A Multi-faceted & Automatic Knowledge Elicitation System for Managing Unstructured Information

Yong Wang

Ph.D

The Hong Kong

Polytechnic University

2010

The Hong Kong Polytechnic University

Department of Industrial and Systems Engineering

A Multi-faceted & Automatic Knowledge Elicitation System for Managing Unstructured Information

Yong Wang

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

June 2006

CERTIFICATE OF ORGINALITY

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.

__________________________(Signed)

___Yong Wang_____________(Name of student)

Abstract

Nowadays, knowledge is becoming a new competitive factor in the knowledge economy. Knowledge work deals with a huge of information and its manipulation. However, many researchers point that most of all potentially usable business information originates in an unstructured form. Unstructured Information Management (UIM) is becoming the current state-of-the-art of technology.

In this thesis, a Multi-faceted and Automatic Knowledge Elicitation System (MAKES) is proposed to manage the mass of unstructured information and support knowledge work. There are four phases in MAKES. The first phase is collecting data automatically and text mining. Multiple patterns of dynamic taxonomies are developed to classify the unstructured information. In the second phase, some knowledge models are adopted to represent knowledge elicited from large amounts of unstructured information. Concept Relationship Model (CRM) illustrates the relationships of concepts elicited from unstructured information. The algorithm named Concept Relationship Exploring Technique (CRET) which is developed to measure the relationship of two concepts. A pattern of knowledge flow can be captured as a Dynamic Knowledge Flow Model (DKFM). Knowledge Capability Model (KCM) evaluates the knowledge capability of a knowledge worker from the traffic of knowledge flow. Thirdly, a multi-faceted navigation platform supports the analysis of knowledge models. Finally, some reports about managing unstructured information and knowledge assets of organizations are produced derived from the knowledge models.

The capability and advantages of the MAKES are demonstrated through the two cases of verification testing. The first case is an application of MAKES in an emergency management system in the Committee of Guang Zhou City Management of the Guang Zhou Municipal Government in China. The efficiency of decision-making when responding to emergency incidents is evaluated based on an evaluation architecture. Another case is applying MAKES to a knowledge audit in an electronic trading firm in Hong Kong. MAKES has significantly improved the efficiency of the knowledge audit process. This research points to a new direction of automatic text mining and to the elicitation of useful organizational knowledge from the very large amount of dynamic and unstructured information that is often neglected in an organization.

Keywords: unstructured information management; text mining; knowledge elicitation; knowledge model; multi-faceted navigation; knowledge audit; emergency management.

Acknowledgements

The author would like to express his sincere gratitude and appreciation to his chief supervisor, Professor W.B. Lee, Chair Professor, Head of Department of Industrial and Systems Engineering, for his guidance and support throughout the author's research study period. The author also wishes to express his sincere thanks to his co-supervisor, Dr. Benny C.F. Cheung, for his guidance and help.

Thanks are due to Dr. Kwok, S.K., from whom the author benefited so much, for his unselfish help and great encouragement. Thanks are also due to the author's fellow research students and friends both within and outside The Hong Kong Polytechnic University for their friendship and help.

The author would like to thank the Committee of Guang Zhou City Management, the Municipal Government and Angus Electronics Co. Ltd., for their support on the case studies. The author would also like to thank the staff of the Microsoft Enterprise Systems Centre for their provision of administration and technical support.

Special thanks are due to the author's wife Ying Wang and his lovely daughter Kexin Wang for their attention, support and love.

Finally, thanks for all people who has supported the author throughout the project.

TABLE OF CONTENT

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Importance of knowledge work

Since the early 1990’s, the importance of Knowledge Management (KM) for creating an organization’s competitive advantage in the knowledge economy has been widely recognized worldwide in the government, public and private sectors. To be successful, enterprises and organizations do not compete only in terms of physical resources and financial capital, but also in terms of intangible knowledge assets that they are capable of creating and leveraging.

About one hundred years ago, Frederick W. Taylor (1911) applied the method of “motion analysis”, and built the milestone theory of “Scientific Management”. The method of Taylor’s scientific management increases the efficiency of labor and production, and builds the paradigm of mass production for the industrial society. During the last three decades, the advances in information and communication technology have changed the competitiveness landscape. Knowledge has replaced labor, land and capital as the major factor of production (Drucker, 1980). The competitiveness among business organizations is now embodied in the knowledge possessed by these organizations. Hence, the new economy is driven by knowledge creation and innovation. Knowledge is becoming a new kind of capital in the growth of the economy (Toffler, 1990).

Knowledge can be explicit or tacit. Explicit knowledge is available in documents, books, white papers, databases, etc. Researchers like Al-Hawamdeh (2003) and Nonaka and Takeuchi (1995) have used the term “explicit knowledge” as synonymous to information. Tacit knowledge resides in the heads of people, such as insight, intuition, experience, memories, etc. Both types of knowledge are important. In this thesis, unless otherwise specified, the word knowledge refers to knowledge that can be made explicit, that is, that which can be codified and recorded.

In knowledge-based enterprises, the main actors are knowledge workers who create and use knowledge to solve daily business problems. New business know-how and decisions often depend on the knowledge generated from the interaction among workers in day to day business processes (Druker, 1980; Nickols, 2000). The term knowledge workers coined by Peter Drucker in 1959, is one who works primarily with information or one who develops and uses knowledge in the workplace. Knowledge work which is based on information and the manipulation of information has largely replaced physical work (Despres and Hiltrop, 1995), and is increasingly regarded as an important contributing factor to business survival and success (Drucker, 1999; Kelloway and Barling, 2000; Holder, 2002).

From the nineteen nineties, Knowledge Management (KM) has been is promoted as a “new” management mind-set in all fast-moving organizations and corporations (Wiig, 1993; O’Dell and Grayson, 1998; Davenport and Prusak, 1998). The purpose of KM is to enhance organizational knowledge processing and sharing (Davenport, 1998; Galagan, 1997). Wiig (2004) defines KM as the systematic, explicit, and deliberate building, renewal, and application of knowledge and other intellectual capital assets to maximize the enterprise’s knowledge-related effectiveness, returns, and success. More and more companies recognize the importance of knowledge management for leveraging the intellectual capital of an organization.

To manage knowledge effectively and securely, Wiig (1993) suggested a framework of three pillars of KM that comprise many valid and relevant perspectives and approaches such as surveys and categorizing knowledge, evaluating knowledge, and synthesizing knowledge and related activities. Nonaka (1991; Nonaka and Takeuchi, 1995; Nonaka, 1994) built the SECI model and provided a framework to support the conversion processes between explicit and tacit knowledge. Inspired by Nonaka's work, the focus of KM has shifted from knowledge codification to knowledge conversion and creation, which becomes the paradigm of the second generation KM that we know today. The concept of intellectual capital (IC) according to Stewart (1997) has also attracted people's attention to the importance of assessing the knowledge assets of the whole organization.

In the past, knowledge was treated as a "thing" that could be transferred and passed around like a bucket of water. Now we realize that knowledge is dynamic and interwoven in the interrelationships and social network of people. Understanding and capturing the flow of knowledge in a social network is one of the main concerns in the third generation KM based on a social complexity framework advocated by Snowden (2002). In research of work patterns of organizations, the traditional work is mainly physical and follows fixed regulations and procedures. It's basically a linear workflow such as in the assembly line of a manufacturing firm. Most of the research about workflow management focuses on how to build a clear work plan and decrease the variety of work to the best one’s ability to improve the efficiency of the workflow.

Nowadays, the main economic value has shifted from physical resources to knowledge production. Knowledge work involves the collection and acquisition of unstructured information from multiple sources, and complex knowledge work requires synthesizing and analyzing and making use of the scattered information from various sources. For example, during customer service offered by manufacturers, the call center needs the helps of other units, such as design, technical support, quality management, and even suppliers. Furthermore, collaborative work among various units is needed to support the customer services. This is a complex non-linear knowledge network that has to be built for the call centre, for it to operate effectively. One of the most obvious characteristics of non-linear work is that the demand of the knowledge needs and its flow among its workers is constantly changing due to both frequent internal and external changes. In this regard, traditional reliance on rigid organization processes and databases (such as organization charts and expertise directory) to locate the knowledge source may not sufficient to respond to the knowledge needs of an organization. There is a need to address how important knowledge assets of an organization can be managed effectively.

1.2 Motivation and scope of study

Most of the knowledge in an organization is unstructured or semi-structured (Waters, 2005). There is a voluminous amount of unstructured knowledge that often represents the key intellectual assets of an organization. Many companies realize that a lot of valuable knowledge is hidden and not attended to in an organization. There are many unstructured knowledge assets such as the knowledge residing in people heads, email communications and various electronic resources and repositories. The unstructured information is not only coming from external knowledge sources but is also created internally such as on emails, intranet, corporate platform, or knowledge portal, blogs, etc. Email messages are one of the important sources of unstructured information nowadays and the quantity of information and knowledge that is embedded them is massive. Emails are used for a multitude of different processes that stretch far from “simple message transfer” and range from storing documents and contact information, to filing bookmarks, literature references and contact information, and to managing tasks and reminders (Venolia et al., 2001; Ducheneaut and Bellotti, 2001; Swaak et al., 2004). The ability to exploit unstructured information has turned into a competitive differentiator, and is one of the biggest challenges in knowledge management. The acquiring, codifying, classifying, storing, analyzing and representing this knowledge into useful knowledge from various electronic resources is a formidable task. Although many companies have established different systems to handle different kinds of unstructured knowledge assets such as email management systems, most of them do not have a well defined taxonomy which can systematically organize and classify the unstructured information from a mass of textual documents and items. It is inconvenient and difficult for staff to share and organize unstructured information effectively using manual methods. Most of the unstructured information of an organization even if captured is seldom analyzed nor utilized.

According to Singh (2007), Information Management (IM) is concerned with managing structured and formalized information, which can easily be identified, organized and distributed, whereas Knowledge Management (KM) basically deals with unstructured, informal information which cannot be easily identified, extracted or managed. Magnus Stensmo (2003) has proposed tools and methodologies for Unstructured Information Management (UIM), but the focus of UIM is on the management of unstructured information and files which are different from managing knowledge, as knowledge cannot be managed directly. Only the information about the knowledge possessed by the people in an organization can be managed (Streafield and Wilson, 1999). Identifying the information possessed by the people in an organization is the first important step in managing the knowledge assets of every organization. The electronic information (such as in e-mails, groupware, portals, blogs and instant messaging) from both external sources and internal sources of an organization grows exponentially with time, and a huge amount of tacit knowledge that hidden in the knowledge flow and social network of an organization is embedded in these sources. The traditional knowledge audit is an attempt to find out what knowledge exists and what is missing, where it is, where and how it is being created and who owns it. However, the process is very labour intensive and time consuming as this requires of huge amount of time to collect and elicit the knowledge items from interviews, surveys, form-filing and document searching, etc. Due to the short life cycle of knowledge in some fast moving organizations and sectors, the drawback of the manual approach when identifying information possessed by people is that the data can rapidly become obsolete. A similar scenario occurs in emergency management in which a huge amount of unstructured information is encountered and processed to enable an intelligent decision to be made. So far, there has been little research reported in the literature on the systematic capturing of unstructured electronic information, its classification and analysis and how it can be used for revealing the patterns and dynamics of knowledge work and knowledge flow among individuals in an organization. It is the aim of this thesis to fill this gap.

