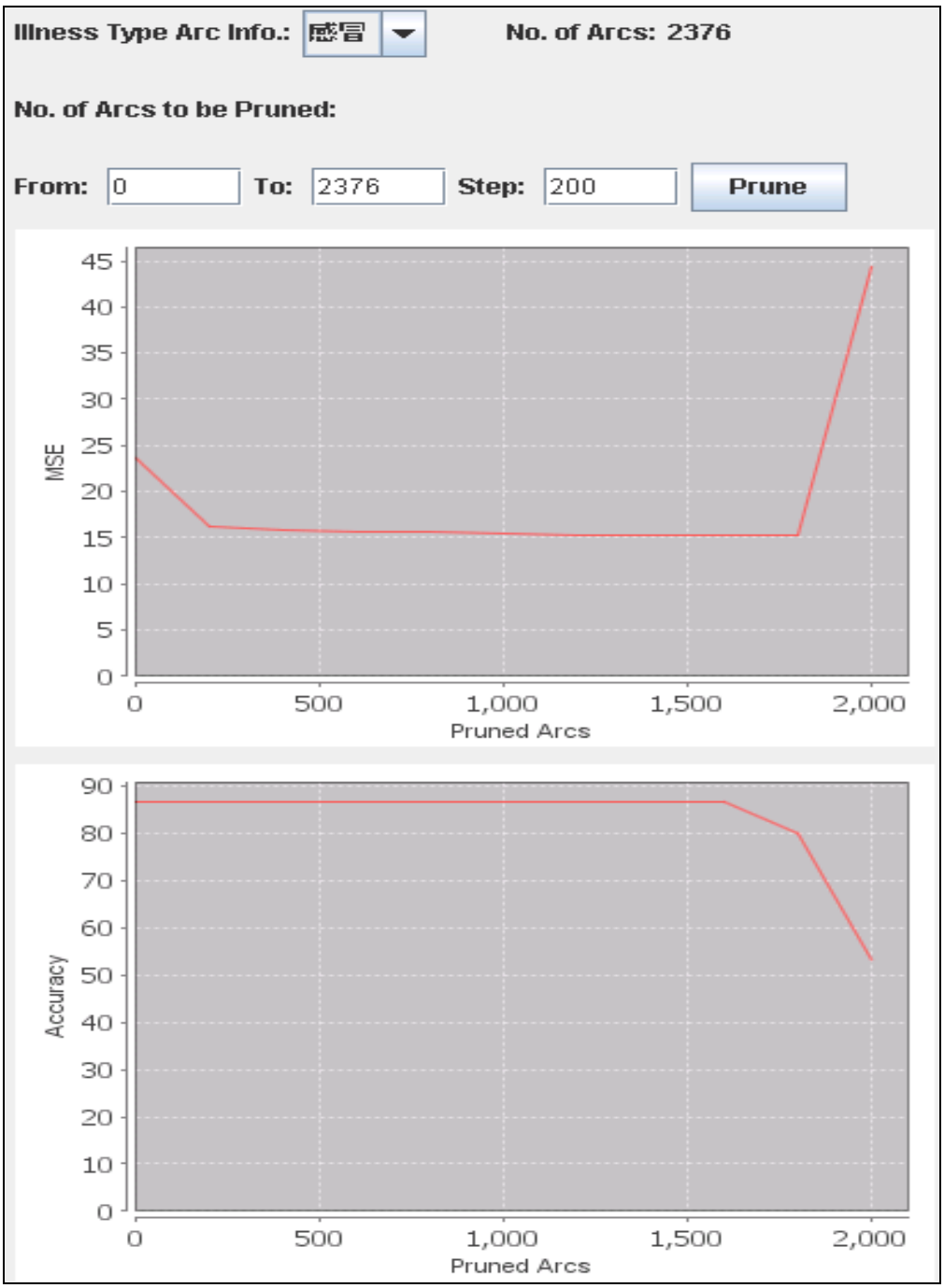


Figure 9.2.5.2.4.1c Explosion of the right side of Figure 9.2.5.2.4.1b when the attributes are assigned different weights

Figure 9.2.5.2.4.1c shows the explosion of the right side of the Figure 9.2.5.2.4.1a GUI when the attributes in the training data set are assigned different weights and then normalized. It shows the higher MSE values; the

maximum number of pruned arcs that the named NN module can tolerate without accuracy decrease. Apparently, assignments of different weights to different attributes would produce very different results from when all the attributes have the same or equal weight.

Cases	Original execution time	Number arcs pruned without affecting accuracy	Execution time after pruning	Increased efficiency (in %)
Flu/感冒 (attributes has equal weights)	18203ms	1600	17987ms	0.012%
Flu/感冒 (attributes has different weights)	16750ms	1600	15843ms	0.054%

Table 9.2.5.2.4.1 Simple comparison of two pruning results

Figure 9.2.5.2.4.2, Figure 9.2.5.2.4.3 and Figure 9.2.5.2.4.4 are the results from three different named NN modules, namely, for the herbs Gan Cao, Jie Geng, and Bo He respectively. The plots are generated based on the pruning algorithm shown in the pseudocode in Figure 9.2.5.2.3.2. They compare the execution times between the "un-pruned" and the "pruned" versions of the same named module, as well as their RMSE differences as a result. From these results we can conclude that the user should limit the number of pruned arcs so as not to increase the RMSE and decrease the NN prediction/inference accuracy. Table 9.2.5.2.4.2 compares the experimental results of three named herbal NN modules, namely, Gan Cao, Jie Geng, and Bo He).

Cases and the three named NN modules	Original execution time	Number arcs pruned without affecting accuracy	Execution time after pruning	Increased efficiency (in %)
甘草 (Gan Cao)	3047ms	1200	2984ms	2.068%
桔梗 (Jie Geng)	3140ms	1200	3000ms	4.459%
薄荷 (Bo He)	3156ms	1200	3015ms	4.468%

Table 9.2.5.2.4.2 Comparison of three herbal pruning results

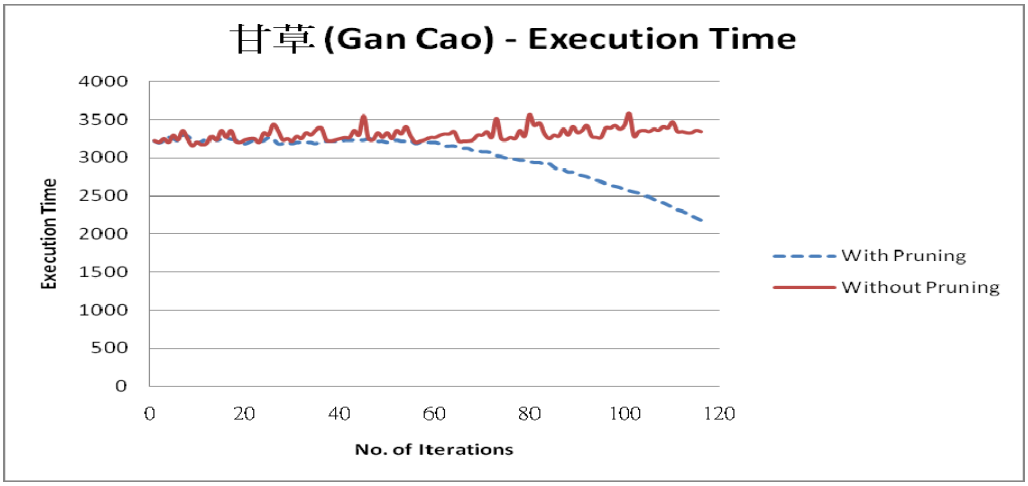


Figure 9.2.5.2.4.2a Execution times of "pruned" and "un-pruned" versions

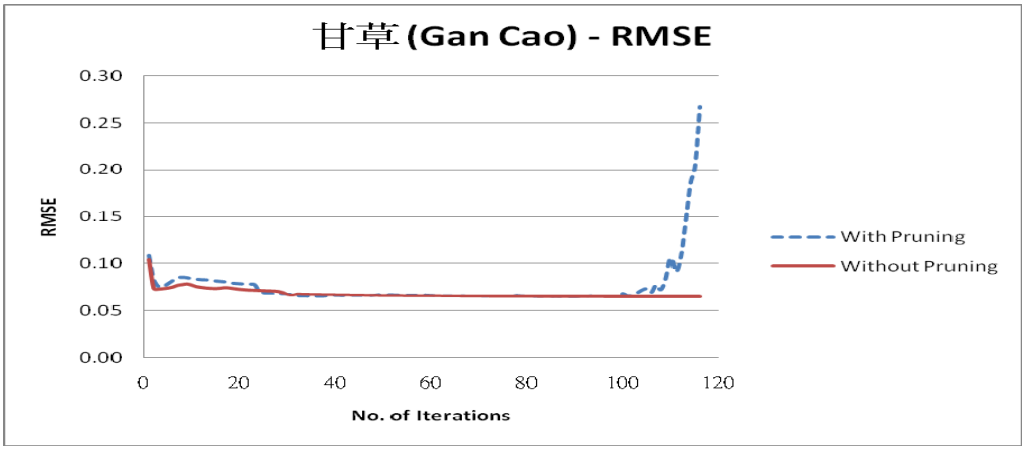


Figure 9.2.5.2.4.2b RMSE values between "pruned" and "un-pruned" versions

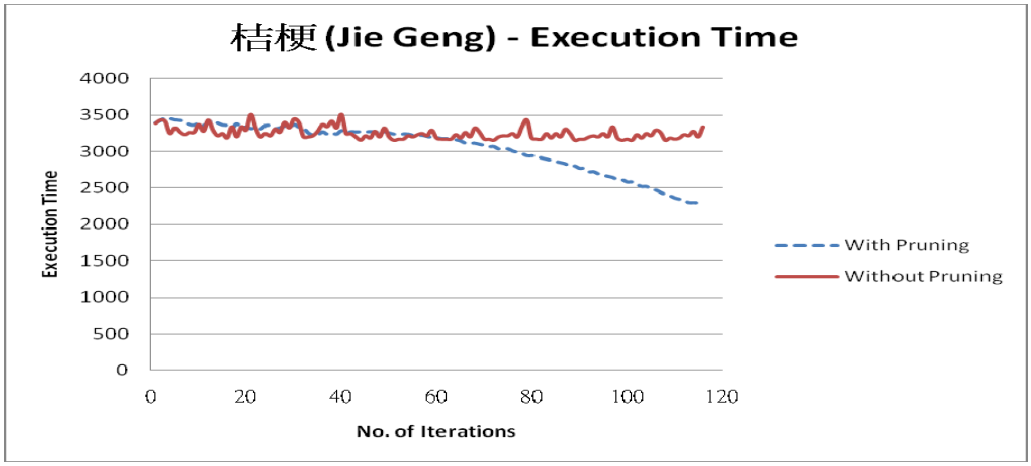


Figure 9.2.5.2.4.3a Execution times of "pruned" and "un-pruned" versions

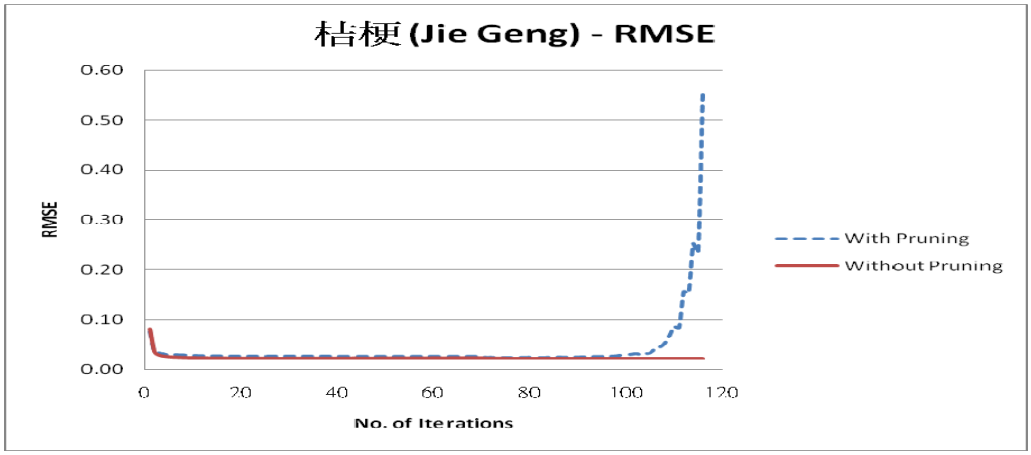


Figure 9.2.5.2.4.3b RMSE values between "pruned" and "un-pruned" versions

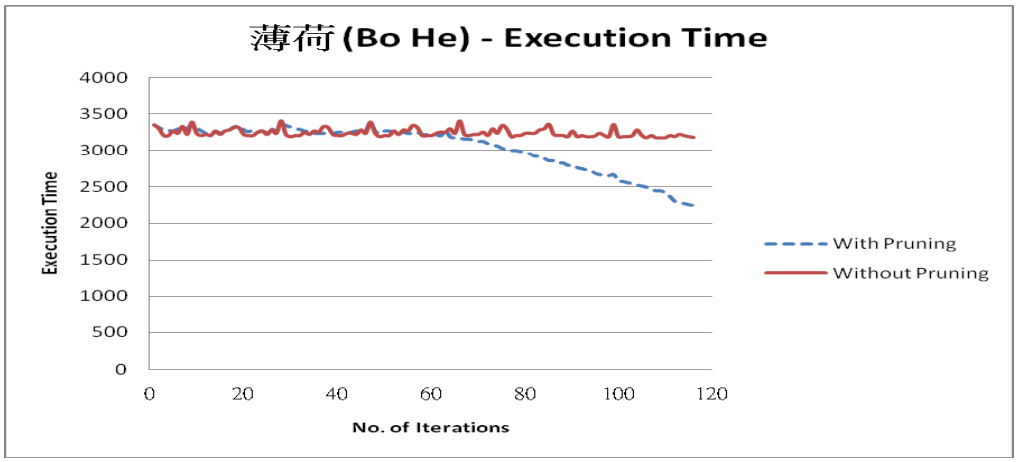


Figure 9.2.5.2.4.4a Execution times of "pruned" and "un-pruned" versions

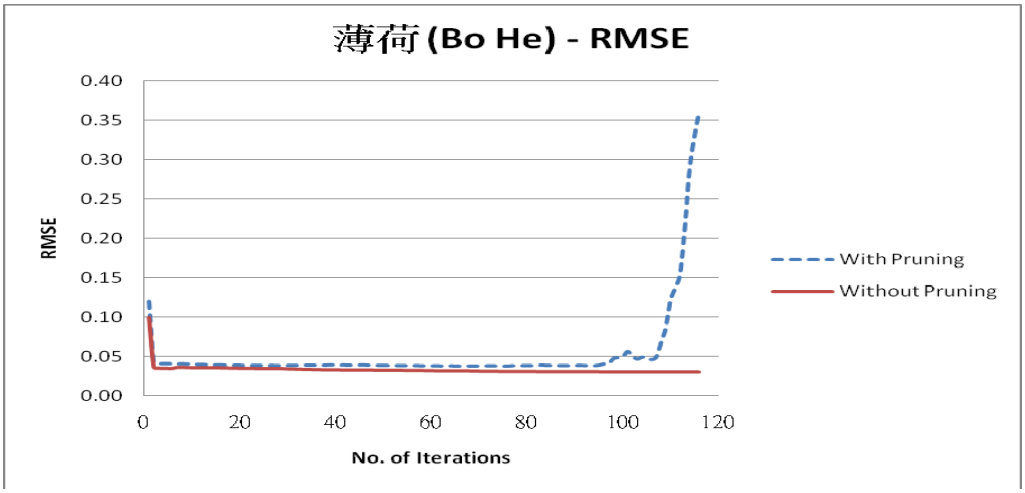


Figure 9.2.5.2.4.4b RMSE values between "pruned" and "un-pruned" versions

The number iterations or pruning cycles serves as the control to help the experimenter visualize how the pruning process is actually progressing. In order to compare different pruning results, the number of arcs to be pruned per iteration may be adjusted, for example, from 200 per iteration down to 20 per one. Adjustments as such provide more scrutiny of details on the RMSE changes, which would affect the accuracy of the finally pruned named NN module.