To manage the unstructured information effectively, it is necessary to construct an architecture which supports knowledge work by applying multiple technologies to acquire, analyze, represent, store and share unstructured information. Technologies for managing unstructured information include statistical and rule-based Natural Language Processing (NLP), Information Retrieval (IR), Machine Learning (ML), ontologies, dynamic taxonomy and automated reasoning (Ferrucci and Lally, 2004a; Stensmo, 2003). The ways to represent and organize knowledge is also an important aspect of managing unstructured information. A scientific representation of knowledge from unstructured information in an explicit form for it to be conveniently stored, shared and transferred is needed. Various methods exit to help reveal patterns of knowledge from unstructured information. For example, a knowledge map is one method to represent, store and display knowledge in the form of graphical patterns so as to easily visualize the knowledge entities and their inter-relationships.

The organization and interconnection of knowledge components leads to knowledge networks. Understanding and capturing knowledge flow in an organization is the critical factor in designing business processes (Newman, 2002). By making the communication flow transparent, business processes can be made more efficient, allowing organizations to make better use of people by freeing them from being buried in conventional multilayer hierarchies and inefficient business processes. The primary objective of capturing knowledge flow is to enable the transfer of capability and expertise from where it resides to where it is needed – across time, space and organizations, as necessary (Nissen, 2002). It is noted that understanding how knowledge flows through an enterprise is critical to the increase of productivity of knowledge work. Until we can understand the phenomenon of knowledge flow, we are unlikely to design effective flow-enhancing interventions (Nissen and Levitt, 2002). Social network mapping or Social Network Analysis (SNA) is one of the most important techniques used in knowledge mapping. SNA maps the relationships and flows between people, groups, organizations, computers, or other information / knowledge processing entities. Social network analysis shows networks of knowledge and patterns of interaction among group members, organizations, and other social entities. SNA provides both a visual and a mathematical analysis of human relationships to identify the roles in organizational networks, central people, peripheral people, boundary spanners and knowledge brokers (Cross et al., 2004). The combination and synthesis of SNA and knowledge flow can provide a comprehensive exploration and audit of knowledge assets in an enterprise. Overall, the complexity of the issues involved in managing unstructured information warrants a system approach to integrate all the emerging technologies in knowledge engineering discussed above to build a flexible platform to effectively convert the unstructured information owned by individuals into useful knowledge assets (such as knowledge capability and knowledge and communication network of dispersed units) of an organization.

As a result, there is a need for the establishment of an effective modular system which is dynamic enough to facilitate searching, navigation, analyzing, discovering and visualizing the organizational knowledge from among a very large number of unstructured information sources. The objectives of this research are thus:

i) To design a Multi-faceted and Automatic Knowledge Elicitation System (MAKES) for managing unstructured information and knowledge. Unstructured knowledge assets are automatically organized, classified and presented in a multi-facet taxonomy map.

ii) To develop the various technological components needed for the MAKES. These include the knowledge mining algorithm to automatically generate and self-maintain the evolution of the multi-faceted taxonomy maps based on the searching criteria, searching keywords, and the behaviors of the knowledge workers, and a purpose-built knowledge elicitation algorithm named the Concept Relationship Exploring Technique (CRET). Different kinds of knowledge models used in the reporting of the organization of knowledge assets are also explored.

iii) To verify and demonstrate the applicability of the MAKES through two case studies, one in emergency management and one in knowledge audit.

1.3 Layout of dissertation

This thesis contains 8 chapters. The background and motivation for the study is addressed in Chapter 1. This includes the importance of knowledge work as a key factor of competitiveness, the problems encountered in unstructured information management and the need for a multi-faceted automatic knowledge elicitation system to manage the dynamic electronic knowledge assets of an organization. In Chapter 2, a review is conducted on the current state-of-the-art of Unstructured Information Management (UIM) and its architecture, various technologies for knowledge elicitation, and the representation and organization of knowledge. Chapter 3 describes the construction of the MAKES through the four phases of information collection, knowledge modelling, multi-faceted navigation and analysis of knowledge flow, and generation of reports inferring from knowledge models. The methodology and the roadmap of the research are addressed. A multi-ties architecture with a multi-agent mechanism to implement MAKES is introduced. The functions and interfaces of various components of MAKES are also described in detail in this Chapter.

Chapter 4 focuses on the architecture of text mining and on the various patterns of dynamic taxonomy of unstructured information deployed in the construction of the MAKES. The thesaurus model and its maintenance are also described. The various knowledge models, Concept Relationship Model (CRM), Dynamic Knowledge Flow Model (DKFM), and Knowledge Capability Model (KCM) which constitute the main components of the system framework of MAKES, are introduced in Chapter5. Multi-faceted navigation for unstructured information, and the generation of reports about knowledge assets in the output tier of MAKES are illustrated in this chapter.

Then, two cases were developed and verified in different settings. Chapter 6 demonstrates how MAKES can be applied in the area of emergency management by the Committee of Guang Zhou City Management of the Guang Zhou Municipal Government in China. The prototype was developed, assembled and tested as a part of the Guang Zhou City Emergency Management System. The relevant information and knowledge about emergency management is uncovered, using a concept relationship model and dynamic knowledge flow model of MAKES. The other case is described in Chapter 7 which reports on the application of MAKES in conducting an audit of the knowledge assets of an electronic trading company in Hong Kong as a reference site for the project. The methods of intellectual property audit and analysis combining a dynamic knowledge flow model and a knowledge capability model are described. Recommendations about knowledge work are suggested. Lastly, the conclusion and suggestions for future work are discussed in Chapter 8.

Chapter 2. Literature Review

In the past few decades, knowledge has been become the core asset of many organizations. Among their various intangible assets, more and more organizations have recognized that a lot of strategic business value (such as networks, key players, process details, business negotiations, etc.) lies in unstructured information, such as e-mails, text documents, audio messages, video information, etc. Unstructured information is a vital part of knowledge sources. Research into unstructured information management will propel effectively the application of knowledge management in organizations to uncover hidden value, and enhance an organization's competitive capability.

2.1 Overview of unstructured information management

Merrill Lynch in 1998 cited estimates that as much as 80% of all potentially usable business information originates in an unstructured form (Aone and Ramos-Santacruz, 2000; Moore, 2002; Cristinaini and Shawe-Taylor, 2000). With this background, adding a structure to unstructured information will enable this information to be found and utilized more easily.

2.1.1 Concept of unstructured information

Grimes (2005) described unstructured data (or unstructured information) as computerized information that either does not have a data model or has one that is not easily usable by a computer program. Conversely, structured information refers to the information that either has a data model or is by nature easily usable by a computer program. Structured information is distinguished from unstructured information as being data which is stored in databases or in documents with semantic annotations. However, unstructured information has an inherent structure which can be inferred from the text, for example, through examining word morphology, sentence syntax, and other small and large-scale patterns. Examples of unstructured information include audio, video, and unstructured text, e.g. the content of an email. Grimes (2005) indicated that software that creates a machine-processable structure needs to exploit the linguistic, auditory, and visual structure that is inherent in all forms of human communication.

However, some data expressed in structured form should still be thought of as unstructured information if its structure is not helpful to the processing task. For example, an HTML Web page is annotated, but the HTML mark-up is not designed to represent the meaning or function to support the automated processing of the content of the page. Tags in XHTML allow machines to process elements, although the tags do not capture or convey the semantic meaning of tagged elements.

Unstructured information includes natural language, written documents, speech, audio, still images, and video (Ferrucci and Lally, 2004a). Approaches to processing unstructured information are becoming one of the most focused on research domains. The capability of using unstructured information will greatly improve the work intelligence of organizations.

Generally, unstructured information falls into two basic categories (Weglarz, 2004):

- Bitmap objects: inherently non-language based, such as image, video or audio files.
- Textual objects: based on a written printed language, such as Microsoft Word documents, e-mails or Microsoft Excel spreadsheets.

Although these objects are classified as being data in nature, the technologies and methodologies for processing bitmap objects are still primary and simple. Most technologies mainly focus on textual objects nowadays. The implementation discussed in this thesis focuses on natural language text.

Naturally, unstructured information lacks an obvious data model to describe characters and attributes of information, which is not easily to be read and processed by machines. Generally, semi-unstructured information has a certain data model which schemas the information so it can be analyzed and understood by machine. However, the data model of semi-unstructured information is not enough to support the automatic processing for information management. For example, email has several fields, e.g. subject, date, sender, receiver, body and etc. XML has a self-description data model. By nature, the content in the semi-structured information is still written using by natural language and is difficult to be understood by machine.

The comparison between unstructured information and structured information is summarized in Table 2.1.

Table 2.1 Comparison of unstructured information and structured information.

Item	**Unstructured information**	**Structured information**
Sources	Corporate Web sites, Email, Meeting notes, SMS, etc,	Databases, Spread Sheets, Forms, etc.
Representation	Natural Language, Movies, speech, images, etc.	Template, frame, field, XML, HTML, EDI, etc.

Unstructured information represents the largest and fastest growing source of knowledge available to businesses and governments world-wide. The amount of unstructured and semi-structured information in enterprises is growing rapidly, doubling every year, by some estimates (Waters, 2005; Moore, 2002). Management of information and knowledge is of two kinds: the management of structured information, and the management of unstructured information.

2.1.2 Managing unstructured information

Unstructured information traditionally is stored as documents in local hard disks or in file servers, or in email systems. The documents include research reports, memos, letters, white papers, presentations, etc. Unstructured information is generally represented in various forms. The lack of structure in unstructured information makes it difficult for it to be collected, accessed, categorized, and searched because such information has no effective association with meta-data.

Unstructured information is unmanaged information. In most organizations, there is a large amount of unstructured content which often represents the key intellectual assets of organizations. Because of the inherent difficulties of understanding unstructured content, many organizations are beginning to tackle this problem and are gaining some benefits. In an effort to control and manage this kind of information with visualization patterns, some technologies have been developed to meet the requirements for knowledge work in organizations.

Recently, Hatch (2007) showed that organizations are now starting to prioritize the use of unstructured data. Morris (2008) thought that the principal challenge with unstructured information is that it needs to be analyzed in order to identify, locate and relate the entities and relationships of interest, and to discover the vital knowledge contained therein.