9.3 Recap

In this chapter experimental results at different stages in my PhD research are presented. These results together chart the flow of my progress as follows:

a) ***XML, RDF and OWL scrutiny***: The aim is to investigate if the three

cognate metadata systems proposed by the W3C can indeed be nested and used unambiguously. The experimental results show that though nesting is supported by these three systems, the XML provides the necessary flexibility for constructing the ontology as the knowledge base for this research. In order to be able to visualize and analyze the different experimental results in a real-time manner, the STV (semantic transitivity visualizer), which confirms if the cross-layer semantic transitivity exists among the ontological layers: bottom - ontology, middle - semantic net and top - user query system is proposed and implmented. Cross-layer semantic transitivity is a must for meaningful herbal discoveries. The metasystem investigation was a success and the selected results are presented in section 9.2.1.

b) ***Essential engineering support***: In this research the TCM ontology that supports all the experiments is an enterprise version of the Nong's Company Limited, a subsidiary of the Hong Kong PuraPharm Group. This version is for real-life clinical practices, especially for the YOT's mobile clinics that have been treating hundreds of patients daily in the Hong Kong SAR since its deployment three years ago. Our argument is that the experimental results using the Nong's enterprise TCM ontology would yield more trustworthy results, at least in the esyes of the TCM cimmunity. Yet, it is useless if a perfect knowledge base is worked upon by a shoddy software system. This uselessness was first noticed by Nong's when they were developing mobile-clinic system, which is now deployed by YOT with great success. For this reason Nong's proposed the novel MI (meta-interface) software engineering approach. In this

approach the user provides only a *specification of icons* (one icon for very clinical function), the the system generator would sutomatically generate the target D/P (diagnosis/prescription) system immediately from the Nong's enterprise TCM ontology. In theory, the generated system is immediately usable, for every function in this system is simply a semantic path in the TCM ontology basis. This wonderful MI proposal was never implemented or tested, but provided a powerful impetus for my PhD research. I visualized that if the MI can be implemented, then I can quickly and correctly construct my prototypes for experiments anytime and anywhere. My drive to implement the MI philosophy led to the success of the novel EOD-ISD approach. All my prototypes for experiments at different stages of my research were, in fact, automatically generated by this novel approach. Some selected experimental results for the EOD-ISD investigation are presented in the section 9.2.2.

c) ***Semantic transitivity visualization***: Any discovery from the ontology construct is meaningless if its three ontological layers are not transitive (i.e. cross-layer semantic transitivity). If such transitivity exists, then for any entity picked from a layer its corresponding representations in the other two layers would appear consistently. To aid the process that verifies that cross-layer semantic transitivity indeed exists, the STV (semantic transitivity visualizer) was proposed. The experimental results presented in section 9.2.3 show how the STV was successfully verified.

d) ***OCOE&CID approach***: This novel approach includes the following elements: i) the novel real-time automatic semantic aliasing (ASA)

mechanism based on the SAME (or in Chinese "同") principle. This principle, which is transformed from the classical TCM knowledge, is the backbnone for high-level discoveries that are usually Type 1; ii) MAT (master aliases table), which is novel proposal to enliven an operational ontology so that it can evolve with time (i.e. the concept of a living ontology); and iii) text mining as a tool to find new scientific reports and incorporate them in the MAT; such new knowledge paths the way for ontological evolution, which, however, must be concluded or sanctioned by consensus certification eventually. Some of the results from the relevant experiments are presented in the section 9.2.4.

e) ***Knowledge classification***: The experience from programming the ASA indicates: i) programming errors can be easily introduced due to the following: misunderstanding; superfluous amount of information for programmer to grasp correctly; inter-continental software engineering collaborations that can make it worse due to multi-representations and thus interpretations of a single concept; ii) it is difficult to maintain the system correctly as it migrates, for not all software parts are reusable; and iii) it is difficult to control the execution time of a software process. My thorough investigation found that the NN (Neural Network) based knowledge classification approach can resolve the above problems. In addition, it was found by [Lin04] previously that the NN execution time can be reduced by prunning in a real-time manner. After a detailed literature search and careful experiments, I found that the NN (backpropagtion) approach is suitable for Type 2 discoveries. The main reason is that the NN construct is generic and we can assign a

named/dedicate NN to deal with a TCM element (e.g. a herb or an illness). With the same set of input parameters the outputs from different named NN modules running in parallel would each compute a relevance index (RI) value quickly. For this research RI provides the basis for making sound Type 2 discovery decisions. Some of the experimental results that verified the NN (backpropagation) as a suitable knowledge classification means for this research are selected and presented in the section 9.2.5.

f) ***Discoveries of individual herbs***: They are considered as low-level or Type 2 compared to the ASA discoveries. The key elements to decide if there is a discovery are the RI values computed by a set of distributed parallel named NN modules. In fact, the use of parallel dedicated NN modules for discoveries of individual herbs is the consequence of the research result in establishing the NN (backpropagation) approach as a suitable knowledge classification technique for my PhD research. Some selected experimental results are presented in the section 9.2.5.2 for demonstration purposes.

g) ***Real-time NN pruning***: The aim is to explore if the Hessian-based pruning technique [Lin04] can be applied effectively to reduce the NN execution time when necessary. All the experimental results in this aspect indicate that the Hessian approach is indeed appropriate in light of aiding faster Type 2 discoveries. Some selected experimental results are presented in the section 9.2.5.2.3.

9.4 Conclusion and Connective Statement

All the experimental results presented in this chapter indictae that all the solutions that we have proposed for this thesis research, namely, "*Web-based Data Mining and Discovery of Useful Herbal Ingredients (WD^{2}UHI)*", in the Traditional Chinese Medicine (TCM) area, are viable. They work together and discover useful TCM herbal ingredients and prescriptions effectively and efficiently. It is therefore logical to walkthrough all these porposed solutions in the next chapter.

9.5 Key References

[Bloehdorn05] S. Bloehdorn, P. Cimiano, A. Hotho and S. Staab, An Ontology-based Framework for Text Mining, LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, May 2005, 87-112

[Coppin04] B. Coppin, Artificial Intelligence Illuminated, Jones & Bartlett Publishers, 2004

[Finney94] R.L. Finney, G.B. Thomas and M.D. Weir, Calculus, Addison-Wesley, 1994

[Gallant92] A.R. Gallant and H. White, On Learning the Derivatives of an Unknown Mapping and Its Derivatives Using Multiplayer Feedforward Networks, Neural Networks, Vol. 5, 1992

[Hagan96] M. Hagan, Neural Network Design, PWS Publishing Company, 1996

[Holzman03] L.E. Holzman, T.A. Fisher, L.M. Galitsky, A. Kontostathis and W.M. Pottenger, A Software Infrastructure for Research in Textual Data Mining, The International Journal on Artificial Intelligence Tools, Vol. 14, No. 4, 2004, 829-849

[JWong09a] J.H.K. Wong, A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), 2009

[Lin04] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, HBP: An Optimization Technique to Shorten the Control Cycle Time of the Neural Network Controller (NNC) that Provides Dynamic Buffer Tuning to Eliminate Overflow at the User Level, International Journal of Computer Systems, Science & Engineering, Vol. 19, No. 2, 2004, 75-84

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

[Yu06] S. Yu, H. Qiang and S. Jing, A Framework of XML-Based Geospatial Metadata System, Proceedings of the APWeb Workshops, 2006, 775-778

Chapter 10 Review of the Proposed Solutions, Achievements and Contributions

As discussed before, this thesis is focused on two types of discoveries conceptually: i) Type 1 – any discovery that falls ***outside*** the context of the current consensus-certified ontology; and ii) Type 2 – any discovery that falls ***within*** the current context of the current consensus-certified ontology. At the same time, we consider prescription discoveries to be "***high-level***" and individual-herb related discoveries to be "***low-level***". The acceptance criteria for new discoveries are as follows:

a) ***Trustworthiness*** - According to RFC 2828 – Internet Security Glossary, 2000, *ftp.isi.edu/in-notes/rfc2828.txt*, trustworthiness is generally defined by the chosen set of parameters depending on the application domain. In the beginning of this research we have included two elements for trustworthiness: i) communication reliability; and ii) information correctness. This led to exhaustive preliminary investigations into the areas covered by these two elements. The early argument in this research was that without reliable communication it is difficult to get the discovered information correctly and timely. I hoped that from the preliminary investigations I could estimate the amount of work involved for the PhD thesis, and if necessary what should be scaled down and excluded to meet the PhD time constraints. My view was supported by my thesis supervisor and the research fund provider, PuraPharm. ***This collective decision, as a result of my exhaustive***

investigation, was to focus mainly on information correctness as related to herbal discoveries. Once this was decided, the next question was how to create a ***formal*** TCM knowledge base so that the Type1 and Type 2 discoveries can be effectively, correctly supported. The credibility of this knowledge base should be based on the following: i) it must contain globally-accepted classical TCM information (e.g. [WHO07, Li16, Circa722BC]); and ii) it should be verifiable in real clinical environments. This led to our decision to use the Nong's enterprise TCM ontology core (onto-core), which has been used successful in world-wide clinical practices, as the knowledge base to support my experiments in the research (with the permission of Nong's, a subsidiary of PuraPharm). The extant Nong's TCM onto-core is a production version that supports the ***D/P (diagnosis/prescription) system*** installed on all the working YOT (Yan Oi Tong) mobile clinics. Since its deployment more than three years ago on the YOT mobile clinics, the D/P system have helped treat hundreds of patients daily in the Hong Kong SAR, without any serious errors reported so far. In this light, the Nong's TCM onto-core has a very high degree of trustworthiness for the clinical TCM professional. Yet, my view is that any discovery is meaningful only if cross-layer semantic transitivity (CST) exists unquestionably in the TCM onto-core. That is why it is a necessary step to be able to check this CST anytime, anywhere by anyone who has TCM knowledge.

b) ***Verifiability*** - Although the empirical clinical data from the YOT mobile clinics generally confirm the trustworthiness of the supporting TCM onto-core to a high degree, it is insufficient to conclude that the CST is perfectly correct in the TCM onto-core. In fact, any CST problems have to be detected and rectified in the next round of physical ontology consensus certification. To address the issue of detecting CST problems, the novel ***semantic transitivity visualizer*** (STV) concept was proposed, implemented, and successfully verified in real clinical environments.

c) ***Evolution*** – The Nong's TCM onto-core is the "*standard knowledge core*" for clinical operation, and is therefore not equipped to evolve automatically to keep abreast of timely changes. There is then the risk of becoming stagnated and outdated, and in order to absorb new scientific reports to enrich the discovery process, the TCM onto-core should be able to evolve. Yet, it is a contradictory requirement; on one hand the TCM onto-core should be the consensus-certified standard of clinical operation and on the other it should evolve to keep abreast of new developments. We cannot destroy the standard of clinical operation of an enterprise by simply allowing unchecked evolution. To resolve this problem a special data structure is needed to accommodate new knowledge, which is data-mined from the open sources (e.g. open web), and supports on-line evolution in a controlled manner. If the newly acquired knowledge from the open sources is inhibited, then the TCM onto-core should automatically revert back to the original consensus-

certified ontological contents. In fact, the newly acquired knowledge may be pruned and added to the extant ontology in the next round of consensus certification, which is a separate, independent process.

d) ***System error minimization*** – It is a known fact that programming errors can be introduced inadvertently in the traditional Waterfall software engineering process. It is particularly serious if the software engineering process involves teams of different cultural backgrounds across several continents (i.e. multisite software development [Wongthongtham09]). One way to resolve this problem effectively is to generate the target system automatically from the given ontology according to the given specification. Conceptually this is what the meta-interface (MI) approach, which was originally proposed by Nong's, argues. The original MI proposal lacks the necessary details for implementation but an empty "shell" of software engineering philosophy that the given specification should be iconic. Then, from the given specification of icons the "generator" should generate the target system in one step automatically. Every icon in the specification should represent a semantic path encoded in the given ontological basis. This "one-step" concept inspires my deeper pursuit of the MI implementation possibility. If this concept could be realized, it would be possible for me to produce immediately usable prototypes for my experiments anytime, anywhere in a single step.

e) ***Performance enhancement*** – The TCM ontological coverage is vast and involves clinical experience over a few thousand years. Data-mining of this coverage basically involves a very large database (VLDB) process that requires efficient tools for support. My literature search reviews that the AI approach seemingly provides an answer to the data-mining problem. The neural network (NN) by backpropagation, for example, provides a generic approach, which can be adapted and trained to suit specific problems. In this sense, the same generic NN construct is reusable. The fringe benefit from the NN approach is system scalability because several copies of the same NN construct can be integrated to form a complex NN structure or these copies can be executed in parallel to gain speed. In fact, the NN module provides additional advantages, as follows: i) it has less chance to suffer from programming errors because of its small size and thus programming simplicity; and ii) the logical definition of an operation point within a NN construct depends on the nature of the training dataset. For example, the same logical point may converge to an "AND" operation for this training dataset but an "OR" operation for another set. In contrast, algorithmic programming in the Waterfall process means two programs to deal with the AND and OR operations separately. One additional advantage of using the NN approach is possible execution time reduction and thus further performance enhancement (i.e. speedup) through the process of dynamic or real-time NN pruning [Lin04].

f) ***Sound principle*** – Discovery of trustworthy and meaningful herbal ingredients needs the support of TCM principles, which must be formal. This formality is the prelude to universal acceptance of the discoveries in the TCM domain.

10.1 Review of Solutions

To achieve the objectives set out for this research and satisfy the aforementioned acceptance criteria, the following solutions were proposed and verified:

a) ***XML as the metadata system*** – The annotations by this system are more flexible and easier to understand. The verification of this point is presented in section 9.2.1.

b) ***Essential engineering support*** - The novel EOD-ISD paradigm is proposed so that the target system/prototype can be automatically and correctly generated from the MI specification of icons and the TCM onto-core. Immediately the target system is deployable for verifications and tests. The verification of this point is presented in the section 9.2.2.

c) ***Cross-layer semantic transitivity verification*** – This is achieved by the STV (semantic transitivity visualizer) solution. Its verification is presented in the section 9.2.3

d) ***OCOE&CID paradigm*** – This solution provides a special data structure, namely the MAT (master aliases table) to support on-line ontological evolution. The MAT, however, does not interfere with operation with

the original TCM onto-core. The verification of this paradigm is in the section 9.2.4.

e) ***Knowledge classification*** – The NN (backpropagation) approach is proposed as the solution because it achieves the following at the same time: i) minimization of software errors that are inadvertently introduced in the Waterfall programming process; ii) reusability, because the same NN construct can be cloned and trained to assume various functional duties; iii) execution speed enhancement because different named/dedicated, trained NN modules can be executed in parallel; and iv) execution time reduction through application of the technique of dynamic NN pruning. The process of discovering ***individual herbal ingredients*** is based on the NN approach. The NN (backpropagation) verifications for different objectives are as follows: i) reusability is presented in the section 9.2.5; ii) discoveries of individual herbs in the section 9.2.5.2; and iii) possibility of real-time pruning to reduce NN execution time in the section 9.2.5.2.3 of Chapter 9.

f) **Automatic semantic aliasing** – This solution associates different entities with respect to the referential entity (RE). The degree of association between any entity and the RE is the unique relevance index RI), which is a value to indicate the degree of similarity. For the expression $\{(e_i, RI_i) \mid RE_j\}$ for $i, j = 1,2,..,n$, the pair (e_i, RI_i) says that RI_i is the relevance index showing the degree of similarity of the entity e_i to the referential entity RE_j.

g) **SIMILARITY/SAME (i.e. Chinese "同") principle** – This is a formal, classical TCM principle that associates different entities in light of

usage and behaviour. This provides the formal argument of how herbs and prescriptions can be linked and discovered.

10.2 Review of Achievements and Significant Contributions

The ***primary achievement*** of this research is making the proposed solutions work together, which achieves the discovery objectives of this research, namely, Type 1 and Type 2 discoveries.

The ***secondary achievement*** of this research is the publication of 16 refereed articles so far, which enrich the telemedicine body of knowledge. The

The findings from the WD^2UHI research have a significant contribution the telemedicine body of knowledge in the following aspects:

a) ***Ontology-based software engineering***: In the novel EOD-ISD paradigm the user needs to provide only the MI specification of icons. With the MI specification, the generator automatically generates the immediately usable target system. This is very useful not only for customizing prototypes for experiments in my research but also extremely useful for customizing D/P system variants that would produce huge commercial gains.

b) ***Cross-layer semantic transitivity visualization***: This technique enables anyone to check the semantic transitivity among the three ontological

layers: ontology at the bottom, semantic net in the middle and user query system at the top. The target ontology-based system is trustworthy if and only if the semantic transitivity is consistently correct.

c) ***Living ontology***: Ontology-based clinical systems are reliable but they are also stagnated with the installation of the previously consensus-certified knowledge. In order to let the ontology to live and evolve by keeping abreast of contemporary findings, newly acquired knowledge must be stored in a non-intrusive support. The proposed MAT solution, which works with text mining that ploughs the open sources (e.g. open web) for new knowledge, is such a temporary storage scheme to enliven the extant ontology.

d) ***Sound formal principle***: Medical discoveries such new TCM prescriptions and herbal ingredients need the support of formal principles that are considered acceptable by the domain experts. In this research, the SAME principle is borrowed from the TCM classical knowledge, and this principle can be adhered to by other TCM telemedicine research.

e) ***Reusable technique***: A medical ontology is usually vast and deep, and any TCM ontology, including the Nong's TCM onto-core, is no exception. In this research, the NN (backpropagation) is a generic construct that can be trained to suit specific discovery definitions. In addition, the NN approach can also provide additional benefits, as

follows: i) NN clones can be reconfigured into a complex construct; ii) NN clones can be trained to assume different discovery purposes and run in parallel; iii) execution time of a NN module can be reduced by real-time pruning; and iv) programming errors can be greatly minimized. The findings from this research have shed light on how the NN (backpropagation) classification technique can benefit telemedicine work, with regard to the above respects.

10.3 Conclusion and Connective Statement

Up to this point I have achieved all the objectives for my PhD research, namely, WD^2UHI, successfully. The proposed solutions in this research have contributed to 16 refereed publications so far. Besides, the findings from the research have undoubtedly contributed to the telemedicine body of knowledge, especially in light of TCM. In fact, the significance of the findings is reflected by the many invitations for my contributions to book chapters, conference papers, and a plenary keynote.

The next logical step is to make the overall project conclusion and suggest what should be pursued in the next step of the research in the near future.

10.4 Key References

[Circa722BC] Yellow Emperor's Canon of Internal Medicine (Huang Di Nei Jing), China, Circa 722 B.C.

[Li16] S.Z. Li, Canon on Materia Medica (Ben Cao Gang Mu), 16th Century, China

[Lin04] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, HBP: An Optimization Technique to Shorten the Control Cycle Time of the Neural Network Controller (NNC) that Provides Dynamic Buffer Tuning to Eliminate Overflow at the User Level, International Journal of Computer Systems, Science & Engineering, Vol. 19, No. 2, 2004, 75-84

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

[Wongthongtham09] P. Wongthongtham, E. Chang, T.S. Dillon and I. Sommerville, Development of a Software Engineering Ontology for Multisite Software Development, IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 8, August 2009, 1205-1217

Chapter 11 Conclusion and Suggested Future Work

The objectives for this thesis research, "***Web-based Data Mining and Discovery of Useful Herbal Ingredients (*** *WD^2UHI* ***)***" have been successfully achieved. The research is in the area of Traditional Chinese Medicine (TCM) and involves deep knowledge in both the IT and TCM domains. The aim is to discover meaningful, useful TCM herbal ingredients and prescriptions, and this involves addressing the following issues:

a) Proposal of a conceptual framework to support *trusted* and meaningful web-based data mining.
b) Proposal of methods and paradigms to achieve two types of discoveries, namely Type 1 and Type 2.
c) Proposal of principle(s) to which the discoveries can strongly adhere to and withstand the rigorous scrutiny in the subsequent process of consensus certification.