Decomposing the whole process of unstructured information into various phases is a right approach to the management of unstructured information. These phases consist of text mining, categorization, information retrieval, portals, taxonomy generation, and so on. Moreover, capability of managing and searching XML data has been developed by many vendors. The common idea in all these technologies is that adding structure to unstructured data can make it easier to process.

According to the Boston-based researchers, the Delphi Group (Delphi Group, 2002), the new generation of information retrieval technologies will play a key role in managing collections of semi-structured and unstructured data by augmenting searching with meta-data creation, content and behavior profiling, and collaboration.

Software vendors who develop the processing software of unstructured data include Autonomy, Inktomi, Convera, Verity, and Stratify. Many traditional vendors who provide the software of content management and document management have also the competence to handle unstructured information. The Interwoven's MetaTagger product automates taxonomy creation and tagging through the technology from Metacode. Vignette is preparing to issue a new adapter to manage unstructured content according to the web site of Vignette.

A host of smaller software vendors, including ClearForest, InStranet, Insigxt, and LingoMotors, are engaged in promoting creative methods and technologies to manage unstructured content. Insigxt has built a browser-based model integrating taxonomy management, classification, information extraction, data visualization, and interactive meta-data. InStranet provides a set of tools, known as classification technologies, which use XML, XSL and HTTP to integrate InStranet with other infrastructure platforms, such as Web content management. These developments are of tremendous value in improving business processes, increasing efficiency, reducing redundancy, and speeding products to market. Hence, if we can structure data and organize it more effectively, it will make an enormous contribution to the effectiveness of a company. Moore (2002) pointed out that the whole issue of structuring data so as to make it intelligible and useful, is a major issue for companies.

Using a search engine is an effective approach to discovering and indexing documents which contain specific terms. The content management system can manage effectively many kinds of content, provide access and version control, both of which are important aspects of information management. Information portal offers an ideal platform on which staffs can find and manage unstructured information, through using a content category and the ability to automate some aspects of the content management process.

2.2 Unstructured Information Management Architecture (UIMA)

The Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004a) is an architecture and framework that helps you to turn unstructured information into structured knowledge. UIMA is an industrial and scalable integrated platform which includes analysis engines and a back-end information processing system. The aim of UIMA is to develop a large scale solution and to use the appropriate technologies to process unstructured information.

2.2.1 Architecture of UIM

UIMA focuses on providing the conceptual foundation and component frames to support discovery, development, composition, and deployment of the ability to analyze unstructured information and integrate it with structured information sources. Delivery of the information to the application of the end users is open-ended and typically very specialized. Consequently, although the delivery phase is a critical aspect of a UIM application, it is not directly addressed by the architecture (Ferrucci and Lally, 2004b).

A UIM application can be thought of as comprising two phases (Figure 2.1): analysis and delivery. In the analysis phase, collections of documents are acquired and analyzed. The results are stored in one or more forms as needed for the delivery phase. In the delivery phase, the results of analysis may be accessed by the application user, possibly together with the original documents or other structured information, through application-appropriate access methods and interface (Ferrucci and Lally, 2004a).

The UIMA and software framework are continuously being researched at IBM Research, in light of requirements coming from many different areas including:

- text and multi-modal analytics,
- machine translation systems,
- transcription systems,
- high throughput analysis systems,
- knowledge integration,
- semantic searching,
- program analysis,
- social network analysis,
- question answering,
- change detection,
- security and
- semantic web applications.

2.2.2 Components of UIMA

An Unstructured Information Management (UIM) application is generally characterized as a software system that analyzes large volumes of unstructured information in order to discover, organize, and deliver relevant knowledge to the end user (Ferrucci and Lally, 2004b).

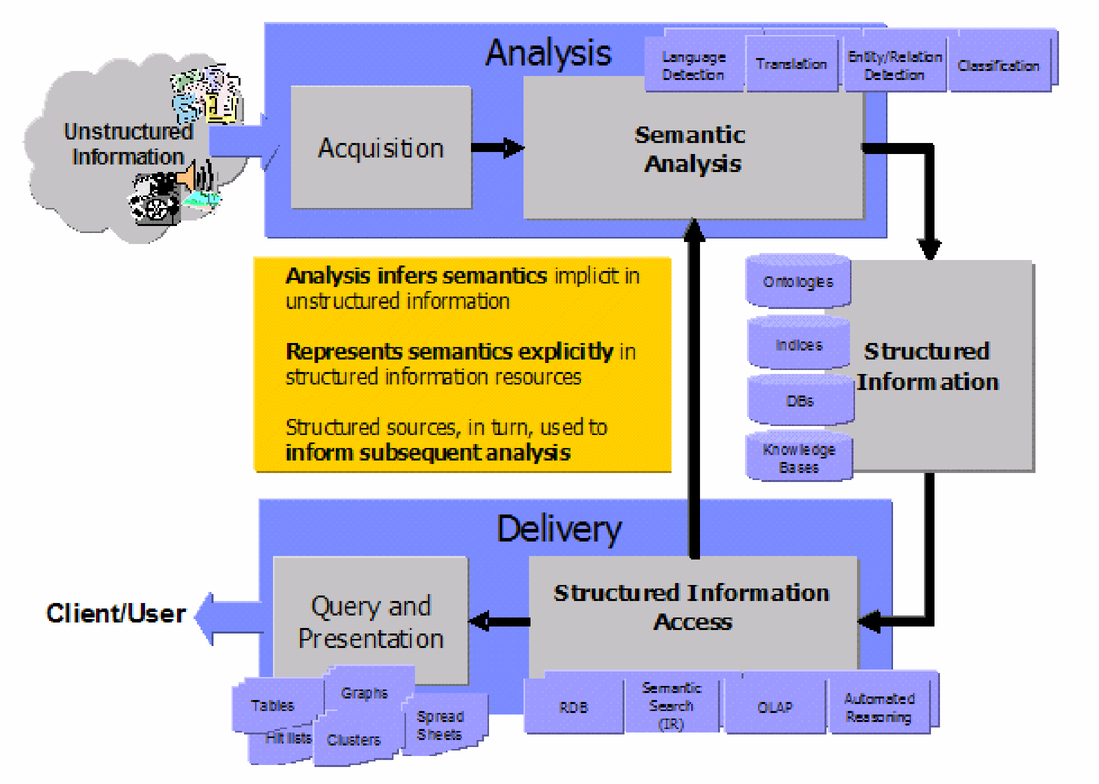


Figure 2.1 The architecture of unstructured information management (Ferrucci and Lally, 2004a).

UIMA is a framework for the development of analysis engines. It provides the component which analyzes the unstructured information stream, such as HTML Web pages. These components can be implemented in a range of situations from the lightweight to those at the top end of the scale. The UIMA annotation format is called CAS (Common Analysis Structure). It is mainly based on the TIPSTER format (Grishman, 1997). Annotations in the CAS are stand-off, for the sake of flexibility. Documents can be processed either at a single document level or at a collection level. Collection in UIMA is handled by the Collection Processing Engine, which has some interesting features such as filtering, performance monitoring and parallelization.

At the heart of UIMA, there is a common representation system, which is called the Common Analysis Structure (CAS). The CAS is used to provide analysis engines with read access to the artifact being analyzed (e.g., document, image, video, and so on) and read/write access to the analysis results or annotations associated with defined regions of the artifact. Regions may correspond to words, sentences or paragraphs in text or frames or parts of frames in a video, for example. The CAS is shared among analysis engines working in concert as part of a larger workflow to process a collection of artifacts; it is passed from one analysis engine to the next, in a flow.

UIMA has many features in common with other software architectures for language engineering such as GATE (Cunningham et al., 2000; Bontcheva et al., 2002) and ATLAS (Laprun et al., 2002). Each of these systems isolates the core algorithms which perform language processing from system services such as data storing, communication between components, and visualization of results. However, the emphasis of UIMA is transferring UIM technologies to products and has led to a richer architecture which allows integrating applications with a host of enterprise products (e.g., WebSphere* Portal Server, Lotus* Workplace) and a variety of middleware and platform options.

UIMA has been used as a platform for IBM's video analysis and search system MARVEL and for a project that acquires, converts, translates, and indexes video news channels, called Tales.

2.3 Knowledge elicitation and relevant technologies

2.3.1 Aims of knowledge elicitation

Knowledge is embedded in organizations in a variety of sources and formats. Knowledge elicitation is a process of extracting information through in-depth interviews and observation, and through the use of some information technologies (Klein, 1999). Nordlander (2005) claimed that knowledge elicitation is a subset of knowledge acquisition that specially refers to retrieving knowledge from human experts through a range of strategies. According to Schreiber et al. (1999), knowledge elicitation consists of a set of technologies and methods which try to elicit domain knowledge through interaction with experts. Most critically, knowledge elicitation is a process of eliciting tacit knowledge that is, bringing out the knowledge present in the conscious and sub-conscious minds, or of assisting the expert to recall and redefine their rules of thumb, work practices, processes, etc. with the help of a knowledge engineer (Morecroft, 1994).

There are many ways of grouping KE methods. One is to group them by the type of interaction with the domain expert. This includes interview, case study, storytelling, observation, prototyping, and etc. According to Cooke (1999), there are many different elicitation technologies. It is important to select the right technologies for each specific situation. Some technologies of artificial intelligence implementations are adopted to analyze unstructured information and elicit knowledge, such as natural language processing, text mining, and information retrieving, etc.

2.3.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a domain of computer science concerned with the interaction between computers and human (natural) languages. Natural language processing gives machines the ability to read and understand the languages that the human beings speak (Luger and Studbblefield, 2004). A natural language generation system is responsible for transforming information from computer databases into readable human language. A natural language understanding system converts samples of human language into more formal representations that are easier for computer programs to manipulate.

Generally, managing unstructured information belongs to the domain of natural language processing. The technologies of NLP include information retrieval, machine learning, ontologies, semantic networks, text mining, reasoning, etc. Basically, these technologies are divided into two categories: those that are statistical and those that are rule-based (Ferrucci and Lally, 2004a; Stensmo, 2003). Some enterprises build content databases and enterprise portals to organize unstructured documents based on integrating technologies for managing unstructured content. A number of rule-based, linguistic, statistical, machine-learning, and hybrid approaches have been developed to mark up terms in text documents. Some other approaches include referring to knowledge sources. Many researchers hope that one day a great powerful natural language processing system will be able to acquire knowledge on its own through reading the text over the Internet. Some applications of natural language processing include information retrieval, text mining, and machine translation (Russell and Norvig, 2003).

Natural language understanding is sometimes referred to as an AI-complete problem, because natural language recognition needs extensive knowledge about the outside world. The definition of “Understanding” is one of the main problems in natural language processing. Hence, NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish the human language (e.g. Chinese, English, and so on) from computer language (such as C++, Java or LISP). Although NLP may include text and speech, processing speech has been evolved into a separate field.

Statistical natural language processing uses stochastic, probabilistic and statistical methods to resolve some of the difficulties discussed above, especially those which arise because longer sentences are highly ambiguous when processed with realistic grammars, yielding thousands or millions of possible analyses. Methods for disambiguation often involve the use of corpora and Markov models. Statistical NLP comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra (Christopher and Schutze, 1999). Statistic-based NLP technologies mainly derive from machine learning and data mining, both of which are belonged to the field of artificial intelligence which learns from data.

Information Retrieval (IR) is concerned with storing, searching and retrieving information. It is a separate field within computer science, but IR relies on some

NLP methods. Some current research and applications seek to bridge the gap between IR and NLP.

2.3.3 Semantic network and semantic web

The main goal of the Semantic Web is to be able to express the meaning of resources that can be found on the Web (Fensel et al., 2003). In order to achieve that objective, Berners-Lee (2000) designed a representational structure with several layers in Figure 2.2.

These layers are: the XML (eXtensible Markup Language from http://www.w3.org/XML) layer represents the structure of data, the RDF (Resource Definition Framework from http://www.w3.org/RDF) layer represents the meaning of data, the ontology layer represents the formal common agreement about the meaning of data, the logic layer enables intelligent reasoning with meaningful data. Berners-Lee (2000) proved that one layer supports the exchange of proofs in an inter-agent communication and enabling common understanding of how the desired information is derived.

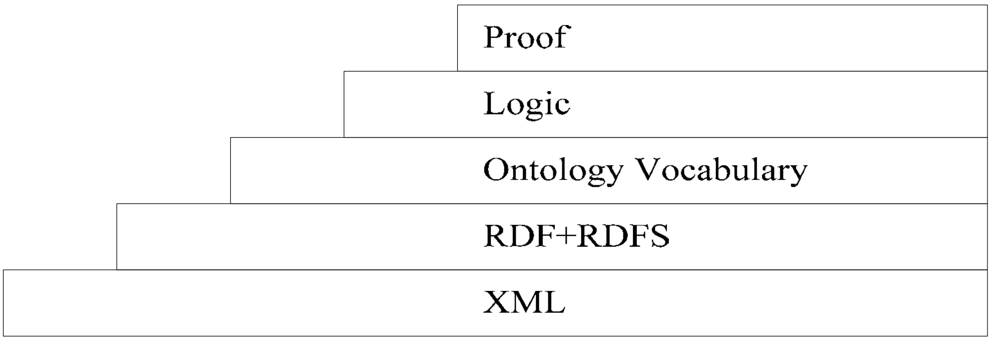


Figure 2.2 Layers of the semantic web architecture (Berners-Lee, 2000).

Ontologies are well-suited for describing the heterogeneous, distributed, semi-structured information sources (e.g., XML documents) that can be found on the Web or on intranets. By defining the shared and common domain theories, ontologies help both people and machines to communicate concisely by supporting the exchange of semantics rather than just syntax. It is therefore important that any semantics for the Web be based upon an explicitly specified ontology.

It is worth noting that the real power of semantic networks will be implemented when many systems have been built. These applications include: collecting content from various sources, integrating and processing information, and exchanging the results with other human or machine agents. Thus, the effectiveness of the Semantic Web will be increased greatly when more machine readable Web content and more automatic services can be offered.

Two important technologies for developing a Semantic Web are already in place, namely XML and RDF. XML lets users create their own tags to annotate Web pages or sections of text on a page. The System can use the tags in sophisticated ways, but to do so, a system programmer must know the meaning of every tag and what the author wants. In other words, XML allows users to add an arbitrary structure to their documents, but says nothing about what the structures mean (Erdmann and Studer, 2000). Typically, the meaning of XML documents is intuitively clear to humans because the semantic markup and tags use terms that are common in the particular domain. However, computers do not have intuition, and the name of a tag does not offer the semantics. Basically, XML lacks a semantic model. It has only a surface model, or tree. So, XML can only play the role of a transport mechanism that provides a readily machine-processable data format.

RDF (from http://www.w3.org/XML/Schema) provides a means for adding semantics to a document. RDF is an infrastructure that enables encoding, exchanging, and reusing structured meta-data. Information is stored in the form of RDF statements, which are machine-understandable. Search engines, intelligent agents, information brokers, browsers, and human users can understand and use that semantic information. RDF is implementation independent and may be serialized in XML (i.e., its syntax is defined in XML). Adding semantic information to Web documents is called semantic annotation (Handschuh et al., 2001). RDF in combination with RDFS (Resource Description Framework Schema from http://www.w3.org/TR/PR-rdf-schema) offers modeling primitives that can be extended to meet the needs of a specific situation. Basic class hierarchies and relations between classes and objects are expressible in RDFS. In general, however, RDFS suffers from a lack of formal semantics for its modeling-primitives, making proper interpretation an error-prone process.

2.3.4 Text mining and Information Extraction (IE)

Text mining is a technology that makes it possible to discover patterns and trends semi-automatically from huge collections of unstructured text (Hearst, 1999; Swanson, 1990; Brusic and Jeleznikow, 1999; Swanson and Smalheiser, 1997). Agrawal and Srikant (1994) thought that text mining was based on the technologies such as natural language processing, information retrieval, information extraction, and data mining. Text mining has proved to be a promising approach for knowledge discovery from text sources.

Text mining is a procedure for discovering knowledge from unstructured text. The steps of text mining can be classified into the following components:

- Gathering of text documents (automated/manual extraction from web sources)
- Text preprocessing (semi-structuring the text by using databases, XML, etc.)
- Natural language processing (entity tagging or labeling, term identification)
- Text categorization (classification or clustering)
- Visualization (interface, graphical representation)
- Analysis (evaluation of extracted information)

Each of the above steps is broad research domains in themselves. The process of text mining needs an efficient integration of these steps for knowledge discovery.

Human languages are diverse and irregular, but most people have learned to understand at least one, and computers can do the same. Text mining software applies linguistic analysis and pattern recognition to identify concepts, terms and entities, for example, names and email addresses. Computers can not create structures, but can extract terms and concepts by applying linguistic models to documents.

Some text- and image-mining software identifies and then tags or extracts concepts, entities, terms or patterns. Some use extracted concepts or patterns to create taxonomy or classification systems to categorize documents. Some apply taxonomies to automate document processing. Some map inter-document links and derive predictive models. These approaches can solve business problems; automating e-mail handling, tuning customer service procedures based on call-center conversations, sifting through medical literature to discover and map disease patterns.

Several text engineering architectures have been proposed to manage text processing over the last decade (Cunningham et al., 2000). General Architecture for Text Engineering (GATE) (Bontcheva et al., 2004) has been essentially designed for information extraction tasks. It aims at reusing NLP tools in built-in components. The interchange annotation format (CPSL – Common Pattern Specific Language) is based on the TIPSTER annotation format (Grishman, 1997).

Based on an external linguistic annotation platform, namely GATE, the KIM platform (Popov and Kiryakov, 2003) can be considered as a "meta-platform". KIM is an architecture based on ontologies, a semantic index, and information retrieval.

2.3.5 Information Retrieval (IR)

The idea of using computers to search for information was proposed by Vannevar Bush (Amit, 2001). In brief, information retrieval (IR) is the science of searching for documents, for information within documents and for meta-data about documents, as well as that of searching relational databases and the World Wide Web. Essentially, information retrieval is the academic discipline which underlies computer-based text search tools. It tends to concentrate on mathematical models and algorithms for retrieval quality.

In 1957, Luhn (1957) was the first to propose using words as indexing units for documents and measuring word overlap as the criterion for retrieval. Several key developments in the field happened in the 1960s. The SMART system developed by Salton (1971) provided researchers with the tools for experimenting with ideas to improve search quality. Cleverdon (1967) designed the Cranfield evaluation methodology for retrieval systems. A large text collection was provided by Text REtrieval Conference (TREC) which is series of evaluation conferences sponsored by various US Government agencies under the auspices of NIST aiming at encouraging research in IR from large text collections (Harman, 1993). And then, many old techniques were modified, and many new techniques were developed to do effective retrieval using large collections (Amit, 2001).

Early IR systems were Boolean systems which allowed users to specify their information needs using a complex combination of Boolean ANDs, ORs and NOTs. But there is no inherent notion of document ranking, and it is very hard for a user to form a good search request using IR systems. Even though Boolean systems usually return matching documents in some order, e.g., ordered by date, or some other document feature, relevance ranking is often not critical in a Boolean system. Considering the leading web search engines: one of their biggest shortcomings is that they cannot understand the meaning of natural language without using linguistic and statistical analysis and cannot provide the hits the user really wants.

Most IR systems rank documents through estimating a numeric score for every document. Some models have been developed. For example, the vector space model (Salton et al., 1975) regards text as a vector of terms in a high dimensional space. To assign a numeric score to a document for a query, the vector space model measures the similarity between the query vector and the document vector. Typically, the angle between vectors is used as a measure of divergence between the vectors, and the cosine of the angle is used as the numeric similarity. Another kind of model is the probabilistic model. Probabilistic models are based on the general principle that documents in a collection should be ranked by decreasing probability of their relevance to a query (Robertson, 1977). Each probability model is based on a different probability estimation technique.

Nowadays, the World Wide Web is being developed rapidly and on a huge scale. The algorithms in IR are further applied for searching on the Internet. In some sense, it can be said that he mass of information on the Web is essentially useless unless this wealth of information can be discovered and consumed by other users (Christopher et al., 2008). Early attempts made web information available by full-text index search engines such as Altavista, Excite and Infoseek, and by taxonomies populated with web pages in categories, such as Yahoo! The former presented the user with a keyword search interface supported by inverted indexes and ranking mechanisms. The latter allowed the user to browse through a hierarchical tree of category labels. However, it is very difficult to accurately classify web pages into a taxonomy accurately because the size of the content on the Web is too large.

The web search engines are challenged when they try to index and search tens of millions of documents, which are much larger than anything prior information retrieval systems needed to handle (Henzinger et al., 2000). Indexing, query serving and ranking at this scale required the harnessing together of thousands of machines to create highly user-friendly systems such as Google. However, the quality and relevance of web search results left much to be desired owing to the idiosyncrasies of content creation on the Web (Sufyan, 2005). This needed the invention of new ranking and spam-fighting techniques in order to ensure the quality of the search results. While classical information retrieval techniques continue to be necessary for web search, they are not by any means sufficient. Whereas, a key aspect is that the classical techniques measure the relevance of a document to a query, there remains a need to gauge the authoritativeness of a document based on cues such as which website hosts it (Christopher et al., 2008).