The above issues can be condensed into <u>*two*</u> main achievable objectives:

a) ***First objective – Contribute to the development of a <u>trusted</u>*** *WD^2UHI* ***platform***: This platform should support useful and meaningful web-based data mining and discovery of useful TCM herbal ingredients. It is trusted because it satisfies two very basic requirements: i) it provides reliable wireline/wireless communication to support correct, pervasive, and responsive client/server interactions, as well as those interactions

among the modules of the partitioned knowledge base (i.e. concept of a distributed data/knowledge base); and ii) it is backed up by a "standard", unambiguous, semantically-transitive knowledge base that supports local and global interoperability to a varying degree. Due to the colossal amount of research work required and the time constraints, it was decided that the ***focus*** of this thesis should be ***mainly*** on the contribution to the ***second basic requirement of the first objective***. This decision was made after serious and extensive preliminary explorations were conducted for the sake of narrowing down the research scope. It is envisioned that the proposed *WD²UHI* conceptual framework and thus the intermediary prototypes for the verification experiments would be characterized by the following:

i) ***3-layer architecture –*** The *bottom layer* is the TCM ontology; the *middle layer* has ***two main items***, namely, the ***semantic net*** which is the machine-processable form of the ontology and the ***parser*** to process it by inference; the *top layer*, which is a query system that abstracts the semantic net, for human understanding and manipulation.

ii) ***An unambiguous and axiomatic TCM ontology core (onto-core) –*** This is a local/enterprise view (or subset) carved out of the global TCM ontology, which is made up of all the available TCM classics, treatises, case histories, and new scientific findings, by the process of consensus certification.

iii) ***On-line evolution*** **–** The platform should absorb new knowledge from the open sources continuously and in a real-time manner. Yet, the absorbed knowledge is only a temporary extension of the enterprise TCM onto-core of the running system (or prototype) because permanent knowledge absorption must be vetted by the consensus certification process instituted by the system owner or enterprise.

iv) ***Ontology-based software engineering (SE)*** **–** A new paradigm in this direction needs to be proposed so that correct WD^2UHI based prototypes can be quickly, correctly and automatically generated in a single step from the named TCM onto-core given as the basis. The advantage of this new SE paradigm is that the semantic transitivity among the three system layers (i.e. bottom ontology, middle semantic net and top query system) in the prototype can be guaranteed. This semantic transitivity indicates if the ontology, and thus the discovered materials, are trustworthy. The new SE paradigm produces *customized telemedicine software systems* (CTSS) from the given TCM onto-core according to the specification provided. Though the specification contents vary with different users, the terminology/lexicon/vocabulary within these specifications should be "standardized" in respect to the given TCM onto-core. Just as a matter of convenience, this specification is called meta-interface (MI) specification or simply MI because the primordial MI SE paradigm was

first proposed by Nong's without the necessary details for implementations (i.e. simply the shell only). To generalize, the ontology-based SE paradigm generates automatically the target *web-based telemedicine system* (WTS) from the given TCM onto-core with the iconic MI specification provided. Thus, the local operational ontology in the target WTS or CTSS (synonyms), is customized from the same named TCM onto-core given, in reality. It is a variant, for different MI specifications produce variants, which are functionally diverse. To make the shell MI paradigm proposed by Nong's usable for this research, the following elements are defined and verified: *semantic transitivity, automatic software generator, master icons library, metadata impact on ontology representation, living ontology (i.e. real-time ontology evolution to support the MI mechanism), standard keyed-in ontological information versus non-standard hand-written information, and support for meaningful herbal discoveries.* These elements have been explained in detail in previous sections. The name of the MI shell concept changes in the course of my PhD research to reflect what has been modified and/or added. As a result, MI has the following cognate synonyms (please refer to the definitions of useful terms in the Preamble section): WTS, CTS, and EOD-ISD.

b) ***Second objective – Propose novel methods to discover herbal ingredients correctly and meaningfully***: In this research, herbal discoveries are divided into two types and two levels. Prescription discovery is considered high-level and individual herbal ingredients as low level. A TCM prescription is usually made up of at least four types of herbs by their capacities/roles: i) the ***principal*** that treats the illness head-on; ii) the ***courtier*** that enhances the curative power of the principle; iii) the ***assistant*** that aids the courtier by pacifying possible ill effects; and iv) the ***messenger*** that brings the curative effect to the "***cause or point***" of the illness spot-on. The relationship between a herbal ingredient and a prescription, however, is not transitive. Conceptually, any discovery is Type 1 should be outside the current ontological context or Type 2 if it is inside the current ontological context.

Logically the two types of discoveries can be defined logically as follows:

Type 1 discovery*: If set P is not part of the knowledge base K (i.e.* $P \notin K$ *) but it does possess all the attributes of one of classes in the predefined class set CL for K, then logically it is "*$< P \in CL > \wedge < P \notin K >$*", where* $\wedge$ *is the logical "AND". This means that P is a new occurrence (Type 1 discovery) with respect to the extant K. The intent of Type 1 discovery is dealing with open knowledge sources such as the open web.*

***Type 2 discovery**: For the population $\Omega = \{s_i\}$ and $i = 1,2,..., j,..,k,...,n$, the following are true logically: $CL_j = (s_j)$; $CL_k = (s_k)$; $CL_j \neq CL_k$; and $< CL_j, CL_k > \in K$. The expression $CL_j = (s_j)$ says that the class CL_j is defined by the set of attributes included in the set s, where the subscript j (i.e. s_j) marks the particular of element in s. The $P_r(CL_j CL_k) = \Theta$ expression shows that the two classes CL_j and CL_k are logically independent. For example, if the two conditions, $\{< X \in CL_j > \wedge < X \in CL_k >\}$ and $X \in K$, are logically ($\wedge$ for logical AND) satisfied by the set X, X is a discovery if either/both of the $< X \in CL_j >$ or/and $< X \in CL_k >$ association(s) was not in K previously and explicitly. This discovery is <u>not</u> Type 1 because $X \in K$ and uncovering of an association, which is intrinsically hidden in K. If the following condition is true, $P_r(UV) = P_r(U \cap V) = P_r(U) + P_r(V) - P_r(U \cup V)$ for $U = CL_j = (s_j)$ and $V = CL_k = (s_k)$, conceptually $P_r(UV)$ or $P_r(U \cap V)$ representss the relevance index (RI) that indicates the degree of similarity between U and V.*

Overall, the success of the *WD²UHI* framework relies on the trustworthiness provided by the following elements:

a) ***Communication***: This covers three modes of client/server interactions: i) interface for the user to invoke the "discovering mechanism"; ii) the text-miner and the open sources; and iii) the partitioned modules of the TCM ontology. Any communication error would reduce the credibility of the information obtained because the latter could be corrupted.

Communication reliability can be gauged as the inverse of the *average number of trials* (ANT) to get a successful transmission in a channel. If the channel has an error probability of δ, the success probability at the j^{th} trial is $P_j = \delta^{j-1}(1-\delta)$, implying $ANT = \sum_{j=1}^{k\to\infty} jP_j \approx \frac{1}{(1-\delta)}$. Therefore, the channel reliability can be gauged as $1/ANT = (1-\delta)$; that is, higher the δ means lower the channel reliability. The exploration of how to harness δ is a wide and deep area of knowledge [Wong08], but it is important for managing a distributed ontology. After careful preliminary exploration in this aspect, my experience showed that any deeper exploration in this direction would exceed the time constraints allowed for my PhD thesis. Therefore, in the subsequent research pursuit it is assumed that the TCM ontology would reside in a single node, which can be one of many in the distributed system.

b) ***TCM ontology***: The TCM ontology in all my experiments is a version of the Nong's enterprise TCM ontology core (onto-core) for clinical practice, used with permission. This TCM onto-core is trusted because it has been validated in real clinical practices by many physicians over the last three years. Its trustworthiness is, in fact, reflected by its "perfect" ontological cross-layer semantic transitivity.

c) ***Software***: This includes: i) the user interface (query system), ii) the semantic net, iii) the parsing mechanism, and vi) the knowledge engineering interface. In order to reduce the introduction of inadvertent

programming errors in the Waterfall software engineering process, the Nong's MI (meta-interface) "shell" paradigm is adopted, enriched, implemented and verified. In the enriched MI approach the user needs to provide only an iconic specification for the "system generator" to generate the target system in one shot automatically from the TCM onto-core. Every function in the target system is implemented for the specific icon in the MI specification. This icon, in fact, is one of the many semantic paths encoded in the TCM onto-core. Maximum system trustworthiness is guaranteed by matching the "perfect software" with the "standard/classical" TCM onto-core. As a result I can construct my prototypes for experiments quickly, accurately, anytime, anywhere. Another mechanism adopted in the research to reduce inadvertently introduced programming errors in the NN (backpropagation) classification technique. Once a generic NN construct is verified, it can be reused in the sense that it can be trained to assume any assigned role. As a result, no additional programming is needed.

d) ***Evolution***: The Nong's TCM onto-core is "static", for it was previously created from classical TCM information by the process of consensus certification. In this sense it does not evolve automatically with time because it needs new rounds of consensus certification as a means of evolution. To allow real-time onto-core evolution a novel mechanism is proposed, supported by the MAT (master aliases table) data structure. When the system is powered on the MAT mechanism will transform, place ad represent the entire TCM onto-core contents in the MAT data

structure as the "booted-up' contents. Then, new information items found by the text miner in the open sources (e.g. open web) will be temporarily appended to the "booted-up" information in the MAT. The appended information, however, does not change the original MAT (i.e. TCM onto-core) contents. If the appended information is disabled, then the system behaves according to the original TCM onto-core. In this way, the trustworthiness of the system ontology is always guaranteed; new appendages only act as temporary value-added information. The appendages, however, provide the new information for updating the extant TCM onto-core in the next round of consensus certification. As shown in Figure 11.1, temporary on-line onto-core evolution only happens in the MAT, and permanent onto-core evolution is based on the off-line consensus certification (CC) cycles. Theoretically, every CC cycle creates a new TCM onto-core by combining the old ontological contents and the pruned information found by text mining and appended in the MAT temporarily.

e) ***Principle(s)***: The principle(s) to decide whether a discovery is valid must come from the knowledge domain so that it is accepted and trusted across the globe. For this reason, we have carefully chosen the SAME principle of the TCM domain to identify discoveries.

f) ***Formality***: We should be able to define/represent a discovery formally. That is why we define the SAME principle logically to gauge the degree of similarity between items by their relevance index (RI).

Axiomatically, the $P_r(UV)$ or $P_r(U \cap V)$ term in the $P_r(U \cup V) = P_r(U) + P_r(V) - P_r(U \cap V)$ expression is the RI between U and V. The novel method proposed in this thesis to compute RI to aid discovery is called *automatic semantic aliasing*. Another formal approach that has been adopted to aid Type 2 discovery is the NN (backpropagation) knowledge classification technique.

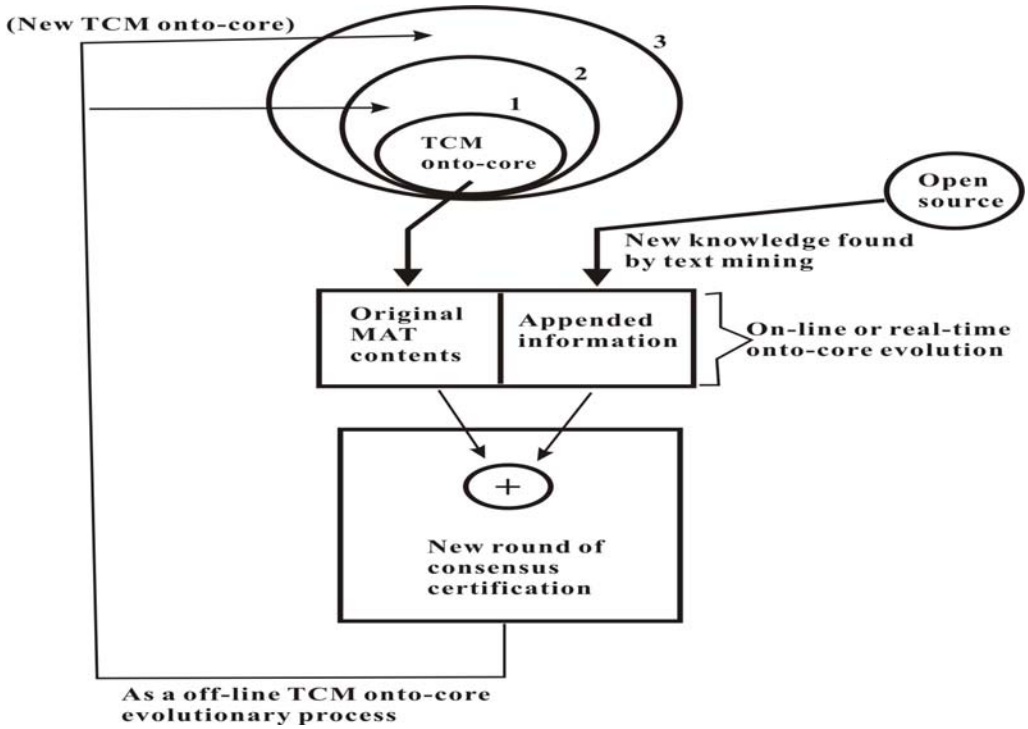


Figure 11.1 TCM onto-core evolutions

11.1 Suggested Future Work

Although this thesis research has successfully achieved the two stated objectives, there are different areas worthy deeper investigations as follows:

a) ***Communication***: For the verifications of all the proposed solution in this thesis it was assumed that the TCM ontology is not distributed. In reality, it is quite natural to parallel a very large data base (VLDB) such as a sizeable TCM ontology for fast response [Wong00]. Before this can be achieved reliably, we need to ensure that the client/server interaction is sufficiently qualitative. The core idea is to reduce the ANT value, which is defined by $ANT = \sum_{j=1}^{k\to\infty} jP_j \approx \frac{1}{(1-\delta)}$, where δ is the channel error probability. Shrinking δ is a nontrivial problem, especially from the Internet traffic point of view [Wong08].

b) ***Principles***: TCM discovery has a vast connotation, and so far we have identified only the SAME principle with the help of practicing TCM physicians. Our experience of proposing the SAME principle tells us that adopting a principle such as this is not an easy matter. The core idea is that such adoption must satisfy the requirements in the TCM domain, and meanwhile it can be axiomatically represented. This direction of pursuit is intrinsically demanding because of its multidisciplinary nature. It is, however, very rewarding because discoveries that are made via such adopted principles is definitely beneficial to the health of mankind.

11.2 Key References

[Wong00] A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, Local Compilation: A Novel Paradigm for Multilanguage-Based and Reliable Distributed Computing over the Internet, Special Issue: Mobile and Wireless Communications and Information Processing, Journal of Simulation, Vol. 75, No. 1, July 2000, 18-31

[Wong08] A.K.Y. Wong, T.S. Dillon, and W.W.K. Lin, Harnessing the Service Roundtrip Time over the Internet to Support Time-Critical Applications – Concept, Techniques and Cases (invited and contracted by Nova Science Publishers, Incorporated, New York, Feburary 2008

Bibliography

[Agrawal94] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, 12-15 September 1994

[Agrawal96] R. Agrawal and J. Shafer, Parallel Mining of Association Rules, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996, 962-969

[AI06] Enterprise Ontology, AIAI, Artificial Intelligence Application Institute, April 2006, http://www.aiai,ed.ac.uk/project/enterprise/enterprise/omtology.html

[ALA99] American Library Association, Task Force on Metadata Summary Report, June 1999

[Ausi00] Australian Bureau of Statistics, Finance, Australia 2000 Special Article – Information Technology and Telecommunications in Australia, 2000, http://www.abs.gov.au/Ausstats/abs@.nsf/0/9053E0EB512D0DDC4CA256F2A0007346F?Open

[Bardram07] J.E. Bardram and H.B. Christensen, Pervasive Computing Support for Hospitals: An Overview of the Activity-Based Computing Project, IEEE Pervasive Computing, Vol. 6, No. 1, January 2007, 44-51

[Bassler98] O.B. Bassler, Leibniz on Intension, Extension, and the Representation of Syllogistic Inference, Synthesis, Springer 1998, 117-139

[Berners98] T. Berners-Lee, Semantic Web Road Map, 1998, http://www.w3c.org/DesignIssues/Semantic.html

[Berners01] T. Berners-Lee, H. James and L. Ora, The Semantic Web, Scientific American Magazine, 17 May 2001

[Beuster02] G. Beuster, Ontologies Talk Given at Czech Academy of Sciences, 2002, http://www.uni-koblenz.de/~gb/papers/2002_intro_talk_ontology_bang/agent_ontologies.pdf

[Bishop95] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995

[Bloehdorn05] S. Bloehdorn, P. Cimiano, A. Hotho and S. Staab, An Ontology-based Framework for Text Mining, LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, May 2005, 87-112

[Boehm08] B. Boehm, Making a Difference in the Software Century, IEEE Computer Society, Vol. 41, No. 3, March 2008, 32-38

[Boyer90] R.S. Boyer and J.S. Moore, A Theorem Prover for a Computational Logic, Lecture Notes in Computer Science, Springer, 1990, 1-15

[Braden98] B. Braden, Recommendations on Queue Management and Congestion Avoidance in the Internet, RFC2309, April 1998

[Bray04] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler and F. Yergeua, Extensible Markup Language (XML) 1.0 (4th

Edition) – Origin and Goals. World Wide Web Consortium, 2004

[BRMT] http://ilrt.org/discovery/2000/08/bized-meta/index.html

[Bultermann04] D.C.A. Bultermann, Is It Time for a Moratorium on Metadata? IEEE MultiMedia, October-December 2004

[Callan03] R. Callan, Artificial Intelligence, Macmillan, May 2003

[Chatranon04] G. Chatranon, M.A. Labrador and S. Banerjee, A Survey of TCP-friendly Router-Based AQM Schemes, Computer Communications, Vol. 27, No. 15, September 2004, 1424-1440

[Cheah07] C. Cheah, Ontological Methodologies - From Open Standards Software Development to Open Standards Organizational Project Governance, Computer Science and Network Security, Vol. 7, No. 3, March 2007

[Chen96] M.S. Chen, J.S. Park, and P.S. Yu, Data Mining for Path Traversal Patterns in a Web Environment, Proceedings of the 16th International Conference on Distributed Computing Systems, Hong Kong, 27-30 May 1996, 385-392

[Circa722BC] Yellow Emperor's Canon of Internal Medicine (Huang Di Nei Jing), China, Circa 722 B.C.