2.3.6 Technologies of Artificial Intelligent (AI)

Artificial Intelligence (AI) refers to machine intelligence and the branch of computer science which aims to create it. John McCarthy, who coined the term AI in 1956, defined it as “the science and engineering of making intelligent machines” (McCorduck, 2004). Poole and Goebel (1998) claimed AI as the study and design of intelligent agents. An intelligent agent is a system that perceives its environment and takes actions which maximize its chances of success (Russell and Norvig, 2003).

AI research is highly technical and specialized, so much so that some critics decry the “fragmentation” of the field (Nilsson, 1998). The main problems of AI research include reasoning, knowledge, planning, learning, communicating, understanding, perception and the ability of manipulating objects. According to Kurzweil (2005), the general intelligence (or “strong AI”) is still a long term goal of some research.

Early AI researchers developed algorithms that imitated the step-by-step reasoning that human beings use when they solve puzzles, play board games or make logical deductions (Luger and Stubblefield, 2004). By the late 80s and 90s, AI research had also developed highly successful methods for dealing with uncertain or incomplete information, employing concepts from probability and economics (Russell and Norvig, 2003).

Artificial intelligence technologies might deliver core knowledge management activities like knowledge discovery (e.g. mining of interest profile, connecting people of common interest in organization), indexing & representation (i.e. the issues of re-organizing and retrieving knowledge), and knowledge fusion such as combining existing knowledge to generate new knowledge (Tsui, 2002).

To exploit and classify accurately knowledge in a specific domain, deep analysis is needed. This may depend on part-of-speech detection, grammatical parsing and named-entity recognition where proper names, organizations and locations are identified. Other engines may specialize in detecting events and times and then others work on detecting relationships between these elements. A variety of techniques may be used to develop these specialized engines including rule-based and statistical machine learning algorithms. Analysis engines may vary along a variety of dimensions including document modality (text, speech, and video), format, natural language, style, and domain. And they may make different performance tradeoffs, for example, precision against speed or recall against precision. The key point is to develop a complete solution which enables people to obtain useful knowledge from unstructured information by integrating a variety of independently developed analysis engines. These must be integrated to perform a comprehensive analysis task and then their results must be funneled into systems that allow users to rapidly find and exploit the discovered knowledge, for example, search, database and/or knowledge bases.

Enterprise content management systems are being adopted widely, and this provides access to unstructured data and the meta-data on top of it. Intelligent systems with semantic layers and taxonomies can connect to the unstructured data. From the semantic construction viewpoint, finding word variants and similar words as in search engine usage, is inadequate. To differentiate between the word “balance” as a verb or a noun is a feat of semantic and linguistic categorization. This approach requires taxonomies, ontologies, and a semantic layer to build concept and category relationships. Weglarz (2004) thought that if we have already identified the context and semantics of our unstructured data, we can bring this information together with our structured data, bridging the two worlds and ultimately providing greater business insight.

The task of text classification is defined as assigning category labels to new documents based on their contents (Remeikis et al., 2004). Several text categorization approaches have been proposed such as neural networks, genetic algorithms and probabilistic models, and support vector machines (Buddeewong and Kreesuradej, 2005). Text classification is an important aspect of information visibility and control. A typical classification workflow might include the steps of: collecting, classifying, searching, reporting, and acting.

2.3.7 A summary of knowledge elicitation techniques

The technologies of knowledge elicitation are summarized in Table 2.2.

Table 2.2 Summary of knowledge elicitation techniques.

Concept/Technology	**Description**
Knowledge elicitation	Knowledge elicitation is a process of eliciting tacit knowledge and relevant technologies (Morecroft, 1994).Some technologies of artificial intelligence implementations are adopted to analyze unstructured information and elicit knowledge, such as natural language processing, text mining, information retrieving, etc. (Cooke, 1999).
Natural Language Processing (NLP)	NLP gives machines the ability to read and understand the natural languages (Luger and Studbblefield, 2004). Managing unstructured information belongs to the domains of NLP. The technologies are divided into two categories: statistical and semantic natural language processing (Ferrucci and Lally, 2004a; Stensmo, 2003).
Semantic Web	The main goal of the Semantic Web is to be able to express the meaning of resources that can be found on the Web (Fensel et al., 2003). Two important

	technologies for developing a Semantic Web are XML and RDF (Erdmann and Studer, 2000).
Text mining and Information Extracting (IE)	Text mining is a domain of technology that makes it possible to discover patterns and trends semi-automatically from huge collections of unstructured text (Hearst, 1999; Swanson, Brusic and Jeleznikow, 1999). Information extracting applies linguistic analysis and pattern recognition to identify concepts, terms and entities. The technology of IE is used to extract entities in text mining.
Information Retrieval (IR)	Information Retrieval (IR) is the science of searching for documents. IR is the academic discipline and tends to concentrate on mathematical models and algorithms for retrieval quality, such as Vector Space Machine (Salton et al., 1975) and Probabilistic Model (Robertson, 1997).
Artificial Intelligence (AI)	AI is the science and engineering of making intelligent machines (McCarthy, 1956). Many technologies are developed in the domain of AI, and these technologies might deliver core knowledge management activities like knowledge discovery, indexing & representation, and knowledge fusion

	(Tsui, 2002).

2.4 Knowledge representation and organization

2.4.1 Scope of knowledge representation

Knowledge representation (Luger and Stubblefield, 2004) and knowledge engineering (Poole et al., 1998) are central to AI research. Many of the problems which are expected to be solved by machines need extensive knowledge about the world. Among the things that AI needs to represent are: objects, properties, categories and relations between objects; situations, events, states and time; causes and effects; and many others. A complete representation of “what exists” is an ontology of which the most general are called upper ontologies (Luger and Stubblefield, 2004; Russell and Norvig, 2003; Nilsson, 1998).

According to Hodge (2000), knowledge representation is fundamentally a surrogate or a substitute for the knowledge itself, which is used to enable an entity to determine a sequence by reasoning about the world. Knowledge representations include a variety of solutions about organizing, managing, retrieving information. It appears in different kinds of forms such as databases, portals and libraries, and ranges from general classification schemas, for example from organizing books on a shelf, to taxonomies, semantic networks and even ontologies. There are various knowledge representation tools such as logic, rules, frames, and semantic networks. Semantic network allows people to define relations between objects. They were developed as a model for human memory and the relationships can be arbitrarily defined by knowledge workers (Quillian, 1968). Konev et al. (2005) thinks that inheritance is one of the main kinds of reasoning employed in semantic nets.

2.4.2 Knowledge model and modelling

Generally, knowledge engineers make use of some schemas to organize and represent knowledge after they acquire the information and knowledge from the knowledge experts. Such schemas are usually referred to as knowledge models. Schreiber et al. (1999) thought that the results of knowledge analysis are documented in a “knowledge model” which contains a specification of the information and knowledge structured in knowledge work.

A model is an intellectual construct in artifact form that provides an abstract, highly formalized, often visual, yet simplified representation of a phenomenon and its interactions (Coffey and Atkinson, 1996; Despres and Chauvel, 2000). Broadly, Satzinger et al. (2000) divided the models into three classes: mathematical models, descriptive models, and graphical models. Mathematical models explain the technical aspects of a system and can be either prescriptive or predictive (Miller, 2006). Descriptive models are in the form of narratives and often use symbolic or mathematic elements to aid understanding. Graphic models use diagrams and symbols to illustrate simple and complex relationships. The models are useful so long as the underlying assumptions are explicit, and it is recognized that they are an abstract representation of reality that may, or may not, be objective (McAdam and McCreedy, 1999).

Modelling is a way to structure objects and their relationships. It includes a set of methods, technologies, and activities. Knowledge Modeling is a cross disciplinary approach to acquiring and validating knowledge and storing knowledge for future use (Shadbolt, 2003). Knowledge Modeling packages combinations of data or information into a reusable format for the purpose of preserving, improving, sharing, aggregating and processing knowledge to simulate intelligence. Makhfi (2009) thought knowledge modeling offers a shift from local proprietary solutions to producing and disseminating embedded knowledge models into larger computational solutions in an effort to generate “applied knowledge.”

In nature, a model is a general concept. It contains entities or objects and their relationships. Thus, the things which contain knowledge entities or objects and describe their relationships can be thought of as knowledge models. Knowledge models are structured representations of knowledge using symbols to represent pieces of knowledge and the relationships between them (Schreiber et al., 1999). Basically, knowledge model structures must be able to represent knowledge so that it can be used for problem-solving. Nowadays, there are many knowledge models which are mainly descriptive and/or graphic models. The knowledge models play a key role in capturing and representing knowledge.

Today, various knowledge models have been suggested according to Sowa (2000), Baral (2003), Luger (2008), Russell and Norvig (2003). Present popular methods for knowledge representation in designing knowledge base systems (KBS) are predicate logic, semantic nets, frames, deductive rules and etc. Many new methods and techniques have also been discussed by Konar (2005), Jones (2008), Munakata (2008) and Rutkowski (2008). These authors describe methods using neural networks and fuzzy logic which are widely used in computational intelligence. Some methods are suitable for representing and processing semantics such as conceptual graphs in Chein and Mugnier (2009), Lehmann (2008), and van Harmelem and Bruce (2008).

2.4.3 Ontology and taxonomy

In philosophy, ontology is a theory about the nature of existence and, in particular, about what types of things can exist; ontology as a discipline studies such theories (Stojanovic et al., 2004). Ontology describes a formal and shared conceptualization of a particular domain of interest (Gruber, 1993). It provides a way of capturing a shared understanding of a domain that can be used both by humans and systems to aid in information exchange and integration.

Ontologies can facilitate interoperability between correlation engines by providing a shared understanding of the domain in question. Ontologies provide an effective means for explicating implicit design decisions and underlying assumptions at system build time. This makes it easier to reason about the intended meaning of the information interchanged between two systems. Hence, inter-operability is a key benefit of the application of ontologies, and many ontology-based approaches to information integration have been developed (Wache et al., 2001).

On the other hand, ontologies provide a formalization of shared understanding which allows machine processability. Machine processability in turn forms the basis for the next generation of the World Wide Web, the so-called Semantic Web, which is itself based on using ontologies for enhancing (annotating) content with formal semantics(Berners-Lee, 2000; Fensel et al., 2003). This definition will enable automatic agents to reason about Web content and to carry out intelligent tasks on behalf of the user.

Additionally, the explicit representation of the semantics of data through ontologies will enable correlation engines to provide a qualitatively new level of services, such as verification, justification, and gap analysis (Stojanovic et al., 2004). These engines will be able to construct a large network of human knowledge and will improve this capability with machine processability. It is important to note that ontologies not only define information, but also add expressiveness and reasoning capabilities.