[CISP99] Cross Industry Standard Process for Data Mining, 1999, http://www.crisp-dm.org/

[Clark05] D. Clark, Position Paper for Rules Concerning Project Management Ontologies, 2005

[Connor94] J.T. Connor, D. Martin and L.E. Atlas, Recurrent Neural Networks and Robust Times Series Prediction, IEEE

Transactions on Neural Networks, Vol.5, No. 2, March 1994, 240–253

[Coplien04] J. Coplien, Organizational Patterns: Beyond Technology to People, Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, 2004

[Coppin04] B. Coppin, Artificial Intelligence Illuminated, Jones & Bartlett Publishers, 2004

[Corazzon04] R. Corazzon, Descriptive and Formal Ontology – A Resource Guide to Contemporary Research, 2004, http://www.formalontology.it/

[Cottrell99] L. Cottrell, M. Zekauskas, H. Uijterwaal and T. McGregor, Comparison of Some Internet Active End-to-End Performance Measurement Projects, 1999, http://www.slac.stanford.edu/comp/net/wan-mon/iepm-cf.html

[Cranefield01] S. Cranefield, S. Haustein, and M. Purvis. UML-based Ontology Modelling for Software Agents, Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada, 28 May – 1 June 2001

[Crovella97] M.E. Crovella and A. Bestrvros, Self-similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Transaction son Networks, Vol. 5, No. 6, December 1997, 835–846

[Davenport98] T.H. Davenport and L. Prusak, Working Knowledge: How Organizations Manage What They Know, Harvard Business School Press, 1998

[Denny04] M. Denny, Ontology Tools Survey, 2004, http://ww.xml.com/lpt/a/2004/07/14/onto.html

[Dillon93] T.S. Dillon and P.L. Tan, Object Oriented Conceptual Models, Prentice Hall, 1993

[DoD03] Criscimagna, NH2003, Interoperability, Vol. 10, Reliability Analysis Center (US DoD Information Analysis Center), 1-16

[Dunn05] C. Dunn and J. Hollander, The REA Enterprise Ontology: Value System and Value Chain Modelling, Enterprise Information Systems: A Pattern-based Approach, McGraw Hill, 2005

[Dymond] http://www.dymondassoc.com/metadata/mos.htm

[Fayyad96] U.M. Fayyad, S.G. Djorgovski and N. Weir, Automating the Analysis and Cataloging of Sky Surveys, in Advances in Knowledge Discovery and Data Mining, eds. Y.M. Fayyad, AAAI/MIT Press, 1996

[Feldman95] R. Feldman and I. Dagan, KDT - Knowledge Discovery in Texts, Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD), Montreal, Quebec, Canada, 20-21 August 1995, 112–117

[Fensel03] D. Fensel, Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, 2nd Edition, Springer, Berlin/Heidelberg, 2003

[Finney94] R.L. Finney, G.B. Thomas and M.D. Weir, Calculus, Addison-Wesley, 1994

[Funahashi89] K. Funahashi, On the Approximation Realization of Continuous Mappings by Neural Networks, Neural Networks, Vol. 2, No. 3, 1989, 183-192

[Gaizauskas03] R. Gaizauskas, An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications, 2003, http://www.itri.bton.ac.uk/projects/euromap/

[Gallant92] A.R. Gallant and H. White, On Learning the Derivatives of an Unknown Mapping and Its Derivatives Using Multiplayer Feedforward Networks, Neural Networks, Vol. 5, 1992

[Ghosh03] A. Ghosh and S. Tsutsui, Advances in Evolutionary Computing: Theory and Applications, Springer, 2003

[Goolsby04] K. Goolsby and F.K. Whitlow, What Causes Outsourcing Failures? Outsourcing Journal, 2004, http://www.outsourcing-journal.com/aug2004-failure.html

[Gruber93a] T.R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, No. 2, 1993, 199-220

[Gruber93b] T.R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Proceedings of the International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation, Padova, Italy, 17 March 1993

[Guarino95] N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, 1995, 25-32

[Guarino00] N. Guarino and C. Welty, Towards a Methodology for Ontology Based Model Engineering, 2000 http://citseer/ist/psu/edu/b12206.htm

[Hagan96] M. Hagan, Neural Network Design, PWS Publishing Company, 1996

[Hamilton02] J.A. Hamilton, J. Rosen, and P.A. Summers, Developing Interoperability Metrics, Auburn University, 2002, http://www.eng.auburn.edu/users/hamiliton/security/spawar/6_Deveoping_Interoperability_Metrics/pdf

[Hearst99] M. Hearst, Untangling Text Data Mining, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), 15 October 1999

[Herman08] I. Herman, Semantic Web Activity Statement, W3C, 7 March 2008

[Hidalgo02] J.M.G. Hidalgo, Tutorial on Text Mining and Internet Content Filtering, Tutorial Notes Online, 2002, http://ecmlpkdd.cs.helsinki.fi/pdf/hidalgo.pdf

[Holzman03] L.E. Holzman, T.A. Fisher, L.M. Galitsky, A. Kontostathis and W.M. Pottenger, A Software Infrastructure for Research in Textual Data Mining, The International Journal on Artificial Intelligence Tools, Vol. 14, No. 4, 2004, 829-849

[Hornik89] K. Hornik, M. Stinchcombe and H. White, Multilayer Feedforward Networks are Universal Approximators, Neural Networks, Vol. 2, No. 5, 1989, 359-366

[Horrocks04] I. Horrocks and P.F. Patel-Schneider, Reducing OWL Entailment to Description Logic Satisfiability, Web Semantics, Vol. 1, No. 4, 2004

[Hotho05] A. Hotho, A. Nurnberger and G. Paab, A Brief Survey of Text Mining, GLDV-Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, 2005, 19-62

[IBM03] IBM and Sandpiper Software Incorporated, Ontology Definition Metamodel – Third Revised Submission to OMG/RFP ad, 4 March 2003, http://www.omg.org/docs/ad/05-8-01.pdf

[JENA] http://jena.sourceforge.net/

[JOONE] http://www.jooneworld.com/

[JWong07] J.H.K. Wong, Advanced and Research Topics in Parallel and Distributed Computing, Technical Report COMP6813, Department of Computing, April 2007

[JWong08a] J.H.K. Wong, T.S. Dillon, A.K.Y. Wong and W.W.K. Lin, Text Mining for Real-time Ontology Evolution, Data Mining for Business Applications, Springer, 2008, ISBN: 978-0-387-79419-8, 143-150

[JWong08b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon**,** An Ontology Supported Meta-Interface for the Development and Installation of Customized Web-based Telemedicine Systems, Proceedings of the 6th IFIP Workshop on Software Technologies for Future Embedded & Ubiquitous Systems (SEUS), Capri Island, Italy, 1-3 October 2008

[JWong08c] J.H.K. Wong, W.W.K. Lin and A.K.Y. Wong, Real-Time Enterprise Ontology Evolution to Aid Effective Clinical Telemedicine with Text Mining and Automatic Semantic Aliasing Support, Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, 11-13 November 2008

[JWong09a] J.H.K. Wong, A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery, International Journal of Computer Systems, Science & Engineering (CSSE), 2009

[JWong09b] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong, Enterprise-ontology-driven TCM (Traditional Chinese Medicine) Telemedicine System Generation, Proceedings of the 4th International Food – New Horizons in Chinese Medicine and Health Food Symposium, Hong Kong, 29-30 October 2009

[JWong09c] J.H.K. Wong, W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, TCM (Traditional Chinese Medicine) Telemedicine with Enterprise Ontology Support – a Form of Consensus-Certified Collective Human Intelligence, IEEE Transactions on Industrial Electronics (TIE), 2009 (to appear)

[Karr99] J.F. Kaar, International Legal Issues Confronting Telehealth Care, Telemedicine Journal, March 1999

[Karray02] F. Karray, F. Gueaieb and A. Al-Sharham, The Hierarchy Expert Tuning of PID Controllers Using Tools of Soft Computer, IEEE Transactions on System, Man and Cybernetics, Vol. 32, No. 1, 2002, 77–90

[Katifori07] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis and E. Giannopoulou, Ontology Visualization Methods – A Survey, ACM Surveys, Vol. 39, No. 4, October 2007

[Ketchen04] D. Ketchen, C. Snow and V. Hoover, Research on Competitive Dynamics: Recent Accomplishments and Future Challenges, Journal of Management, Vol. 30, No. 6, December 2004, 779-804

[Knuth97] D.E. Knuth, The Art of Computer Programming, Addison-Wesley, 1997

[Kodratoff99] Y. Kodratoff, Knowledge Discovery in Texts: A Definition and Applications, Lecture Notes in Computer Science, 1999, 1609 – 1629

[Kogut02] P. Kogut, S. Cranefield, L. Hart, M. Dutra, K. Baclawski, M. Kokar, and J. Smith, UML for Ontology Development, Knowledge Engineering Review Journal Special Issue on Ontologies in Agent Systems, Vol. 17, No. 1, March 2002, 61-64

[Kramer89] A.H. Kramer and A. Sangiovanni-Vincentelli, Efficient Parallel Learning Algorithms for Neural Networks, Advances in Neural Information Processing Systems 1, 1989, 40–48, ISBN 1-558-60015-9

[Kumar96] R. Kumar, Research methodology, A Step-by-step Guide for Beginners, Melbourne: Longman Australia, 1996

[LaoZi] http://www.iep.utm.edu/l/laozi.htm

[Lewis96] T. Lewis, The Next 10000 Years: Part 1, IEEE Computer Society, Vol. 29, No. 4, 1996, 64-70

[Li16] S.Z. Li, Canon on Materia Medica (Ben Cao Gang Mu), 16th Century, China

[Lin04] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, HBP: An Optimization Technique to Shorten the Control Cycle Time of the Neural Network Controller (NNC) that Provides Dynamic Buffer Tuning to Eliminate Overflow at the User Level, International Journal of Computer Systems, Science & Engineering, Vol. 19, No. 2, 2004, 75-84

[Lin06a] W.W.K. Lin, A.K.Y. Wong and T.S. Dillon, Application of Soft Computing Techniques to Adaptive User Buffer Overflow Control on the Internet, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 36, No. 3, May 2006, 397-410

[Lin06b] W.W.K. Lin, A.K.Y. Wong and R.S.L. Wu, Applying Fuzzy Logic and Genetic Algorithms to Enhance the Efficacy of the PID Controller in Buffer Overflow Elimination for Better Channel Response Timeliness over the Internet, Concurrency and Computation: Practice & Experience, Vol. 18, No. 7, June 2006, 725-747

[Lin08] W.W.K. Lin, J.H.K. Wong and A.K.Y. Wong, Applying Dynamic Buffer Tuning to Help Pervasive Medical

Consultation Succeed, Proc. of the 1st International Workshop on Pervasive Digital Healthcare (PerCare), Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications, Hong Kong, 17–21 March 2008, 675-679

[Lin09] W.W.K. Lin, J.H.K. Wong and A.K.Y. Wong, A Novel Real-Time Traffic Sensing (RTS) Model to Improve the Performance of Web-based Industrial Ecosystems, IEEE Transactions on Industrial Electronics (TIE), 2009

[LISP] http://en.wikipedia.org/wiki/Lisp

[Lopez99] F. Lopez, Overview of Methodologies for Building Ontologies, 1999, http://www.ontology.org/maim/presentations/madrid/analysis.pdf

[Lu00] H. Lu, L. Feng and J. Han, Beyond Intra-transaction Association Analysis: Mining Multi-dimensional Inter-Transaction Association Rule, ACM Transactions on Information Systems, Vol. 18, No. 4, October 2000, 423-454

[Manning99] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 978-02-6213-360-9

[Marrone07] P. Marrone, The Complete Guide All You Need to Know about Joone, 2007

[Mitra94] S. Mitra and S.K. Pal, Self-Organizing Neural Network as a Fuzzy Classifier, IEEE Transactions on Systems, Man and Cybernetics, Vol. 24, No. 3, 1994, 385-399

[Molnar99] S. Molnar, T.D. Dang, and A. Vid acs, Heavy Tailedness, Long-range Dependence and Self-similarity in Data Traffic, Proceedings of the 7th International Conference on Telecommunication Systems Modeling and Analysis, Nashville, Tennessee, USA, March 1999

[Nahm02] U. Nahm and R. Mooney, Text Mining with Information Extraction, Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002

[Ng08] S.C.S. Ng and A.K.Y. Wong, RCR – A Novel Model for Effective Computer-Aided TCM (Traditional Chinese Medicine) Learning over the Web, Proceedings of the International Conference on Information Technology in Education (CITE), Wuhan, China, July 2008

[NNWJ] http://www.nnwj.de/backpropagation-net.html

[Osterweil08] L.J. Osterweil, C. Ghezzi, J. Kramer and A.L. Wolf, Determining the Impact of Software Engineering Research on Practice, IEEE Computer Society, March 2008, 39-49

[Paxson95] V. Paxson and S. Floyd, Wide-area Traffic: The Failure of Poisson Modelling, IEEE/ACM Transactions on Networking, Vol. 3, No. 3, 1995, 226 – 244

[Paxson99] V. Paxson. Bro: A System for Detecting Network Intruders in Real-Time, Computer Networks, Vol. 31, No. 23-24, 1999, 2435-2463

[Pedrycz98] W. Pedrycz, Fuzzy Set Technology in Knowledge Discovery, Fuzzy Sets and Systems, Vol. 98, No. 3, 1998, 279-290

[PetriNets] http://www.petrinets.info/

[Podlipnig03] S. Podlipnig and L. Bozormenyi, A Survey of Web Cache Replacement Strategy, ACM Computing Surveys, Vol. 5, No. 54, December 2003, 374 - 398

[PTeC06] Service Description: YOT Chinese Medicine Vehicle Information System Project, PolyU Technology & Consultancy Company Limited, 2006

[PTeC07] Service Description: A Feasibility Study on the Effective Generalization of the Present PP-N's Diagnostic/Prescription (D/P) System into Mobile-Business Framework, PolyU Technology & Consultancy Company Limited, 2007

[Ren02] F. Ren, Y. Ren and X. Shan, Design of a Fuzzy Controller for Active Queue Management, Computer Communications, Vol. 25, 2002, 874–883

[RFC2828] Internet Security Glossary, 2000, ftp.isi.edu/in-notes/rfc2828.txt

[Rifaieh06] R. Rifaieh and A. Benharkat, From Ontology Phobia to Contextual Ontology Use in Enterprise Information System, Web Semantics & Ontology, ed. D. Taniar and J. Rahayu, Idea Group Incorporated, 2006

[Sarawagi08] S. Sarawagi, Information Extraction, FnT Databases, Vol. 1, No. 3, 2008

[Sarle97] W. Sarle, Neural Networks Frequently Asked Questions, 1997, ftp://ftp.sas.com/pub/neural/FAQ.html

[Singhal01] A. Singhal, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, 2001, 35–43, http://singhal.info/ieee2001.pdf

[Standish04] The 3rd Quarter Research Report: Chaos Demographics, The Standish Group International, 2004, http://standishgroup.com/sample_research/ darts_sample.php

[Taniar06] D. Taniar and J.W. Rahayu, Web Semantics & Ontology, Idea Group Publishing, 2006

[Thomas03] S.F. Thomas and M.L. Gillenson, Mobile Commerce: What It Is and What It Could Be, Communications ACM, Vol. 46, No. 12, December 2003, 33-34

[UMLS] http://umls.nlm.nih.gov/

[Uschold07] M. Uschold, M. King, S. Moralee and Y. Zorgios, The Enterprise Entology, Artificial Intelligence Applications Institute, University of Edinburg, UK, 2007, http://citesee.ist.psu.edu/cache/papers/cs/11430/ftp:zSzzSzftp.aiai.ed..ac.ukzSzpubzSzdocumentszSz1998zSz98-ker-ent-ontology.pdf/uschold95enterprise.pdf

[W3Ca] W3C, Ontology Definition MetaModel, 2005, http://www.omg.org/docs/ad/05-08-01.pdf#search='Ontology%20Definition%20Metamodel

[W3Cb] W3C, Web Service Architecture (Working Paper), http://www.w3.org/TR/ws-arch/

[W3Cc] W3C Web site, http://www.w3.org/

[W3Schools] W3 Schools, http://www.w3schools.com/

[Walt06] C. van der Walt and E. Barnard, Data Characteristics that Determine Classifier Performance, Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, 2006, 160-165