Many languages for representing ontologies have been defined, including Ontology Interchange Language (OIL) from http://www.ontoknowledge.org/oil, DARPA Agent Markup Language (DAML) from http://www.daml.org/2001/03/ reference.html, and Ontology Web Language (OWL) from http://www.w3.org/TR/ owl0-ref. Meanwhile, the KArlsruhe Ontology (KAON) language is based on RDFS, but provides a clean separation of the modeling primitives from the ontology itself (Motik et al., 2002). KAON provides means for modeling meta-classes and incorporating several commonly used modeling primitives, such as transitive, symmetric, and inverse properties, as well as cardinalities (Stojanovic et al., 2004; Ferrucci and Lally, 2004a).

In current usage within “Knowledge Management”, taxonomies are seen as less broad than ontologies as ontologies apply a larger variety of relation types (Suryanto and Compton, 2000). Taxonomy is an effective methodology which provides a way of classifying and representing knowledge. Blake (2002) defines taxonomy as consisting of entities (objects or headings), relationships, links, grouping, tagging and navigation, and facilitating efficient searching, browsing, alerting, and effective content management. Basically, taxonomy can be seen as a complex structured knowledge beyond the knowledge repository (Conway and Sligar, 2002; Scott and Kreulen, 2002). The components of taxonomy include controlled vocabulary, meta-data, and classifications. Mathematically, a hierarchical taxonomy is a tree structure of classifications for a given set of objects. The progress of reasoning in this hierarchical taxonomy proceeds from the general to the more specific (Grossi et al., 2005). Linnaean taxonomy is the system most familiar to non-taxonomists. It uses the formal taxonomic ranks Kingdom, Phylum, Class, Order, Family, Genus, and Species. The lower ranks (superfamily to subspecies) are strictly regulated, whereas taxonomy at the higher ranks is a result of consensus in the scientific community. The taxonomy is exclusively based on cluster analysis and neighbor joining to best-fit numerical equations that characterize all measurable quantities of a number of things.

Since knowledge is growing hugely and rapidly, and is becoming more complex, dynamic taxonomies have been proposed as a model of knowledge management to describe and access complex and heterogeneous information and knowledge bases. Dynamic taxonomies provide more effect and efficient classifications for knowledge so as to enhance knowledge work on knowledge management. Dynamic classification enables users to view all possible categories of information together with the ability and tools to view, cross-correlate, mix and match the categories.

Users gain the freedom and creativity to decide for themselves how they would like to organize their knowledge space, classify information in that space, and apply the logic they have uniquely created to classify new, incoming information as it arrives (Sacco, 2000).

2.4.4 Knowledge map and concept map

Knowledge maps are created by transferring certain aspects of knowledge into a graphical format that is easily understandable (Kim et al., 2003). Developing a knowledge map includes locating important knowledge in the organization and then publishing the some sort of list or picture that shows where to find it.

To develop enterprise knowledge asset sharing, a knowledge map is used and provides the interactive platform related with knowledge-bases in real time for the different users. Due to the simplicity of the notation used and of its semantics (Gómez et al., 2000) and the visualization for user intuitionistic cognition (Gordon, 2000; Gordon and Edge, 1997), the studies related to knowledge maps have been carried out in recent years far and wide. Gómez et al. (2000) presented the knowledge map as an aid for addressing holistic testing, at the same time. This is a navigation map or index catalog of expertise knowledge (Davenport and Prusak, 1998) and is created to assess an organization’s knowledge position where cataloging of its existing intellectual resources is required (Zack, 1999). Woo et al. (2004) provided the concept of the dynamic knowledge map that was used as a web-based knowledge navigator when the tacit knowledge of experts needed to be reused as a tool which manipulates knowledge in knowledge management, a knowledge map displays the distribution of knowledge assets and the relationships among knowledge sources and users.

A knowledge map gives a useful blueprint for implementing a knowledge management system and summarizes the captured knowledge (Kim et al., 2003). The concrete knowledge objects, such as concept/attribute/value (Gómez et al., 2000) and social constructed knowledge (Woo et al., 2004), need to be expressed in the knowledge node of a knowledge map.

A concept map is a web diagram for detecting, collecting and sharing information. Concept maps identify the way we think, the way we see relationships among different concepts (Walker, 2002). Concept maps provide a way to represent information visually. Concept maps are diagrams showing the relationships among concepts. They are graphical tools for organizing and representing knowledge. Concept maps include concepts which are denoted by nodes and the relationships among concepts which are denoted by lines connecting two concepts. The technique of concept mapping was developed by Novak and Cañas, (2006). In particular, constructivists hold that prior knowledge is used as a framework to learn new knowledge.

The fundamentals of concept mapping are based on Ausubel learning theory (Novak, 1998) which in itself is based on the assumption that meaningful learning occurs when new concepts are linked to familiar concepts existing in the learner's cognitive structure and can be applied to all subject matter. The use of concept maps is becoming widespread in the domains of mathematics and of science education. Peled et al. (1993) suggested that the concept map should be recommended as a means of producing meaningful learning in the analysis of scientific articles as well as enhancing the integration of theory and practice. Khan (1993) thought concept maps were also an effective means of bridging the gap between conceptual and procedural knowledge. The use of concept maps enhances the people's conceptual understanding (Baldissera, 1993; Khan, 1993; Markham et al., 1993; Peled et al., 1993; Tveita, 1993; Vquez and Caraballo, 1993; von Minden and Nardi, 1993). Concept maps can be used to develop an understanding of knowledge, to explore new information, gather new knowledge and information, access knowledge and information, share knowledge and information, produce problem solving options from written documents, web sites, web searches, multimedia presentations, etc.

There are main four categories of concept maps. These are distinguished by their different formats for representing information (College of ACES, 2004). The Spider concept map is organized by placing the central theme or unifying factor in the center of the map. Hierarchy concept maps present information in a descending order of importance. The most important information is placed at the top. Distinguishing factors determine the placement of the information. Flowchart concept maps organize information in a linear format. System concept maps organize information in a format which is similar to a flowchart with the addition of "INPUTS" and "OUTPUTS".

2.4.5 Social Network Analysis (SNA)

Whereas concept maps link up the relationship between concepts, social network links up the relationship between people. The understanding of social networks is important as a lot of unstructured information is embedded in various forms of social network that are not shown in the formal organization charts. A social network is a social structure made of nodes which are generally individuals or organizations (Hill and Dunbar, 2002). Social Network Analysis (SNA) is based on an assumption of the importance of relationships among interacting units or nodes (Wasserman and Faust, 1994). SNA is an interdisciplinary methodology developed mainly by sociologists and researchers in social psychology, further developed in collaboration with mathematics, statistics, and computing. This led to a rapid development of formal analyzing techniques which made it an attractive tool for other disciplines like economics, marketing or industrial engineering (Scott, 2002). Social network analysis displays the knowledge networks and interaction patterns among group members, organizations, and other social entities. Krebs (1998) thought SNA is the mapping and measuring of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. These relations, defined by linkages among units/nodes, are a fundamental component of SNA (Scott, 2002; Wassermann and Faust, 1994). SNA provides both a visual and a mathematical analysis of human relationships. Cross et al. (2004) thought that network structure can facilitate or impede the effectiveness of knowledge workers. Generally, the success or failure of society and of an organization often depends on its inherent structure.

Podolny and Page (1998) have argued that this could be shown by the fact that the field is increasingly more interested in the outcomes of networking rather than the reasons for networking. Oliver and Ebers (1998) have shown that there are four research paradigms dominating the field of inter-organizational networking: social networks, power and control, institutional theory, and institutional economics and strategy. It is within the social network category that the use of network analysis has been the most important. Owen-Smith et al. (2002) compared the organization and structure of scientific research in the United States and Europe by building networks of R&D cooperation. Breschi and Lissoni (2003) expanded the study of Trajkenberg and Hendersson (1993) and found that social proximity has the stronger relevance for the degree of knowledge spillovers than geographical proximity.

Network analysis should be used to describe networks and to attempts to link these descriptions to network outcomes; but not to outcomes for specific actors located in networks (Doreian, 2001). Cross et al. (2004) pointed out that typical roles in organizational networks include central people, peripheral people, boundary spanners and knowledge brokers. The relationship between the centralities of all nodes can reveal much about the overall network structure (Krebs, 1998). In SNA, there are four most important concepts and measures about the analysis of network structures: network density, centrality, betweenness, and centralization. Meanwhile, there are four measuring factors concerned with the analysis of dynamic performance of a network: robustness, efficiency, effectiveness and diversity. These measures are often used in any network analysis and an understanding of them is the foundation for the comprehension of empirical work. Most of the definitions are adapted so that they use the terminology previously defined by Scott (2002) and Burt (1992).

The results of social network analysis can be used at the level of individuals, departments or organizations to identify teams and individuals playing central roles, identify isolated teams or individuals, and detecting information bottlenecks. Then, the opportunities to improve knowledge flow are found in order to accelerate the flow of knowledge and information across functional and organizational boundaries and improve the effectiveness of formal communication channels.

Social networks in an organization represent a complex system in which relationships are changing all the time, and you can never accurately predict the results of an intervention. Social network analysis is a tool that can be used, with discretion and sensitivity, on an ongoing basis in the context of continuous organizational improvement (Anklam, 2003).

2.4.6 Knowledge flow in enterprise

In an enterprise, knowledge flows through various social networks. A knowledge flow is the passing of knowledge between nodes according to certain rules or principles. Knowledge flow is the critical factor which enables the business processes (Newman, 2002). It is worth to note that understanding how knowledge flows through an enterprise is critical to improving the productivity of knowledge work. Nissen and Levitt (2002) thought that we are unlikely to design effective flow-enhancing interventions until we can understand the phenomenon of knowledge flow. Zhuge (2006) pointed out that a knowledge node in knowledge flow is a team member or role, or a knowledge portal or process, and a node can generate, learn, process, understand, synthesize, and deliver knowledge. The interaction of knowledge nodes or entities leads to knowledge networks. Knowledge networks are the basic components of most social relationships, ranging from families and friendships to companies, professional organizations, and governments. Newman (2002) thought knowledge flows were sequences of transformations performed by agents on knowledge artifacts in support of specific actions or decisions in knowledge network.

Since Nonaka (1994) built the SECI model to describe patterns of knowledge flow, a number of theoretical models have been developed to describe various aspects of the knowledge flow phenomenon (Augier et al., 2001; Dixon, 2000; King and Ko, 2001; O'Leary, 2001; Schultze and Boland, 2000; Swap et al., 2001). On the other hand, optimizing and reengineering of the knowledge flow enable organizations and individuals to be more creative, innovative, and responsive to change. Schutt (2003) derived 11 factors that lead to the optimization of the work of a knowledge worker. By making the communication flow transparent, business processes can be made more efficient, allowing organizations to make better use of people by freeing them from being buried in conventional multilayer hierarchies and inefficient business processes (Malone, 2004).