[Wand99] Y. Wand, V.C. Storey, and R. Weber, An Ontological Analysis of the Relationship Construct in Conceptual Modeling, ACM Transactions on Database Systems, Vol. 24, No. 4, 1999, 495-528

[WEKA] http://www.cs.waikato.ac.nz/ml/weka/

[WHO07] WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region, World Health Organization, 2007, ISBN 978-92-9061-248-7

[Witmer04] G. Witmer, Dictionary of Philosophy of Mind-Ontology, May 2004, http://www.artsci.wustl.edu/~philos/MindDict/ontology.html

[Witten00] I.H. Witten, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Academic Press, 2000

[Wong00] A.K.Y. Wong, W.W.K. Lin and T.S. Dillon, Local Compilation: A Novel Paradigm for Multilanguage-Based and Reliable Distributed Computing over the Internet, Special Issue: Mobile and Wireless Communications and Information Processing, Journal of Simulation, Vol. 75, No. 1, July 2000, 18-31

[Wong01] A.K.Y. Wong, T.S. Dillon, W.W.K. Lin and T.W. Ip, M^2RT: A Tool Developed for Predicting the Mean Message Response Time for Internet Channels, Journal of Computer Networks, Vol. 36, 2001, 557-577

[Wong03] A.K.Y. Wong, M.T.W. Ip and R.S.L. Wu, A Novel Dynamic Cache Size Adjustment Approach for Better Data Retrieval Performance over the Internet, Computer Communications, Vol. 26, 2003, 1709-1720

[Wong08] A.K.Y. Wong, T.S. Dillon, and W.W.K. Lin, Harnessing the Service Roundtrip Time over the Internet to Support Time-Critical Applications – Concept, Techniques and Cases (invited and contracted by Nova Science Publishers, Incorporated, New York, Feburary 2008

[Wongthongtham04] P. Wongthongtham, E. Chang and T.S. Dillon, Ontology-based Multi-agent System to Multi-site Software Development, Proceedings of the Workshop on Quantitative Techniques for Software Agile Process, Newport Beach, California, USA, November 2004

[Wongthongtham06a] P. Wongthongtham, E. Chang, T.S. Dillon and I. Sommerville, Ontology-based Multi-site Software Development Methodology and Tools, Journal of Systems Architecture, Vol. 52, No. 11, 2006, 640-653

[Wongthongtham06b] P. Wongthongtham, Ontology and Multi-agent-based Systems for Human Disease Studies, PhD Thesis, Curtin University, 2006

[Wongthongtham09] P. Wongthongtham, E. Chang, T.S. Dillon and I. Sommerville, Development of a Software Engineering Ontology for Multisite Software Development, IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 8, August 2009, 1205-1217

[Wu02] R.S.L. Wu, A.K.Y. Wong and T.S. Dillon, Comparing Four Novel Scalable Split/Aggregate Algorithms (Mobile Agent Based) for Distributes Mining of Multimedia Association Rules over the Internet, Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, Vol. 2, 24-27 June 2002, 760-766.

[Wu06a] R.S.L. Wu, A.K.Y. Wong and T.S. Dillon, A Novel Dynamic Cache Size Tuning Model with Relative Object Popularity for Fast Web Information Retrieval, Journal of Supercomputing, 2006

[Wu06b] R.S.L. Wu, W.W.K. Lin and A.K.Y. Wong, Harnessing Wireless Traffic is an Effective Way to Improve Mobile Internet Performance, Proceedings of the 1st Australian Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless), Sydney, Australia, March 2006

[Yann98] L. Yann, B. Leon, G.B. Orr and K. Muller, Efficient BackProp, Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, Springer, 1998

[Yergeua04] F. Yergeua, T. Bra, J. Paoli, S. Sperberg-McQueen and E. Maler, Extensible Markup Language (XML) 1.0 (3rd ed.), W3C Recommendation, 2004

[Yeung98] D.S. Yeung and A.K.Y. Wong, The OORHS: A Conceptual Framework that Provides Easy and Reversible Distributed Programming, International Journal of Computer Systems, Science and Engineering, Vol. 13, No. 15, 1998, 289 - 301

[Yu87] P.S. Yu, C.M. Krishna and Y.H. Lee, An Adaptive Optimization Model with Applications to Testing, Computer Performance and Reliability, 1987, 503-515

[Yu06] S. Yu, H. Qiang and S. Jing, A Framework of XML-Based Geospatial Metadata System, Proceedings of the APWeb Workshops, 2006, 775-778

[Zhao03] W. Zhao, R. Chellappa, P.J. Phillips and A. Rosenfeld, Face Recognition: A Literature Survey, ACM Computing Surveys, Vol. 35, No. 4, 2003, 339–458

Appendix I Formal Agreement of Thesis Submission

PuraPharm培力

PuraPharm International (H.K.) Ltd.
培力(香港)健康產品有限公司

1st December 2009

It is our congratulations to Mr. Jackei Ho Kei WONG for completing his Ph.D. We agree with his contents and the accomplishment of the stated objectives. His Ph.D. research is a TCS (Teaching Company Scheme) based, contractual agreement between the Hong Kong Polytechnic University and the PuraPharm Group of the Hong Kong SAR. The objectives defined and fulfilled in the present thesis are decided upon the preliminary investigations carried out by Jackei. We agreed to discontinue deeper investigation into the "trusted communication" aspects in the research and postpone it for future investigation. The reason is that the expected research effort for this part alone would far exceed the time constraints imposed on Jackei's Ph.D. pursuit. The collective decision was reached by the company and the Chief Supervisor, after assessing all the available facts. "Trust communication" is, nevertheless, useful for boosting the company's mobile business gain. Therefore, it should be pursed in the future work, as a continuation of the direction pinpointed by Jackei's thesis, in any appropriate form(s) to be defined. The intellectual property of the research results in Jackei's thesis is stipulated by the TCS agreement. That is, any of the results in any form should be affirmed by consents by the TCS partners.

Dr. Wilfred Wan Kei LIN, Ph.D., MACM, MIEEE
Industrial co-supervisor
IT Manager of PuraPharm
(on behalf of PuraPharm)

Suites 4103-08, Jardine House, 1 Connaught Place, Central, Hong Kong.
G.P.O. Box 10122, Hong Kong
Telephone: (852) 2840-1840 • Facsimile: (852) 2840-0778
E-mail: info@purapharm.com
Website: www.purapharm.com

香港中環康樂廣場1號怡和大廈4103-08室
香港郵政總局信箱10122
電話：(852) 2840-1840 • 傳真：(852) 2840-0778
電子郵件：info@purapharm.com
網址：www.purapharm.com

Appendix I Formal Agreement of Thesis Submission

PuraPharm培力

PuraPharm International (H.K.) Ltd.
培力(香港)鈺康產品有限公司

1st December 2009

It is our congratulations to Mr. Jackei Ho Kei WONG for completing his Ph.D. We agree with his contents and the accomplishment of the stated objectives. His Ph.D. research is a TCS (Teaching Company Scheme) based, contractual agreement between the Hong Kong Polytechnic University and the PuraPharm Group of the Hong Kong SAR. The objectives defined and fulfilled in the present thesis are decided upon the preliminary investigations carried out by Jackei. We agreed to discontinue deeper investigation into the "trusted communication" aspects in the research and postpone it for future investigation. The reason is that the expected research effort for this part alone would far exceed the time constraints imposed on Jackei's Ph.D. pursuit. The collective decision was reached by the company and the Chief Supervisor, after assessing all the available facts. "Trust communication" is, nevertheless, useful for boosting the company's mobile business gain. Therefore, it should be pursed in the future work, as a continuation of the direction pinpointed by Jackei's thesis, in any appropriate form(s) to be defined. The intellectual property of the research results in Jackei's thesis is stipulated by the TCS agreement. That is, any of the results in any form should be affirmed by consents by the TCS partners.

Dr. Wilfred Wan Kei LIN, Ph.D., MACM, MIEEE
Industrial co-supervisor
IT Manager of PuraPharm
(on behalf of PuraPharm)

Suites 4103-08, Jardine House, 1 Connaught Place, Central, Hong Kong
G.P.O. Box 10122, Hong Kong
Telephone : (852) 2840-1840 • Facsimile : (852) 2840-0778
E-mail : info@purapharm.com
Website : www.purapharm.com

香港中環康樂廣場1號怡和大廈4103-08室
香港郵政信箱10122
電話：(852) 2840-1840 • 傳真：(852) 2840-0778
電子郵件：info@purapharm.com
網址：www.purapharm.com

Appendix II Cross-Validation, Error Estimation and Sixty Different Herbal Items for the Experiments

Cross-validation:

Cross-validation is the statistical practice of partitioning a sample data into N subsets. One subset is selected to validate NN, which was trained by the rest (N-1) sets [Kohavi95, Chang92]. Cross-validation is a model evaluation method that is better than residual approach. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen [Schneider97]. One way to overcome this problem is not to use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed will be used to test the performance of the learned model on "new" data.

There are four common types of cross-validation:

a) *Holdout Validation (HOV):* In this type, the data are never crossed over. Observations are chosen randomly from the initial sample to form the validation data, and the remaining observations are retained as the training data. In most of the cases, only less than a third of the initial sample is used for validation data [DTree].

b) Leave-one-out Cross-validation: Leave-one-out cross-validation (LOOCV) uses only one single observation from the original sample as the validation data, and the remaining observations as the training data. The validation is repeated until every observation in the sample has been used once as the validation data. This is the same as a K-fold cross-validation (to be discussed later), with K equal to the number of observations in the original sample [Cawley03]. The advantage of LOOCV is that all the data can be repeatedly used for both training and testing. This helps maintain statistical robustness by eliminating the "dominant effect" of a single outlier that can perturb the data set [Liang05]. The disadvantage of LOOCV is the high computational cost entailed by the *n* trainings of the classifier [Riegera08].

c) Repeated Random Sub-sampling Validation (RRSSV): The dataset is randomly split into training and validation data. For each such split, the classifier is retrained with the training data and then validated by the remaining data. The results from each split can then be averaged. The advantage over K-fold cross-validation is that the proportion of the training/validation split is not dependent on the number of iterations (fold). However, some observations may never be selected in the validation sub-samples whereas others may be selected more than once (i.e. validation subsets may overlap).

d) K-fold Cross-validation (KFCV): The original sample is partitioned into K sub-samples [Kohavi95]. Among the K sub-samples, a single one is

retained as the validation data for the model testing, and the remaining K-1 sub-samples are used as training data. The cross-validation process is then repeated K times (according to the fold assigned), with each of the K sub-samples used exactly once as the validation data. The K results from the folds can be averaged (or combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. The value of K (the fold) is commonly set to 10 (10-fold). The disadvantage is that the fold can be quite unrepresentative, biasing the estimator [Diamantidis00] as a result.

The four cross-validation types are compared in Table app.1.

	Advantage	**Disadvantage**
HOV	Easy to implement	The data is not crossed at all
LOOCV	All the data can be repeatedly used for both training and testing	The high computational cost entailed by the *n* trainings of the classifier
RRSSV	The proportion of the training/validation split is not dependent on the number of iterations (folds)	Some observations may never be selected in the validation subsample, whereas others may be selected more than once
KFCV	All observations are used for both training and validation, and each observation is used for validation exactly once	(not obvious for our purpose)

Table app.1 Comparison of four cross-validation approaches

K-fold Cross-validation as the Choice:

K-fold cross validation was chosen, for it is generally more accurate than the hold-out validation [Blum99]. For this reason it is often used for

generalizing error estimation, model selection, and learning algorithms comparison [Bengio03].

Error Estimation:

a) Error Estimation – It accommodates three common types of errors that can be used as NN performance indicators:

i) *Mean Absolute Error (MAE)*: It measures the error between the forecast and the expected outcome [Schaeffer80]; $MAE = \frac{1}{n}\sum_{i=1}^{n}| f_i - y_i | = \frac{1}{n}\sum_{i=1}^{n}| e_i |$. It is an average of the absolute errors where f_i is the prediction and y_i the true value.

ii) *Root Mean Squared Error (RMSE)*: This measures the difference between a prediction (or reference) and a measurement [Sheiner81]. If O is the conceptual predicted output, m_i the measured output of a process at the i^{th} cycle, then $RMSE = \sqrt{\sum_{i=1}^{N}(O - m_i)^2}$ for $i \le N$.

iii) *Relative Absolute Error (RAE)*: The absolute error is the magnitude of the difference between the exact value and the approximation. The RAE is the absolute error divided by the

magnitude of the exact value. For the given value v and its approximation, v_{approx} $\eta = \frac{|v_{approx} - v|}{v}$ holds.

b) Accuracy and Precision: In statistics, accuracy is the degree of closeness of a measured/calculated quantity to its actual (true) value. Precision is the degree to which further measurements/calculations show the same or similar results [Taylor97]. In this section, different measurement methods for accuracy and precision will be discussed:

i) *Precision/Recall*: Precision is the proportion of the examples, which truly have class x, among all those that were classified as class x. It is the probability that a (randomly selected) retrieved document is relevant [Yates99]. Recall is the same as TP Rate – the probability that a (randomly selected) relevant document is retrieved in a search [Yates99].

ii) *Kappa Statistic*: This measures the agreement of prediction with the true class. The complete agreement is represented by the value 1.0, and the equation is: $k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$. Pr(a) is the relative observed agreement among the "raters", and Pr(e) is the hypothetical probability of chance agreement.

iii) *TP/FP Rate*: TP Rate (True Positive Rate) is the proportion of examples, which were classified as class x, among all examples

which truly have class x (i.e. how much of the class was captured). FP Rate (False Positive Rate) is the proportion of examples, which were classified as class *x*, but belong to a different class, among all examples which are not of class *x*.

iv) *Precision/Recall*: Precision is the proportion of the examples, which truly have class *x* among all those that were classified as class *x*. It is the probability that a (randomly selected) retrieved document is relevant [Yates99]. Recall is the same as TP Rate – the probability that a (randomly selected) relevant document is retrieved in a search [Yates99].

v) *F-Measure*: This measures the effectiveness of retrieval with respect to a user, who attaches the same importance to recall as precision; the equation for this measure is: $\frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecsision + \mathrm{Re}call}$.

(Remarks: lt - Latin; and py - pinyin in Putonghua)

感冒	不寐	便秘
甘草 (lt – Radix Glycyrrhizae) (py - Gan Cao)	知母 (lt – Rhizoma Anemarrhenae) (py - Zhi Mu)	大黃 (lt – Radix et Rhizoma Rhei) (py - Dai Huang)
桔梗 (lt – Radix Platycodi) (py - Jie Geng)	白朮 (lt – Rhizoma Atractylodis) (py - Bai Zhu)	當歸 (lt- Radix Angelicae) (py - Dang Gui)
薄荷 (lt –Herba Menthae) (py - Bo He)	甘草(炙) (lt - Radix Glycyrrhizae (roasted)) (py - Gan Cao (Zhi))	陳皮 (lt – Oericarpium Citri Reticulatae) (py - Chen Pi)
白芷 (lt – Radix Angelicae Dahuricae (py - Bai Zhi)	酸棗仁 (lt – Seman Ziziphi Spinosae) (py - Suan Zao Ren)	黨參 (lt – Radix Codonopsis Pilosulae) (py - Dang Shen)
生薑 (lt – Rhizoma Zingiberis Recens) (py - Sheng Jiang)	合歡皮 (lt – Cortex Albizziae) (py - He Huan Pi)	火麻仁 (lt – Fructus Cannabis) (py - Huo Ma Ren)
防風 (lt – Radix Ledebouriellae) (py - Fang Feng)	川芎 (lt – Rhizoma Ligustici) (py - Chuan Xiong)	麥冬 (lt – Radix Ophiopogonis) (py - Mai Dong)
連翹 (lt – Fructus Forsythiae) (py - Lian Qiao)	白芍 (lt – Radix Paeoniae Alba) (py - Bai Shao)	黃蓍 (lt – Radix Astragali seu Hedysari) (py - Huang Shi/Qi)
牛蒡子 (lt – Fructus Arctii) (py - Niu Bang Zi)	茯苓 (lt – Poria) (py - Fu Ling)	玄參 (lt – Radix Scrophulariae) (py - Xuan Shen)
荆芥 (lt – Herba Schizonepetae) (py - Jing Jie)	地骨皮 (lt – Cortex Lycii Radicis) (py - Di Gu Pi)	黃芩 (lt – Radix Scutellariae) (py - Huang Qin)
葛根 (lt – Radix Puerariae) (py - Ge Gen)	大棗 (lt – Fructus Ziziphi Jujubae) (py - Da Zao)	枳實 (lt – Fructus Aurantii Immaturus) (py - Zhi Shi)
天花粉 (lt – Radix Trichosanthis) (py - Tian Hua Fen)	(牡)丹皮 (lt – Cortex Moutan) (py – (Mu) Dan Pi)	梔子 (lt – Fructus Gardeniae) (py - Zhi Zi)
柴胡 (lt – Radix Bupleuri) (py - Chai Hu)	首烏藤 (lt – Caulis Polygoni Multiflori) (py - Shou Wu Teng)	地黃 (lt – Radix Rehmanniae); raw (py - Di Huang)
金銀花 (lt – Flos Loncierae) (py - Jin Yin Hua)	熟地黃 (lt – Radix Rehmanniae); cooked (py - Shu Di Huang)	澤瀉 (lt – Rhizoma Alismatis) (py - Ze Xie)
辛夷 (lt – Flos Magnoliae) (py - Xin Yi)	遠志 (lt – Radix Polygalae) (py - Yuan Zhi)	厚樸 (lt – Cortex Magnoliae Officinalis) (py - Hou Po)
荆芥穗 (lt – Herba Schizonepetae) (py - Jing Jie Sui)	枸杞子 (lt – Fructus Lycii) (py - Gou Qi Zi)	肉蓯蓉 (lt – Herba Cistanches) (py - Rou Cong Rong)
羌活 (lt – Rhizoma seu Radix Notopterygii) (py - Qiang Huo)	丹參 (lt – Radix Salviae Miltiorrhizae) (py - Dan Shen)	山茱萸 (lt – Fructus Corni) (py - Shan Zhu Yu)
蘆根 (lt – Rhizoma Phragmitis) (py - Lu Gen)	黃柏 (lt - Cortex Phellodendri) (py - Huang Bo)	(幹)山藥 (lt – Rhizoma Dioscoreae) (py – (Gan) Shan Yao)
浙貝母 (lt – Bulbus Fritillariae Thunbergii) (py - Zhe Bei Mu)	人參 (lt – Radix Ginseng) (py - Ren Shen)	薏苡仁 (lt – Cemen Coicis) (py - Yi Yi Ren)
淡竹葉 (lt – Herba Lophatheri) (py - Dan Zhu Ye)	山楂(lt – Fructus Crataegi) (py - Shan Zha)	砂仁 (lt – Fructus Amomi) (py - Sha Ren)
蒼耳子 (lt – Fructus Xanthii) (py - Cang Er Zi)	石菖蒲 (lt – Rhizoma Acori Graninei) (py - Shi Chang Pu)	苦杏仁 (lt – Semen Armeniacae Amarum) (py - Ku Xing Ren)