Improving enterprise performance mainly depends upon the rapid and efficient transfer from one organization, location or time of application, to another. From the technical viewpoint, the dynamic dependence points immediately to the design of Information Systems (IS) – along with corresponding organization and process characteristics (Leavitt, 1965; Davenport, 1993) – to enhance knowledge flow. But knowledge is distinct from information and data, for example, it enables direct and appropriate action（Davenport et al., 1998; Teece, 1998). A few extant IS even address knowledge as the focus or object of flow, according to Nissen (1999).

Indeed in this light, the IS field does not have the benefit of a strong theory of knowledge flow, as Alavi and Leidner (2001) noted, that there exist large gaps in the body of knowledge in this area. So how does knowledge flow through the modern enterprise, and what kind of management, for example, IS development, training, organizational change, workflow reconfiguration, can be made to enhance the flow of knowledge? A number of theoretical models has been developed to describe various aspects of the knowledge-flow phenomenon (Augier et al., 2001; Dixon, 2000, King and Ko, 2001; Perkmann, 2002; O'Leary, 2001; Schultze and Boland, 2000; Swap et al., 2001), but few provide insight into the phenomenon itself; that is, there is a paucity of models that has been developed to describe how knowledge flows through the enterprise.

Von Hippel (1994) took a notable step in the direction of understanding how knowledge flows, as he examined causal factors for the relative marginal costs—characterized by the term stickiness—associated with transferring tacit and explicit knowledge for technical-innovation problem solving. Szulanski (1996, 2000) went further by tying his "stickiness" notion to four different stages of the knowledge-transfer process (i.e., initiation, implementation, ramp-up, integration). Numerous life cycle models (Nissen et al., 2000) adopt a similar staged view of knowledge flow. Nonaka (1994) went further still, as he introduced a model describing a "spiral" of dynamic interaction between tacit and explicit knowledge along an epistemological dimension, and he characterizes four processes (i.e., socialization, externalization, combination, integration) that enable individual knowledge to be "amplified" and effect organizational knowledge "crystallization" along the ontological dimension. A later and related work (Nonaka et al., 1996) identifies enabling "triggers" for and provides additional workplace examples of each knowledge-flow process (e.g., the "trigger" for socialization is building a "field" of interaction).

Building upon these theoretical steps, Nissen (2002) integrated and extended the research above to develop a phenomenological model of enterprise knowledge flow. This model makes the time of the knowledge flow explicit and supports a multidimensional representational framework that enables a new approach to analysis and visualization of diverse knowledge flow patterns in the enterprise. However, throughout this research, important dynamic interactions between model elements remain obscured through descriptive models based upon natural language texts and figures.

2.4.7 A summary of technologies of knowledge representation

The technologies of knowledge representation are summarized in Table 2.3.

Table 2.3 Summary of technologies of knowledge representation.

Concept/Technology	**Description**
Knowledge representation	Knowledge representation is fundamentally a surrogate or substitute for the knowledge itself, which is used to enable an entity to determine a sequence by reasoning about the world (Hodge, 2000). Knowledge

	representations include a variety of solutions about organizing, managing, and retrieving information. There are various knowledge representation tools such as logic, rules, frames and semantic networks.
Knowledge modeling and knowledge models	Knowledge modeling is a cross disciplinary approach to acquiring and validating knowledge and storing knowledge for future use (Shadbolt, 2003). Knowledge models are structured representations of knowledge using symbols to represent pieces of knowledge and the relationships between them (Schreiber et al., 1999). Knowledge models play a key role in capturing and representing knowledge.
Ontology and Taxonomy	Ontology describes a formal and shared conceptualization of a particular domain of interest (Gruber, 1993). It provides a way of capturing a shared understanding of a domain that can be used both by humans and systems to aid in information exchange and integration. Taxonomy can be seen as a complex structured knowledge beyond the knowledge repository (Conway and Sligar, 2002; Scott and Kreulen, 2002). Taxonomy includes entities, relationships, links, grouping, tagging and navigation, and facilitating efficient searching, browsing, alerting,

	and content management.
Knowledge map and concept map	Knowledge maps are created by transferring certain aspects of knowledge into a graphical format that is easily understandable (Kim et al., 2003). A concept map is a web diagram for detecting, collecting and sharing information and identifies the relationships among concepts (Walker, 2002).
Social Network Analysis (SNA)	A social network is a social structure made of nodes which are generally individuals or organizations (Hill and Dunbar, 2002). Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other entities (Krebs, 1998). SNA provides both a visual and a mathematical analysis of human relationships. Cross et al. (2004) pointed out that typical roles in organizational networks include central people, peripheral people, boundary spanners and knowledge brokers. In SNA, there are four most important concepts and four measuring factors concerned with the analysis of dynamic performance of a network.
Knowledge flow	Knowledge flows are sequences of transformations performed by agents on knowledge artifacts in support

	of specific actions or decisions in knowledge network (Newman, 2002). A number of theoretical models have been developed to describe various aspects of the knowledge flow phenomenon (Augier et al., 2001; Dixon, 2000; King and Ko, 2001).

2.5 Application and development of UIM

In analysis of unstructured content, Unstructured Information Management (UIM) applies various technologies, including statistics- and rule-based natural language processing, information retrieving, machine learning, ontologies, reasoning, etc. UIM applications may consult structured sources to help resolve the semantics of the unstructured content (Ferrucci and Lally, 2004b). For example, a database of chemical names can help in focusing the analysis of medical abstracts. A database of terrorist organizations and their locations can help in analyzing documents for terror-related activities. A UIM application generally produces structured information resources that unambiguously represent content derived from unstructured information input. These structured resources are made accessible to the end user through a set of application-appropriate access methods. A simple example is a search index and query processor that makes documents quickly accessible by topic and ranks them according to their relevance to key concepts specified by the user. A more complex example is a formal ontology and inference system that, for example, allows the user to explore the concepts, their relationships, and the logical implications contained in a collection consisting of millions of documents.

There are many industrial applications of UIM. Examples of applications of include:

- Shaving off just seconds per call to find the right technical documentation in call-centers can save millions.
- Rapidly detecting emerging trends in problem-reports coming in from all over the globe can avoid recalls and save companies and their customers millions if not billions.
- Detecting otherwise unrealized drug interactions through analyzing the linkages in of medical abstracts can help prevent disaster as well as help discover new drugs or cures.
- Analyzing communications linked to terrorist networks in the form of multi-lingual text or other modalities can help uncover plots threatening national security before they happen.
- Analyzing SEC reports to help evaluate corporate financial positions
- and other applications.

Applications like these, which rely on the rapid discovery of vital knowledge, require the analysis of unstructured information. This is all the information that has NOT been carefully encoded in enterprise databases but rather exists as natural language text, speech or video. An example is an application that processes millions of medical abstracts to discover critical drug interactions. Another example is an application that processes tens of millions of documents to discover evidence of probable terrorist activities (Roush, 2003).

There are two application areas which must process a large amount of scattered information but have been relatively less studied: emergency management and enterprise knowledge audit. Emergency management or disaster management is the discipline of dealing with and avoiding risks (Haddow and Bullock, 2004). It is a discipline that involves preparing for disaster before it occurs, disaster response (e.g. emergency evacuation, quarantine, mass decontamination, etc.), and supporting, and rebuilding society after natural or human-made disasters have occurred.

Emergency management software is the software used by local, state and federal emergency management personnel to deal with a wide range of disasters (including natural or human-made hazards) and can take many forms. Training software such as simulators are often used to help prepare first responders, word processors can keep form templates handy for printing and analytical software can be used to perform post-hoc examinations of the data captured during an incident. All of these systems are interrelated, so that the results of the after-incident analysis can then be used to program the training software to better prepare for a similar situation in the future. Crisis Information Management Software (CIMS) is the software found in emergency management operation centers (EOC) that supports the management of crisis information and the corresponding response by public safety agencies (Ashcroft, 2001).

Although emergency management software had existed prior to the 9/11 incidence in New York, after 2001 there was a significant increase in focus on emergency management. A 2001 study by the National Institute of Justice (NIJ) compared software features from 10 vendors. In 2004, the Institute for Security

Technology Studies published a report addressing the interoperability of different software, which has remained a strong focus in the development of software for the Emergency Management field. To support National Incident Management System implementation, the Department of Homeland Security established the National Incident Management System Support Center (NIMS SC) and the Supporting Technology Evaluation Program (STEP) in 2005. In 2007 a study similar to the National Institute of Justice report was conducted by the United States Air Force (USAF) (Robillard and etc., 2007). In 2008 the United States Air Force and the University of Colorado Center for Homeland Security surveyed several hundred Emergency Management personnel hoping to prioritize user requirements (Robillard, 2008).

Common features of the software include Geographic Information Systems (GIS), weather and plume modeling, resource management, and Command, Control, and Communication (C3) functions. The Federal Emergency Management Agency (FEMA) supports evaluation of software through the National Incident Management System Supporting Technology Evaluation Program (NIMS STEP). The National Preparedness Directorate Incident Management Systems Integration Division (NPD-IMSI) identifies criteria for this program to evaluate against. These criteria are derived primarily from the National Incident Management System according to Federal Emergency Management Agency (2009).

To handle the emergencies, some functional departments have to collect the relevant knowledge to advise or make timely decisions. In reality, the knowledge they want are scattered in many places including law documents, regulations, emergency preparedness, principles and dictionaries, research papers and the experience of experts. Such documents are often prepared for different purposes and located in unpredictable sources. In emergency management, decision-makers are confronted with an explosive amount of information that is disseminated among different authorities, external sources (such press media and web), and other people within a short period of time. Management of information and knowledge has become an increasingly important part of emergency management activity due to increasing community reliance on complex and sometimes vulnerable technological systems and infrastructure, and the need for emergency management staff to apply technological systems to manage risk more effectively.

A knowledge audit (K-Audit) is a systematic examination and evaluation of organizational knowledge assets. It is usually recommended in industries as an important first step, prior to the launching of any knowledge management programme (Choy et al., 2004; Liebowitz et al., 1999; 2003). People view the K-Audit as being the business needs assessment, cultural assessment, and an examination of what knowledge is needed, available, missing, applied, and contained. The knowledge audit is the critical stage in the KM programme (Liebowitz et al., 1999) providing, accurate identification, quantification, measurement and assessment of the sum total of tacit and explicit knowledge in the organization. However, the most a knowledge audit can do is to handle the structured part. There has been no research report on how the major part of knowledge that resides in an unstructured form can be audited and known.

2.6 Issues of unstructured information management

Although there are many concepts and technologies in unstructured information management, some limitations are worth noted in the research work in the thesis.

Morris (2008) thought that the principal challenge with unstructured information is that it needs to be analyzed in order to identify, locate and relate the entities and relationships of interest, and to discover the vital knowledge contained therein.

One of the shortcomings of the leading web search engines that they cannot understand the meaning of natural language without using linguistic and statistical analysis and cannot provide the hits the user really wants.