Table app.2 Sixty different herbal items for the experiments

Key References

[Bengio03] Y. Bengio and Y. Grandvalet, Estimators of Variance of K-fold Cross-validation, CRM Workshop on Advances in Machine Learning, University of Montreal, 2003

[Blum99] A Blum, A Kalai and J Langford, Beating the Hold-out: Bounds for K-fold and Progressive Cross-validation, Proceedings of the 12th Annual Conference on Computational Learning Theory, 1999, 203–208

[Cawley03] G.C. Cawley and N.L.C. Talbot, Efficient Leave-one-out Cross Validation of Kernel Fisher Discriminant Classifiers, Pattern Recognition, Vol 36, 2003, 2585–2592

[Chang92] J. Chang, Y. Luo and K. Su, GPSM: a Generalized Probabilistic Semantic Model for ambiguity resolution. Proceedings of the 30th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, 28 June-02 July 1992, 177-184

[Diamantidis00] N.A. Diamantidis, D. Karlis and E.A. Giakoumakis, Unsupervised Stratification of Cross-validation for Accuracy Estimation, Artificial Intelligence, Vol 116, No. 1-2, January 2000, 1-16

[DTree] http://decisiontrees.net/node/36

[Kohavi95] R. Kohavi, A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, Proceedings of the

14th International Joint Conference on Artificial Intelligence, Vol. 2, No. 12, 1995, 1137–1143

[Liang05] Y. Liang and A. Kelemen, Associating Phenotypes with Molecular Events: Recent Statistical Advances and Challenges Underpinning Microarray Experiments, Functional and Integrative Genomics, Vol. 6, No. 1, 2005, 1-13

[Riegera08] J.W. Riegera, C. Reicherta, K.R. Gegenfurtnerb, T. Noesselta, C. Braunc, H.J. Heinzea, R. Krusee and H. Hinrichsa, Predicting the Recognition of Natural Scenes from Single Trial MEG Recordings of Brain Activity, NeuroImage, Vol. 42, No. 3, September 2008, 1056-1068

[Schaeffer80] D.L. Schaeffer, A Model Evaluation Methodology Applicable to Environmental Assessment Models, Ecological Modelling, Vol. 8, 1980, 275-295

[Schneider97] J. Schneider and A.W. Moore, A Locally Weighted Learning Tutorial using Vizier 1.0, Tutorial Notes, The Robotics Institute, School of Computer Science, Carnegie Mellon University, 1 February 1997

[Sheiner81] L.B. Sheiner and S.L. Beal, Some Suggestions for Measuring Predictive Performance, Journal of Pharmacokinetics and Pharmacodynamics, 1981, 503-512

[Taylor97] J.R. Taylor, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, University Science Books, 1997, 128-129

[Yates99] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. New York: ACM Press, Addison-Wesley, 1999

Appendix III The isolated Influenza sub-ontology in XML

<?xml version="1.0" encoding="Big5" ?>

<感冒> (<Cold>)

<風寒證> (<Wind Cold Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defense-exterior >)

<怕冷> (<Fear of Cold>)

<怕冷重>id="1"</怕冷重>

(<Seriously Fear of Cold>id=”1”</Seriously Fear of Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<發熱輕>id="1"</發熱輕>

(<Mild Fever>id=”1”</Mild Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<無汗>id="1"</無汗>

(<Absence of Sweating>id=”1”</Absence of Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭痛四肢痠痛>id="1"</頭痛四肢痠痛>

(<Headache and Limbs Pain>id=”1”</Headache and Limbs Pain>)

</頭身> (</Head and Body>)

</衛表> (</Defence-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻塞流清涕多嚏>id="1"</鼻塞流清涕多嚏>

(<Nasal Congestion, Clear Sniffle and Profuse Sneezing>id=”1”

</Nasal Congestion, Clear Sniffle and Profuse Sneezing>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽癢>id="1"</咽癢> (<Throat Itching>id=”1” </Throat Itching>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽聲重>id="1"</咳嗽聲重>

(<Profuse Coughing>id=”1”</Profuse Coughing>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰稀薄色白>id="1"</痰稀薄色白>

(<White Clear Phlegm>id=”1”</White Clear Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<口不渴或渴喜熱飲>id="1"</口不渴或渴喜熱飲>

(<Not Thirsty Nor Fancy Hot Drinks>id=”1”</Not Thirsty Nor Like Hot Drinks>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌苔薄白>id="1"</舌苔薄白> (<Thin White Fur>id=”1”</Thin White Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈象浮或兼緊>id="1"</脈象浮或兼緊>

(<Floating Tight Pulse>id=”1”</Floating Tight Pulse>)

</脈象> (/<Pulse>)

</風寒證> (</Wind Cold Syndrome >)

<風熱證> (<Wind Heat Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defence-exterior >)

<怕冷> (<Fear of Cold>)

<怕冷輕>id="2"</怕冷輕>

(<Slightly Fear of Cold>id=”2”</Slightly Fear of Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<發熱重>id="2"</發熱重> (<Profuse Fever>id=”2”</Profuse Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<有汗>id="2"</有汗> (<Sweating>id=”2”</Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭脹痛>id="2"</頭脹痛>

(<Distending Headache>id=”2” </Distending Headache>)

</頭身> (</Head and Body>)

</衛表> (</Defence-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻塞流黃濁涕>id="2"</鼻塞流黃濁涕>

(<Nasal Congestion and Yellow Turbid Sniffle>id=”2”</Nasal Congestion and Yellow Turbid Sniffle>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽疼痛紅腫>id="2"</咽疼痛紅腫>

(<Sore Throat and Swelling>id=”2” </Sore Throat and Swelling>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽啞氣粗>id="2"</咳嗽啞氣粗>

(<Cough with Hoarseness Sound>id=”2”</Cough with Hoarseness Sound>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰稠黏色黃或白>id="2"</痰稠黏色黃或白>

(<Yellow or White Turbid Phlegm>id=”2”</Yellow or White Turbid Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<口乾欲飲>id="2"</口乾欲飲>

(<Thirsty and like to Drink>id=”2”</Thirsty and like to Drink>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<薄白而乾或薄黃尖邊紅>id="2"</薄白而乾或薄黃尖邊紅>

(<Dry Thin White Fur or Red Tips and Margins of Tongue with Thin Yellow Fur>id=”2” </Dry Thin White Fur or Red Tips and Margins of Tongue with Thin Yellow Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈浮數>id="2"</脈浮數>

(<Floating and Rapid Pulse>id=”2”</Floating and Rapid Pulse>)

</脈象> (</Pulse>)

</風熱證> (</Wind Heat Syndrome>)

<暑濕證> (<Summerheat-dampness Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defence-exterior >)

<怕冷> (<Fear of Cold>)

<微惡寒>id="3"</微惡寒>

(<Mild Aversion to Cold>id=”3”</Mild Aversion to Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<身熱>id="3"</身熱> (<Body Fever>id=”3”</Body Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<少汗>id="3"</少汗> (<Mild Sweating>id=”3”</Mild Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭昏重脹痛肢體痠楚>id="3"</頭昏重脹痛肢體痠楚>

(<Heavy-headedness, Headache and Limbs Pain>id=”3”

</Heavy-headedness, Headache and Limbs Pain>)

</頭身> (</Head and Body>)

</衛表> (</Defence-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻流濁涕>id="3"</鼻流濁涕>

(<Turbid Sniffle>id=”3”</Turbid Sniffle>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽痛>id="3"</咽痛> (<Sore Throat>id=”3”</Sore Throat>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽>id="3"</咳嗽> (<Cough>id=”3”</Cough>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰黏或黃或白>id="3"</痰黏或黃或白>

(<Yellow or White Turbid Phlegm>id=”3”</Yellow or White Turbid Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<心煩口渴>id="3"</心煩口渴>

(<Vexation and Thirsty>id=”3”</Vexation and Thirsty>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<苔淡黃膩或黃膩>id="3"</苔淡黃膩或黃膩>

(<Pale Yellow or Yellow Slimy Fur>id=”3”</Pale Yellow or Yellow Slimy Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈濡數>id="3"</脈濡數>

(<Rapid and Soggy Pulse>id=”3”</Rapid and Soggy Pulse>)

</脈象> (</Pulse>)

</暑濕證> (</Summerheat-dampness Syndrome>)

</感冒> (</Cold>)

Appendix IV Partial RDF-annotated code for partial DOM tree in Chinese

```
<?xml version="1.0" encoding="big5" ?          /* Chinese only */>
<!DOCTYPE rdf:RDF [<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">]>

<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
        xmlns:tcm="c:/testSchema#"
        xml:base="file:///./#">

        <tcm:病名 rdf:ID="感冒" />

        <tcm:證型 rdf:ID="風寒證">
                <rdfs:subClassOf rdf:resource="#感冒"/>
                <tcm:兼證 rdf:ID="風寒證兼證">口不渴或渴喜熱飲</tcm:兼證>
                <tcm:舌苔 rdf:ID="風寒證舌苔">舌苔薄白</tcm:舌苔>
                <tcm:脈象 rdf:ID="風寒證脈象">脈象浮或兼緊</tcm:脈象>
        </tcm:證型>

        <tcm:主證 rdf:ID="風寒證主證">
                <rdfs:subClassOf rdf:resource="#風寒證"/>
        </tcm:主證>

        <tcm:衛表 rdf:ID="風寒證衛表">
                <rdfs:subClassOf rdf:resource="#風寒證主證"/>
```

<tcm:怕冷 rdf:ID="風寒證怕冷">怕冷重</tcm:怕冷>

<tcm:發熱 rdf:ID="風寒證發熱">發熱輕</tcm:發熱>

<tcm:汗 rdf:ID="風寒證汗">無汗</tcm:汗>

<tcm:頭身 rdf:ID="風寒證頭身">頭痛四肢痠痛</tcm:頭身>

</tcm:衛表>

<tcm:肺 rdf:ID="風寒證肺">

<rdfs:subClassOf rdf:resource="#風寒證主證"/>

<tcm:鼻 rdf:ID="風寒證鼻">鼻塞流清涕多嚏</tcm:鼻>

<tcm:咽 rdf:ID="風寒證咽">咽癢</tcm:咽>

<tcm:咳 rdf:ID="風寒證咳">咳嗽聲重</tcm:咳>

<tcm:痰 rdf:ID="風寒證痰">痰稀薄色白</tcm:痰>

</tcm:肺>

<tcm:證型 rdf:ID="風熱證">

<rdfs:subClassOf rdf:resource="#感冒"/>

<tcm:兼證 rdf:ID="風熱證兼證">口乾欲飲</tcm:兼證>

<tcm:舌苔 rdf:ID="風熱證舌苔">薄白而乾或薄黃尖邊紅</tcm:舌苔>

<tcm:脈象 rdf:ID="風熱證脈象">脈浮數</tcm:脈象>

</tcm:證型>

<tcm:主證 rdf:ID="風熱證主證">

<rdfs:subClassOf rdf:resource="#風熱證"/>

</tcm:主證>

<tcm:衛表 rdf:ID="風熱證衛表">

<rdfs:subClassOf rdf:resource="#風熱證主證"/>

<tcm:怕冷 rdf:ID="風熱證怕冷">怕冷輕</tcm:怕冷>

<tcm:發熱 rdf:ID="風熱證發熱">發熱重</tcm:發熱>

<tcm:汗 rdf:ID="風熱證汗">有汗</tcm:汗>

<tcm:頭身 rdf:ID="風熱證頭身">頭脹痛</tcm:頭身>

</tcm:衛表>

<tcm:肺 rdf:ID="風熱證肺">

<rdfs:subClassOf rdf:resource="#風熱證主證"/>

<tcm:鼻 rdf:ID="風熱證鼻">鼻塞流黃濁涕</tcm:鼻>

<tcm:咽 rdf:ID="風熱證咽">咽疼痛紅腫</tcm:咽>

<tcm:咳 rdf:ID="風熱證咳">咳嗽啞氣粗</tcm:咳>

<tcm:痰 rdf:ID="風熱證痰">痰稠黏色黃或白</tcm:痰>

</tcm:肺>

<tcm:證型 rdf:ID="暑濕證">

<rdfs:subClassOf rdf:resource="#感冒"/>

<tcm:兼證 rdf:ID="暑濕證兼證">心煩口渴</tcm:兼證>

<tcm:舌苔 rdf:ID="暑濕證舌苔">苔淡黃膩或黃膩</tcm:舌苔>

<tcm:脈象 rdf:ID="暑濕證脈象">脈濡數</tcm:脈象>

</tcm:證型>

<tcm:主證 rdf:ID="暑濕證主證">

<rdfs:subClassOf rdf:resource="#暑濕證"/>

</tcm:主證>

<tcm:衛表 rdf:ID="暑濕證衛表">

<rdfs:subClassOf rdf:resource="#暑濕證主證"/>

<tcm:怕冷 rdf:ID="暑濕證怕冷">微惡寒</tcm:怕冷>

<tcm:發熱 rdf:ID="暑濕證發熱">身熱</tcm:發熱>

<tcm:汗 rdf:ID="暑濕證汗">少汗</tcm:汗>

<tcm:頭身 rdf:ID="暑濕證頭身">頭昏重脹痛肢體痠楚</tcm:頭身>

</tcm:衛表>

<tcm:肺 rdf:ID="暑濕證肺">

<rdfs:subClassOf rdf:resource="#暑濕證主證"/>

<tcm:鼻 rdf:ID="暑濕證鼻">鼻流濁涕</tcm:鼻>

<tcm:咽 rdf:ID="暑濕證咽">咽痛</tcm:咽>

<tcm:咳 rdf:ID="暑濕證咳">咳嗽</tcm:咳>

<tcm:痰 rdf:ID="暑濕證痰">痰黏或黃或白</tcm:痰>

</tcm:肺>

</rdf:RDF>

Appendix V Partial RDF schema in Chinese

```
<?xml version="1.0" encoding="big5" ?>
<!DOCTYPE rdf:RDF [
        <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
]>

<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:tcm="c:/testSchema#">

        <rdfs:Class rdf:ID="病名" />

        <rdfs:Class rdf:ID="證型">
                <rdfs:subClassOf rdf:resource="#病名"/>
        </rdfs:Class>

        <rdfs:Class rdf:ID="主證">
                <rdfs:subClassOf rdf:resource="#證型"/>
        </rdfs:Class>

        <rdfs:Class rdf:ID="衛表">
                <rdfs:subClassOf rdf:resource="#主證"/>
        </rdfs:Class>

        <rdfs:Class rdf:ID="肺">
                <rdfs:subClassOf rdf:resource="#主證"/>
        </rdfs:Class>
```

<rdf:Description rdf:ID="兼證">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>兼證</rdfs:label>

</rdf:Description>

<rdf:Description rdf:ID="舌苔">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>舌苔</rdfs:label>

</rdf:Description>

<rdf:Description rdf:ID="脈象">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>脈象</rdfs:label>

</rdf:Description>

<rdf:Description rdf:ID="怕冷">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>怕冷</rdfs:label>

</rdf:Description>

<rdf:Description rdf:ID="發熱">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>發熱</rdfs:label>

```
		</rdf:Description>

		<rdf:Description rdf:ID="汗">
			<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Property"/>
			<rdfs:label>汗</rdfs:label>
		</rdf:Description>

		<rdf:Description rdf:ID="頭身">
			<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Property"/>
			<rdfs:label>頭身</rdfs:label>
		</rdf:Description>

		<rdf:Description rdf:ID="鼻">
			<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Property"/>
			<rdfs:label>鼻</rdfs:label>
		</rdf:Description>

		<rdf:Description rdf:ID="咽">
			<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Property"/>
			<rdfs:label>咽</rdfs:label>
		</rdf:Description>

		<rdf:Description rdf:ID="咳">
			<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Property"/>
```

<rdfs:label>咳</rdfs:label>

</rdf:Description>

<rdf:Description rdf:ID="痰">

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>

<rdfs:label>痰</rdfs:label>

</rdf:Description>

</rdf:RDF>

Appendix VI Partial OWL-annotated code for partial DOM tree in Chinese

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE rdf:RDF [
        <!ENTITY eg   'urn:tcm:eg/'>
        <!ENTITY rdf  'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
        <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
        <!ENTITY xsd  'http://www.w3.org/2001/XMLSchema#'>
        <!ENTITY owl  "http://www.w3.org/2002/07/owl#" >
]>
<rdf:RDF
        xmlns:rdf="&rdf;"
        xmlns:rdfs="&rdfs;"
        xmlns:xsd="&xsd;"
        xmlns:owl="&owl;"
        xml:base="urn:tcm/#"
        xmlns="&eg;"
        xmlns:tcm="urn:tcm/#"
>
        <owl:Class rdf:ID="病名">
        </owl:Class>
        <owl:Class rdf:ID="證型">
                <rdfs:subClassOf rdf:resource="病名"/>
        </owl:Class>
        <owl:Class rdf:ID="主證">
                <rdfs:subClassOf rdf:resource="證型"/>
        </owl:Class>
        <owl:Class rdf:ID="衛表">
```

<rdfs:subClassOf rdf:resource="主證"/>

</owl:Class>

<owl:Class rdf:ID="肺">

<rdfs:subClassOf rdf:resource="主證"/>

</owl:Class>

<owl:ObjectProperty rdf:ID="兼證">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="舌苔">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="脈象">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="怕冷">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="發熱">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="汗">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="頭身">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="鼻">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="咽">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="咳">

</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="痰">

</owl:ObjectProperty>

<owl:Class rdf:ID="感冒">

<rdf:type rdf:resource="病名"/>

</owl:Class>

<owl:Class rdf:ID="風寒證">

<rdf:type rdf:resource="證型"/>

<rdfs:subClassOf rdf:resource="感冒"/>

<兼證 rdf:ID="風寒證兼證">口不渴或渴喜熱飲</兼證>

<舌苔 rdf:ID="風寒證舌苔">舌苔薄白</舌苔>

<脈象 rdf:ID="風寒證脈象">脈象浮或兼緊</脈象>

<建議處方 rdf:resource="荊防敗毒散" />

</owl:Class >

<owl:Class rdf:ID="風寒證主證">

<rdf:type rdf:resource="主證"/>

<rdfs:subClassOf rdf:resource="風寒證"/>

</owl:Class>

<owl:Class rdf:ID="風寒證衛表">

<rdf:type rdf:resource="衛表"/>

<rdfs:subClassOf rdf:resource="風寒證主證"/>

<怕冷 rdf:ID="風寒證怕冷">怕冷重</怕冷>

<發熱 rdf:ID="風寒證發熱">發熱輕</發熱>

<汗 rdf:ID="風寒證汗">無汗</汗>

<頭身 rdf:ID="風寒證頭身">頭痛四肢痠痛</頭身>

</owl:Class>

<owl:Class rdf:ID="風寒證肺">

<rdf:type rdf:resource="肺"/>

<rdfs:subClassOf rdf:resource="風寒證主證"/>

<鼻 rdf:ID="風寒證鼻">鼻塞流清涕多噓</鼻>

<咽 rdf:ID="風寒證咽">咽癢</咽>

<咳 rdf:ID="風寒證咳">咳嗽聲重</咳>

<痰 rdf:ID="風寒證痰">痰稀薄色白</痰>

</owl:Class>

<owl:Class rdf:ID="風熱證">

<rdf:type rdf:resource="證型"/>

<rdfs:subClassOf rdf:resource="感冒"/>

<兼證 rdf:ID="風熱證兼證">口乾欲飲</兼證>

<舌苔 rdf:ID="風熱證舌苔">薄白而乾或薄黃尖邊紅</舌苔>

<脈象 rdf:ID="風熱證脈象">脈浮數</脈象>

<建議處方 rdf:resource="銀翹散" />

</owl:Class>

<owl:Class rdf:ID="風熱證主證">

<rdf:type rdf:resource="主證"/>

<rdfs:subClassOf rdf:resource="風熱證"/>

</owl:Class>

<owl:Class rdf:ID="風熱證衛表">

<rdf:type rdf:resource="衛表"/>

<rdfs:subClassOf rdf:resource="風熱證主證"/>

<怕冷 rdf:ID="風熱證怕冷">怕冷輕</怕冷>

<發熱 rdf:ID="風熱證發熱">發熱重</發熱>

<汗 rdf:ID="風熱證汗">有汗</汗>

<頭身 rdf:ID="風熱證頭身">頭脹痛</頭身>

</owl:Class>

<owl:Class rdf:ID="風熱證肺">

<rdf:type rdf:resource="#肺"/>

<rdfs:subClassOf rdf:resource="#風熱證主證"/>

<鼻 rdf:ID="風熱證鼻">鼻塞流黃濁涕</鼻>

<咽 rdf:ID="風熱證咽">咽疼痛紅腫</咽>

<咳 rdf:ID="風熱證咳">咳嗽啞氣粗</咳>

<痰 rdf:ID="風熱證痰">痰稠黏色黃或白</痰>

</owl:Class>

<owl:Class rdf:ID="暑濕證">

<rdf:type rdf:resource="證型"/>

<rdfs:subClassOf rdf:resource="感冒"/>

<兼證 rdf:ID="暑濕證兼證">心煩口渴</兼證>

<舌苔 rdf:ID="暑濕證舌苔">苔淡黃膩或黃膩</舌苔>

<脈象 rdf:ID="暑濕證脈象">脈濡數</脈象>

<建議處方 rdf:resource="新加香薷飲" />

</owl:Class>

<owl:Class rdf:ID="暑濕證主證">

<rdf:type rdf:resource="主證"/>

<rdfs:subClassOf rdf:resource="暑濕證"/>

</owl:Class>

<owl:Class rdf:ID="暑濕證衛表">

<rdf:type rdf:resource="衛表"/>

<rdfs:subClassOf rdf:resource="暑濕證主證"/>

<怕冷 rdf:ID="暑濕證怕冷">微惡寒</怕冷>

<發熱 rdf:ID="暑濕證發熱">身熱</發熱>

<汗 rdf:ID="暑濕證汗">少汗</汗>

<頭身 rdf:ID="暑濕證頭身">頭昏重脹痛肢體痠楚</頭身>

</owl:Class>

```
        <owl:Class rdf:ID="暑濕證肺">
                <rdf:type rdf:resource="肺"/>
                <rdfs:subClassOf rdf:resource="暑濕證主證"/>
                <鼻 rdf:ID="暑濕證鼻">鼻流濁涕</鼻>
                <咽 rdf:ID="暑濕證咽">咽痛</咽>
                <咳 rdf:ID="暑濕證咳">咳嗽</咳>
                <痰 rdf:ID="暑濕證痰">痰黏或黃或白</痰>
        </owl:Class>
        <owl:Class rdf:ID="荊防敗毒散">
        </owl:Class>
        <owl:Class rdf:ID="新加香薷飲">
        </owl:Class>
        <owl:Class rdf:ID="銀翹散">
        </owl:Class>
        <owl:TransitiveProperty rdf:ID="建議處方">
                <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
        </owl:TransitiveProperty>
</rdf:RDF>
```

Appendix VII Cold/Influenza sub-ontology in XML

<感冒> (<Cold>)

<風寒證> (<Wind Cold Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defense-exterior >)

<怕冷> (<Fear of Cold>)

<怕冷重>id="1"</怕冷重>

(<Serious Fear of Cold>id=”1”</Serious Fear of Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<發熱輕>id="1"</發熱輕>

(<Mild Fever>id=”1”</Mild Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<無汗>id="1"</無汗>

(<Absence of Sweating>id=”1”</Absence of Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭痛四肢痠痛>id="1"</頭痛四肢痠痛>

(<Headache and Limbs Pain>id=”1”</Headache and Limbs Pain>)

</頭身> (</Head and Body>)

</衛表> (</Defense-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻塞流清涕多嚏>id="1"</鼻塞流清涕多嚏>

(<Nasal Congestion, Clear Sniffles and Profuse Sneezing>id=”1”

</Nasal Congestion, Clear Sniffles and Profuse Sneezing>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽癢>id="1"</咽癢> (<Throat Itching>id=”1” </Throat Itching>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽聲重>id="1"</咳嗽聲重>

(<Profuse Coughing>id=”1”</Profuse Coughing>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰稀薄色白>id="1"</痰稀薄色白>

(<White Clear Phlegm>id=”1”</White Clear Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<口不渴或渴喜熱飲>id="1"</口不渴或渴喜熱飲>

(<Not Thirsty Nor Fancy Hot Drinks>id=”1”</Not Thirsty Nor Like Hot Drinks>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌苔薄白>id="1"</舌苔薄白> (<Thin White Fur>id=”1”</Thin White Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈象浮或兼緊>id="1"</脈象浮或兼緊>

(<Floating Tight Pulse>id=”1”</Floating Tight Pulse>)

</脈象> (/<Pulse>)

</風寒證> (</Wind Cold Syndrome >)

<風熱證> (<Wind Heat Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defense-exterior >)

<怕冷> (<Fear of Cold>)

<怕冷輕>id="2"</怕冷輕>

(<Slightly Fear of Cold>id=”2”</Slightly Fear of Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<發熱重>id="2"</發熱重> (<Profuse Fever>id=”2”</Profuse Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<有汗>id="2"</有汗> (<Sweating>id=”2”</Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭脹痛>id="2"</頭脹痛>

(<Distending Headache>id=”2” </Distending Headache>)

</頭身> (</Head and Body>)

</衛表> (</Defense-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻塞流黃濁涕>id="2"</鼻塞流黃濁涕>

(<Nasal Congestion and Yellow Turbid Sniffles>id=”2”</Nasal Congestion and Yellow Turbid Sniffles>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽疼痛紅腫>id="2"</咽疼痛紅腫>

(<Sore Throat and Swelling>id=”2” </Sore Throat and Swelling>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽啞氣粗>id="2"</咳嗽啞氣粗>

(<Cough with Hoarseness Sound>id=”2”</Cough with Hoarseness Sound>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰稠黏色黃或白>id="2"</痰稠黏色黃或白>

(<Yellow or White Turbid Phlegm>id=”2”</Yellow or White Turbid Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<口乾欲飲>id="2"</口乾欲飲>

(<Thirsty and like to Drink>id=”2”</Thirsty and like to Drink>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<薄白而乾或薄黃尖邊紅>id="2"</薄白而乾或薄黃尖邊紅>

(<Dry Thin White Fur or Red Tips and Margins of Tongue with Thin Yellow Fur>id=”2” </Dry Thin White Fur or Red Tips and Margins of Tongue with Thin Yellow Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈浮數>id="2"</脈浮數>

(<Floating and Rapid Pulse>id=”2”</Floating and Rapid Pulse>)

</脈象> (</Pulse>)

</風熱證> (</Wind Heat Syndrome>)

<暑濕證> (<Summerheat-dampness Syndrome>)

<主證> (<Main Symptom>)

<衛表> (<Defense-exterior >)

<怕冷> (<Fear of Cold>)

<微惡寒>id="3"</微惡寒>

(<Mild Aversion to Cold>id=”3”</Mild Aversion to Cold>)

</怕冷> (</Fear of Cold>)

<發熱> (<Fever>)

<身熱>id="3"</身熱> (<Body Fever>id=”3”</Body Fever>)

</發熱> (</Fever>)

<汗> (<Sweating>)

<少汗>id="3"</少汗> (<Mild Sweating>id=”3”</Mild Sweating>)

</汗> (</Sweating>)

<頭身> (<Head and Body>)

<頭昏重脹痛肢體痠楚>id="3"</頭昏重脹痛肢體痠楚>

(<Heavy-headedness, Headache and Limbs Pain>id=”3”

</Heavy-headedness, Headache and Limbs Pain>)

</頭身> (</Head and Body>)

</衛表> (</Defense-exterior >)

<肺> (<Lung>)

<鼻> (<Nose>)

<鼻流濁涕>id="3"</鼻流濁涕>

(<Turbid Sniffles>id=”3”</Turbid Sniffles>)

</鼻> (</Nose>)

<咽> (<Throat>)

<咽痛>id="3"</咽痛> (<Sore Throat>id=”3”</Sore Throat>)

</咽> (</Throat>)

<咳> (<Cough>)

<咳嗽>id="3"</咳嗽> (<Cough>id=”3”</Cough>)

</咳> (</Cough>)

<痰> (<Phlegm>)

<痰黏或黃或白>id="3"</痰黏或黃或白>

(<Yellow or White Turbid Phlegm>id=”3”</Yellow or White Turbid Phlegm>)

</痰> (</Phlegm>)

</肺> (</Lung>)

</主證> (</Main Symptom>)

<兼證> (<Other Symptom>)

<心煩口渴>id="3"</心煩口渴>

(<Vexation and Thirsty>id=”3”</Vexation and Thirsty>)

</兼證> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<苔淡黃膩或黃膩>id="3"</苔淡黃膩或黃膩>

(<Pale Yellow or Yellow Slimy Fur>id=”3”</Pale Yellow or Yellow Slimy Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<脈濡數>id="3"</脈濡數>

(<Rapid and Soggy Pulse>id=”3”</Rapid and Soggy Pulse>)

</脈象> (</Pulse>)

</暑濕證> (</Summerheat-dampness Syndrome>)

</感冒> (</Cold>)

Appendix VIII Insomnia sub-ontology in XML

<不寐> (<Sleep Inability/Insomnia>)

<痰熱內擾> (<Internal Harassment of Phlegm-heat>)

<主症> (<Main Symptom>)

<胸悶心煩不寐>id="1"</胸悶心煩不寐>

(<Oppression in the Chest, Vexation and Sleep Inability>id=”1”

</Oppression in the Chest, Vexation and Sleep Inability>)

<泛惡>id="1"</泛惡> (<Malign Flood>id=”1”</Malign Flood>)

<噯氣>id="1"</噯氣> (<Belching>id=”1”</Belching>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<頭重目眩>id="1"</頭重目眩>

(<Heavy Headedness and Dizzy Vision>id=”1”</Heavy Headedness and Dizzy Vision>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<苔黃膩>id="1"</苔黃膩>

(<Slimy and Yellow Fur>id=”1”</Slimy and Yellow Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<滑數>id="1"</滑數>

(<Slippery and Rapid Pulse>id=”1”</Slippery and Rapid Pulse>)

</脈象> (</Pulse>)

</痰熱內擾> (</Internal Harassment of Phlegm-heat>)

<陰虛火旺> (<Yin Deficiency with Effulgent Fire>)

<主症> (<Main Symptom>)

<心悸心煩不寐>id="2"</心悸心煩不寐>

(<Palpitations, Vexation and Sleep Inability>id=”2” </Palpitations, Vexation and Sleep Inability>)

<腰痠足軟>id="2"</腰痠足軟>

(<Lumbar Aching and Foot Soft>id=”2”</Lumbar Aching and Foot Soft>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<頭暈耳鳴>id="2"</頭暈耳鳴>

(<Dizziness and Tinnitus>id=”2”</Dizziness and Tinnitus>)

<健忘遺精>id="2"</健忘遺精>

(<Forgetfulness and Seminal Emission>id=”2”<Forgetfulness and Seminal Emission>)

</兼症> (</Other Symptom>)

<脈象> (<Pulse>)

<沉細>id="2"</沉細> (<Sunken and Fine Pulse>id=”2”</Sunken and Fine Pulse>)

</脈象> (</Pulse>)

</陰虛火旺> (</Yin Deficiency with Effulgent Fire>)

<心脾兩虛> (<Dual Deficiency of the Heart-spleen>)

<主症> (<Main Symptom>)

<多夢易醒>id="3"</多夢易醒>

(<Profuse Dreaming and Easy Wake Up>id=”3”</Profuse Dreaming and Easy Wake Up>)

<心悸健忘>id="3"</心悸健忘>

(<Palpitations and Forgetfulness>id=”3”</Palpitations and Forgetfulness>)

<神疲食少>id="3"</神疲食少>

(<Lassitude of Spirit and Loss of Appetite>id=”3”</Lassitude of Spirit and Loss of Appetite>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<面色少華>id="3"</面色少華>

(<Pale White Complexion>id=”3”</Pale White Complexion>)

<四肢倦怠>id="3"</四肢倦怠>

(<Limbs Overexertion>id=”3”</Limbs Overexertion>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌淡>id="3"</舌淡> (<Pale Tongue>id=”3”</Pale Tongue>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<細>id="3"</細> (<Fine Pulse>id=”3”</Fine Pulse>)

</脈象> (</Pulse>)

</心脾兩虛> (</Dual Deficiency of the Heart-spleen>)

</不寐> (</Sleep Inability/Insomnia >)

Appendix IX Constipation sub-ontology in XML

<便秘> (<Constipation>)

<熱秘證> (<Heat Constipation>)

<主症> (<Main Symptom>)

<大便乾結>id="1"</大便乾結>

(<Hard Bound Stool>id=”1”</Hard Bound Stool>)

<腹部脹滿>id="1"</腹部脹滿> (<Abdominal Fullness>id=”1”<Abdominal Fullness>)

<口乾>id="1"</口乾> (<Dry Mouth>id=”1”</Dry Mouth>)

<口臭>id="1"</口臭> (<Fetid Mouth Odor>id=”1”</Fetid Mouth Odor>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<小便短赤>id="1"</小便短赤>

(<Difficult Painful Urination>id=”1”</Difficult Painful Urination>)

<口舌生瘡>id="1"</口舌生瘡>

(<Sore in Mouth and Tongue>id=”1”</Sore in Mouth and Tongue>)

<身熱面赤>id="1"</身熱面赤> (<Body Fever>id=”1”</Body Fever>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌質紅>id="1"</舌質紅> (<Red Tongue>id=”1”</Red Tongue>)

<苔燥>id="1"</苔燥> (<Dry Fur>id=”1”</Dry Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<滑數>id="1"</滑數> (<Slippery and Rapid Pulse>id=”1”</Slippery and Rapid Pulse>)

</脈象> (</Pulse>)

</熱秘證> (</Heat Constipation>)

<氣秘證> (<Qi Constipation>)

<主症> (<Main Symptom>)

<大便秘結>id="2"</大便秘結>

(<Hard Bound Stool>id=”2”</Hard Bound Stool>)

<噯氣頻作>id="2"</噯氣頻作>

(<Frequent Belching>id=”2”</Frequent Belching>)

<脅腹脾滿>id="2"</脅腹脾滿>

(<Hypochondrium, Abdominal and Spleen Fullness>id=”2”</Hypochondrium, Abdominal and Spleen Fullness>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<腹中脹滿而痛>id="2"</腹中脹滿而痛>

(<Abdominal Fullness and Pain>id=”2”</Abdominal Fullness and Pain>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌苔薄膩>id="2"</舌苔薄膩>

(<Thin Tongue and Slimy Fur>id=”2”</Thin Tongue and Slimy Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<弦>id="2"</弦> (<String-like Pulse>id=”2”</String-like Pulse>)

</脈象> (</Pulse>)

</氣秘證> (<Qi Constipation>)

<氣虛證> (<Qi Deficiency>)

<主症> (<Main Symptom>)

<臨廁努挣乏力挣則汗出短氣>id="3"</臨廁努挣乏力挣則汗出短氣>

(<Sweating and Shortness of Breath during Bowel Movement >id=”3”</Sweating and Shortness of Breath during Bowel Movement >)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<面色白>id="3"</面色白> (<Pale Complexion>id=”3”</Pale Complexion>)

<神疲氣祛>id="3"</神疲氣祛>

(<Lassitude of Spirit and Eliminate Qi>id=”3”</Lassitude of Spirit and Eliminate Qi>)

<便後乏力>id="3"</便後乏力>

(<Lack of Strength after Bowel Movement >id=”3”</Lack of Strength after Bowel Movement>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<舌淡嫩>id="3"</舌淡嫩> (<Pale and Soft Tongue>id=”3”</Pale and Soft Tongue>)

<苔薄>id="3"</苔薄> (<Thin Fur>id=”3”</Thin Fur>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<虛>id="3"</虛> (<Vacuous Pulse>id=”3”</Vacuous Pulse>)

</脈象> (</Pulse>)

</氣虛證> (</Qi Deficiency>)

<血虛證> (<Blood Deficiency>)

<主症> (<Main Symptom>)

<大便秘結如栗>id="4"</大便秘結如栗>

(<Hard Bound Stool>id=”4”</Hard Bound Stool>)

<面色萎黃無華>id="4"</面色萎黃無華>

(<Pale Yellow Complexion>id=”4”</Pale Yellow Complexion>)

</主症> (</Main Symptom>)

<兼症> (<Other Symptom>)

<頭暈目眩>id="4"</頭暈目眩> (<Dizziness>id=”4”<Dizziness>)

<心悸>id="4"</心悸> (<Palpitations>id=”4”<Palpitations>)

</兼症> (</Other Symptom>)

<舌苔> (<Tongue Fur>)

<唇舌淡>id="4"</唇舌淡> (<Pale Mouth and Tongue>id=”4”</Pale Mouth & Tongue>)

</舌苔> (</Tongue Fur>)

<脈象> (<Pulse>)

<細>id="4"</細> (<Fine Pulse>id=”4”</Fine Pulse>)

</脈象> (</Pulse>)

</血虛證> (</Blood Deficiency>)

</便秘> (</Constipation>)

Appendix X Emails of Acceptance

intechweb.org

OFFICE
In-Tech
Kirchengasse 43/3
A-1070 Vienna, Austria

In-Tech
Hotel 80
51000 Rijeka, Croatia
PHONE
+385 51 179 9650
+43 676 43 80631
FAX
+385 51 494 536
EMAIL
lazinica@in-techweb.org
WEB
intechweb.org

NOTIFICATION OF ACCEPTANCE

DATE
February 10, 2008

TO
Wilfred Lin
The Hong Kong Polytechnic University
PQ806, Mong Man Wai Building, The Hong Kong
Polytechnic University, Hung Hom, Kowloon,
Hong Kong
Hong Kong NA Hong Kong

In the name of Editorial Colleqiums it is my pleasure to inform you that your chapter proposal under the tentative title "Dynamic Cache Size Tuning for Happier, Return E-shoppers in Mobile Businesses" written by Jackei H.K. Wong, Allan K.Y. Wong and Wilfred W.K. Lin is accepted for publishing in the book "E-Commerce", ISBN 978-953-7619-X-X.

Sincerely yours,
Aleksandar Lazinica

IN-TEH
D.O.O. RIJEKA
EXCELLENCE IN EDUCATION AND PUBLISHING

From: tieedit@auburn.edu
To: wilfred.lin@ ,wilfred.lin@
Subject: Transactions on Industrial Electronics - Manuscript No: 09-0197-TIE.R1
Body: 16-Jun-2009

Dear. Wilfred Lin,

The reviewing process of your paper No: 09-0197-TIE.R1, entitled "A Novel Real-Time Traffic Sensing (RTS) Model to Improve the Performance of Web-based Industrial Ecosystems" has been completed.

Based on the opinions of the reviewers and the Associate Editor in charge, your manuscript has been provisionally accepted for publication in the IEEE Transactions on Industrial Electronics. Please accept my congratulations!

Before submitting the final version please kindly update the reference section with recently published work (since the submision of the original manuscript). Searching through the literature for TIE papers you may find http://tie.ieee-ies.org/tie/abs/index.htm more convenient than IEEE Xplore.

While preparing your final version of manuscript, please print out and follow exactly instructions on http://tie.ieee-ies.org/tie/final_sub.html. The final manuscript and all publication items must be uploaded in one ZIP (RAR) file from your Author Center on Manuscript Central. See http://tie.ieee-ies.org/tie/final_sub.html for details.

IMPORTANT – Please be sure that you follow exactly the instructions for preparation of the zip file (http://tie.ieee-ies.org/tie/final_sub.html). You may ask another person to check if everything is in order. With new submission process of final paper through Manuscript Central neither you nor we have an option to undo your final submission. As consequence, corrections to your zip file may delay publication of your paper up to a couple of months. Please do not submit the zip file unless you are 100% sure that everything is correct.

Your kind cooperation will be greatly appreciated.

Sincerely yours,
Prof. Bogdan Wilamowski
Editor-in-Chief
Transactions on Industrial Electronics
http://tie.ieee-ies.org/tie/

From: Tharam S Dillon [mailto:tharam2@]
Sent: Wednesday, April 01, 2009 1:20 PM
To: Jackei Wong
Subject: [IJCSSE] C1439 Notification - Accepted

Dear Jackei Wong,

I am pleased to inform you that your paper C1439 entitled "A Novel Approach to Achieve Real-time TCM (Traditional Chinese Medicine) Telemedicine Through the Use of Ontology and Clinical Intelligence Discovery)" has been accepted for publication in the International Journal of Computer Systems Science and Engineering.

Please check the 'Information for Authors' for the preparation of all required files for final publication. In particular, note the requirement for image files. Please prepare a single compressed file (e.g. zip) which contains a copy of your paper in PDF format, and all text/image source files that are needed for publishing. To upload your file, please do the following:

1. Go to: http://csse.debii.curtin.edu.au/index.php/csse

2. Enter your login details

3. Click [Author]

4. Click [Accepted], if you currently have more than one active submission; make sure you select the correct paper.

5. Go to the end of the page, and follow the instruction in the section of 'Revised/Final Version Upload' to upload your files.

6. Once completed the upload, click the [Notify Editor] mail icon to notify the editor about the revised submission.

You should also make sure the 'Submission Metadata' is correct on the 'Summary' page. These details are use in the final version and the online TOC; incorrect details may result in your paper not reference properly by other readers and authors.

We will be in contact with you soon in case the journal requires a copyright form for your publication. Any enquiries thereafter regarding the publication of your journal, you may contact the Executive Editor of CRL Publishing, Mr. Jeremy Thompson, his email address is: csse@crlpublishing.co.uk

Congratulations again on the acceptance of your paper. If you have any further questions, please feel free to get in touch.

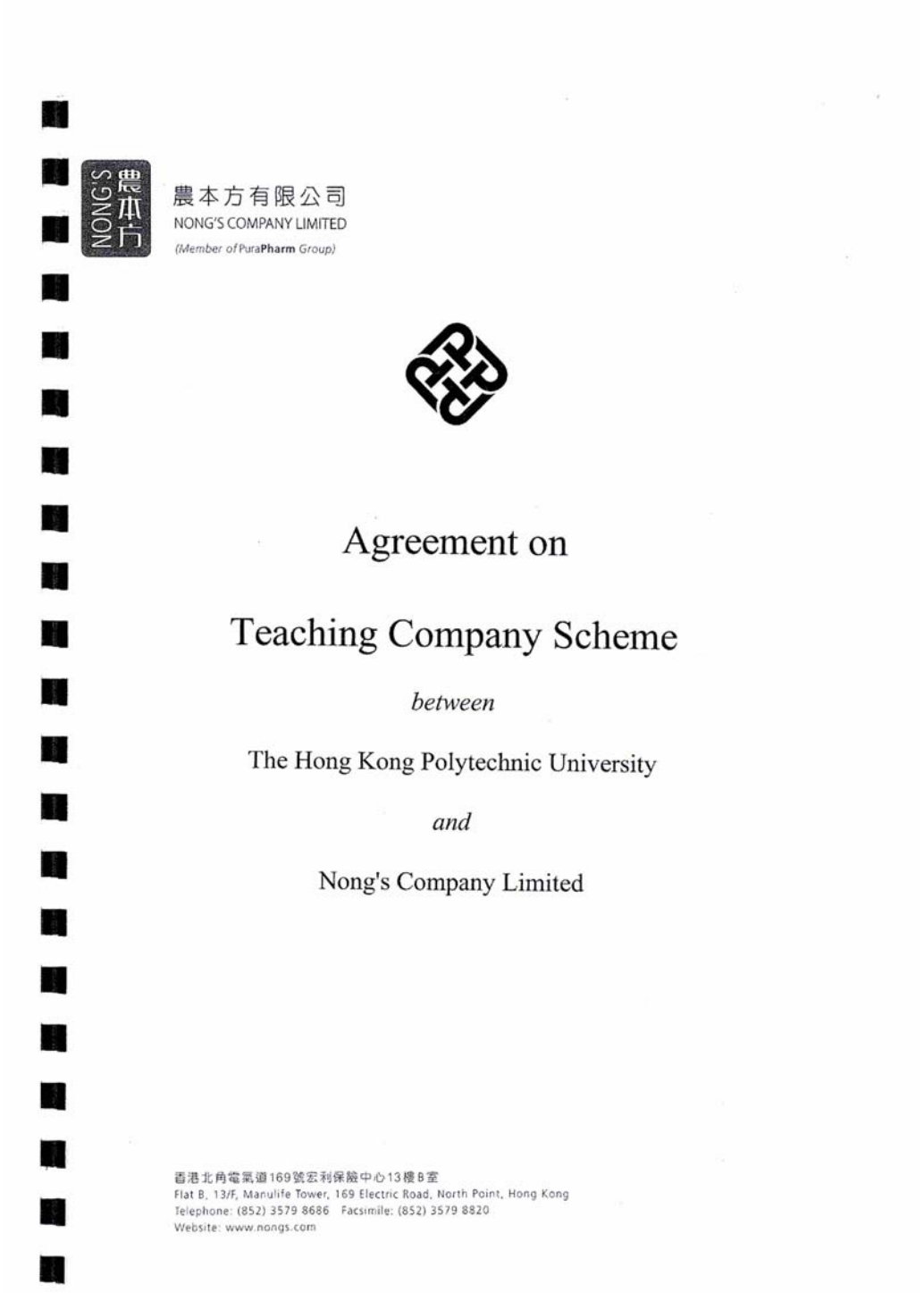
-----Comments------
Please make the following modification for the final version:
NIL

Yours sincerely

Professor T.S. Dillon
Editor-in-Chief

IJCSSE - http://csse.debii.curtin.edu.au/

Appendix XI TCS Contract (Extracts)



農本方有限公司
NONG'S COMPANY LIMITED
(Member of PuraPharm Group)

Agreement on

Teaching Company Scheme

between

The Hong Kong Polytechnic University

and

Nong's Company Limited

香港北角電氣道169號宏利保險中心13樓B室
Flat B, 13/F, Manulife Tower, 169 Electric Road, North Point, Hong Kong
Telephone: (852) 3579 8686 Facsimile: (852) 3579 8820
Website: www.nongs.com

農本方有限公司
NONG'S COMPANY LIMITED
(Member of PuraPharm Group)

THIS AGREEMENT is made the 17th day of October 2006.

BETWEEN

(1) THE HONG KONG POLYTECHNIC UNIVERSITY (the "PolyU") whose registered address situates at Hung Hom, Kowloon, Hong Kong

AND

(2) NONG'S COMPANY LIMITED (the "Partner") whose registered address situates at Suites 4103-08, Jardine House, 1 Connaught Place, Central, Hong Kong,

hereinafter collectively referred to as the "Parties".

WHEREAS

1. The Teaching Company Scheme Programme ("Programme") is designed and conducted by the PolyU in support of industry and business aiming to bring ideas through research activities to results that can be used by companies to support their growth and development; and

2. The Parties agree to jointly develop a Teaching Company Scheme under the Programme subject to the terms and conditions of this Agreement.

NOW, BOTH PARTIES HAVE HEREBY AGREED WITH EACH OTHER THAT:

(A) In this Agreement, unless otherwise specified:

(a) words importing the plural include the singular and vice versa;

(b) words importing a gender include every gender;

(c) the headings do not affect the interpretation of this Agreement;

(d) an Appendix forms part of this Agreement; and

(e) any reference to a Clause or an Appendix is a reference to a Clause of or an Appendix to this Agreement.

1. THE TEACHING COMPANY SCHEME

香港北角電氣道169號宏利保險中心13樓B室
Flat B, 13/F, Manulife Tower, 169 Electric Road, North Point, Hong Kong
Telephone: (852) 3579 8686 Facsimile: (852) 3579 8820
Website: www.nongs.com

1.1 For the mutual interest and benefits to the Parties and subject to the terms and conditions of this Agreement, the Parties agree to jointly develop a Teaching Company Scheme ("TCS") under which the Parties shall provide facilities and guidance to the TCA (as defined in Clause 2.1) in a research project of high intellectual content and of application and/or research value to the Parties according to agreed-upon plan and objectives ("Research Project"). The specifications of the TCS and the Research Project are stipulated in Appendix I (such specifications may be modified or varied with the mutual agreement of the Parties from time to time).

1.2 This Agreement shall commence on the date hereof until the expiry of the Implementation Period stipulated in Appendix I, unless extended for such further period (if any) as the Parties may mutually agree or terminated in accordance with the provisions contained in this Agreement.

2. TEACHING COMPANY ASSOCIATE

2.1 A researcher will be * recruited / appointed by the Partner at its sole discretion as the Teaching Company Associate ("TCA") to conduct the Research Project.

2.2 Upon completion of the Research Project, subject to the satisfactory completion of a thesis and applicable regulations of the PolyU for the Degree of * Master of Philosophy / Doctor of Philosophy offered by the PolyU, the TCA may be conferred a postgraduate degree by the PolyU.

2.3 The qualification of the TCA as well as the nature of the Research Project shall satisfy the PolyU Research Committee's requirement for the registration of a higher degree programme.

2.4 It shall be the sole responsibility of the TCA to register at his own expense for the higher degree of * MPhil / PhD sought in accordance with the PolyU's regulations. The PolyU reserves the exclusive right to withdraw the registration of the TCA from such higher degree programme in accordance with the PolyU's regulations, provided that such decision will only be taken after full consultation with the Partner.

3. RESPONSIBILITY OF THE PARTNER

The Partner shall, subject to the terms and conditions of this Agreement:

3.1 at its own expense, provide all the necessary facilities reasonably available to support the TCA's research activities in the Research Project subject to operational needs and availability of resources.

3.2 appoint an industrial supervisor from among its own staff to guide and support the TCA's research activities in the Research Project.

3.3 pay the TCA direct all the studentships that the TCA is entitled to receive during the term of this Agreement.

3.4 provide, in collaboration with the PolyU's academic director, continuous project support to the TCA on the development, execution, writing-up, submission and examination of the TCA's academic research thesis for his academic award programme.

3.5 provide adequate time allowance for the TCA to cater for his work in the Research Project, in acknowledgment of the necessary commitment of the TCA as a registered * full-time / part-time higher degree student.

4. RESPONSIBILITY OF THE POLYU

The PolyU shall:

4.1 provide necessary and appropriate guidance and supervision of a high academic standard for the TCA's research activities in the Research Project through the PolyU's academic director who will maintain typically weekly contact with the TCA. Whenever deemed necessary, support and supervision may also be provided by other academic staff of the PolyU. For the purpose of this Agreement, the academic director and/or other members, staff or students of the PolyU who participate or are involved in the Research Project are collectively referred to as the "PolyU Research Team".

4.2 allow, whenever necessary or reasonably requested by the TCA, the TCA to use facilities available at the PolyU to conduct his research activities for the Research Project subject to rules and regulations of the PolyU as stated in the PolyU student handbook.

4.3 allow, where permitted by applicable regulations as stated in the PolyU student handbook or reasonably requested by the TCA, the TCA to use campus and library facilities at the PolyU during the course of the Research Project.

IN WITNESS WHEREOF, the Parties have executed this Agreement as of the date first above written.

For and on behalf of THE HONG KONG POLYTECHNIC UNIVERSITY	For and on behalf of NONG'S COMPANY LIMITED
Dr. LUI Sun-wing Vice President Partnership Development	Name : Mr Abraham Chan Title : Director

香港北角電氣道169號力寶中心13樓E室
Flat E, 13/F, Manulife Tower, 169 Electric Road, North Point, Hong Kong
Telephone: (852) 3579 8688 Facsimile: (852) 3579 8820
Website: www.nongs.com