There have been no cases reported on how to apply UIM in managing dynamic knowledge and knowledge flow.

In this thesis, the concepts relationship mapping, dynamic taxonomy and the multi-faceted knowledge elicitation approach are introduced for the first time in the managing of unstructured information in emergency management and knowledge audit. Two respective prototypes are built.

Chapter 3. Research Methodology

The aim of this thesis is to develop an automatic system to manage unstructured information. A Multi-faceted and Automatic Knowledge Elicitation System (MAKES) is proposed to be capable of acquiring and analyzing knowledge from a great amount of unstructured information automatically. It will then display the organizational knowledge assets in the form of a visualized knowledge map.

3.1 Research method

3.1.1 Comparing between the conventional and MAKES approach for managing unstructured information

There are some differences between the conventional approach and the MAKES approach for unstructured information management. These differences enable to form the particular research approach of MAKES.

One of the major differences between the conventional approach and the MAKES approach proposed in this thesis for managing unstructured information is that the latter can provide a larger coverage of concepts. Most unstructured information can be classified by different criteria according to the author, date created, subject field, format, or language, etc. Multi-dimensional taxonomies are used for the treatment of different classification criteria. Various classification scenarios can be worked on independently or corporately.

Another major difference is that MAKES is supported by Artificial Intelligence (AI) technologies, so as to achieve automatic classification, intelligent searching and navigation, personalization and self-maintenance. An automatic classification algorithm is used to classify the new explicit knowledge created in an organization, according to different predefined rules under appropriate concepts. The advantage of MAKES is that it can reduce dramatically the amount of time and human effort that is spent in the classifying of a huge number of knowledge items.

Meanwhile, most traditional search engines provide a list of search results which may show many thousands of search records without any organization or classification. This is not user friendly enough for the knowledge workers to locate the right item from a large number of records. In MAKES, a multi-faceted taxonomy map is used for the classification of the search results based on the searching keywords. A personalized taxonomy can also be generated based on the behaviors or requirements of the user. A simple example is that the CEO may have authorization to see the whole taxonomy of the organization. However, the staff in the marketing department may be able to access only the taxonomy of the marketing department and have no access rights to any other department's taxonomy.

A traditional taxonomy is fixed and static after it has been built, and requires human intervention for any subsequent changes. The construction of knowledge models is time consuming and expensive. Moreover, the elicitation and interpretation of knowledge models relies on implicit human reasoning which can not be done by computer. In the MAKES approach, knowledge elicitation is accomplished by a new

AI-based knowledge elicitation and mining algorithm named the Concept Relationship Exploring Technique (CRET) as proposed in this thesis.

MAKES also allows the self-maintenance of the taxonomy and concepts which is not available in the conventional approach to manage unstructured information. New knowledge are analyzed and updated continuously and they may become new categories or abstractions in the taxonomy. Table 3.1 provides a summary of the differences between conventional taxonomy and multi-faceted taxonomy.

Table 3.1 Summary of differences between conventional and MAKES approaches for managing unstructured information.

Characteristics	**Conventional**	**MAKES**
Dimension of Categorization of knowledge	Single dimension	Multi-dimension
Knowledge representation	Only one knowledge entity	Several different knowledge entities at any levels of abstraction
Taxonomy structure	Static (unchangeable)	Dynamic (changeable)
AI support	No	Yes
Automatic classification	No	Yes
Intelligent searching and navigation	No	Yes

Personalization of taxonomy	No	Yes
Automatic Knowledge Elicitation	No	Yes
Self-maintenance of taxonomy	No	Yes

3.1.2 Case study approach

There are many research methods, e.g. surveys, ethnographies, experiments, quasi-experiments, economic and statistical modeling, histories, research syntheses, and developmental methods. The overall idea is that different research methods can and often do serve complementary functions.

Case study research excels at bringing us to an understanding of a complex issue or object and can extend experience or add strength to what is already known through previous research. Case studies emphasize detailed contextual analysis of a limited number of events or conditions and their relationships. Researchers have used the case study research method for many years across a variety of disciplines. Social scientists, in particular, have made wide use of this qualitative research method to examine contemporary real-life situations and provide the basis for the application of ideas and extension of methods. Yin (2003) defines the case study research method as an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used.

Critics of the case study method believe that the study of a small number of cases can offer no grounds for establishing reliability or generality of findings. Others feel that the intense exposure to study of the case biases the findings. Some dismiss case study research as useful only as an exploratory tool. Yet researchers continue to use the case study research method with success in carefully planned and crafted studies of real-life situations, issues, and problems. Reports on case studies from many disciplines are widely available in the literature.

The case study approach can be applied in at least two situations (Yin, 2004). The first type of applications is where the case study approach can be employed in either a descriptive question (what happened?) or an explanatory question (how or why did something happen?) (Shavelson and Townes, 2002). The second type of applications is used to make direct observations and collect data in natural settings, compared to relying on “derived” data (Bromley, 1986).

The target of this thesis is to research the approach of effective management of unstructured information through developing MAKES. The functions of MAKES focus on organizing, representing, retrieving, storing unstructured information and the output are mainly useful information and knowledge. However, the usage and effectiveness of MAKES is difficult to be evaluated and tested through theoretical reasoning or inference alone. In this thesis, the effectiveness of output of information and knowledge from MAKES is evaluated through applying it in two real-life applications through the case study approach to test is applicability to solve real world problems. The experience gained will be useful for further refinement of the methodology as a continuous improvement process.

Hence, this thesis adopts a case study approach to analyze and evaluate MAKES in two important application domains: namely emergency management and enterprise knowledge audit. Through these two cases, MAKES is implemented, run and evaluated.

3.2 Construction of MAKES

In the MAKES, it is mainly by constructing knowledge models to manage unstructured information. Several knowledge models are applied to represent, store and use information and knowledge of organizations. Based on multi-faceted analysis, recommendations and advices about knowledge assets and knowledge work are produced in some relevant reports.

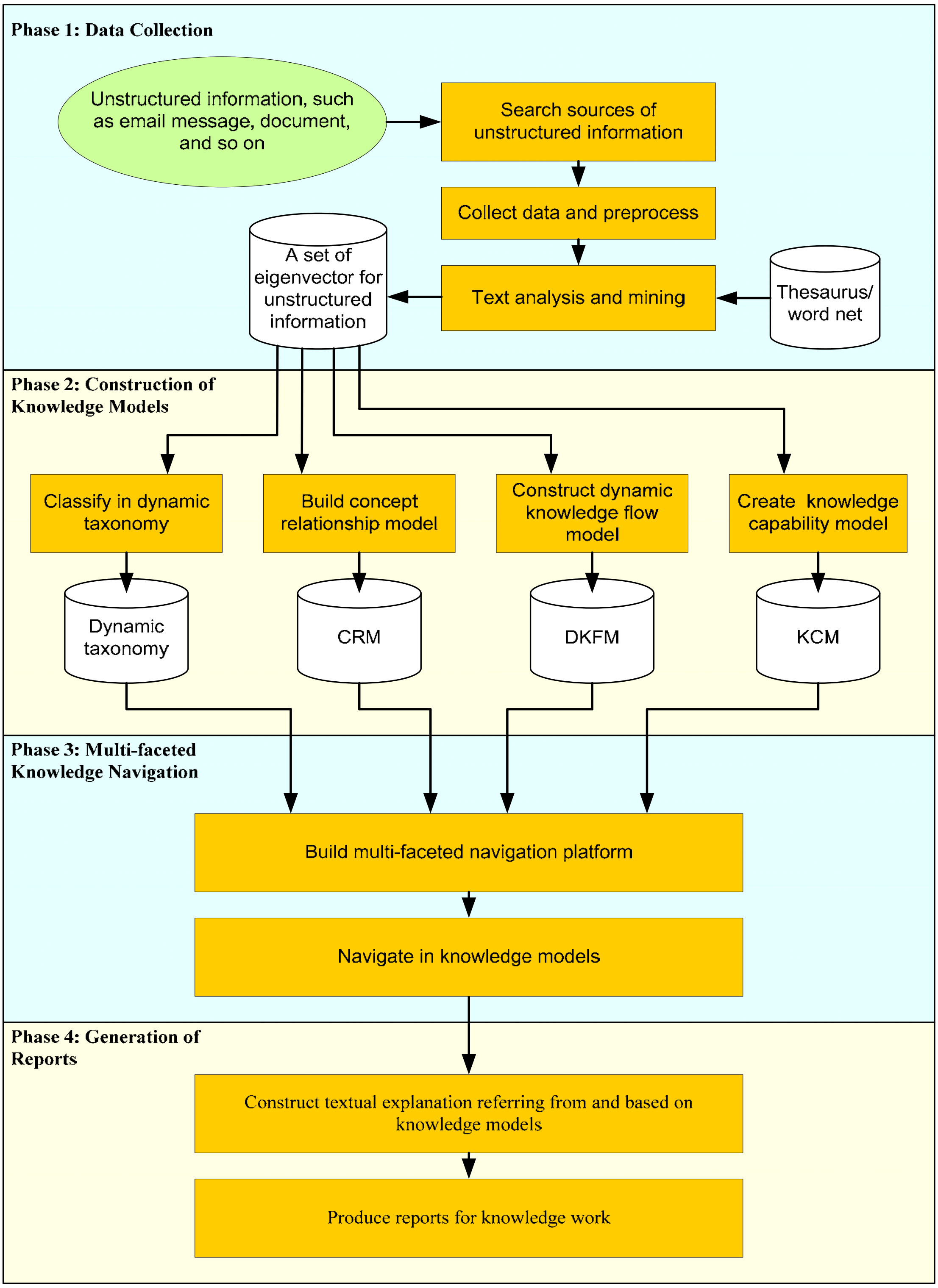

Phase 1: Data Collection
Unstructured information, such as email message, document, and so on
Search sources of unstructured information
Collect data and preprocess
Text analysis and mining
A set of eigenvector for unstructured information
Thesaurus/ word net
Phase 2: Construction of Knowledge Models
Classify in dynamic taxonomy
Build concept relationship model
Construct dynamic knowledge flow model
Create knowledge capability model
Dynamic taxonomy
CRM
DKFM
KCM
Phase 3: Multi-faceted Knowledge Navigation
Build multi-faceted navigation platform
Navigate in knowledge models
Phase 4: Generation of Reports
Construct textual explanation referring from and based on knowledge models
Produce reports for knowledge work

Figure 3.1 The process of MAKES for managing unstructured information.

3.2.1 Phase 1: information collection for acquisition and preprocessing of unstructured information from multiple knowledge sources

The first phase of MAKES is concerned with collecting, searching and preprocessing information from various unstructured sources. Generally, two ways of data collection are proposed. One is collecting and analyzing data automatically through text mining, another way is to collect and analyze the data manually. In this thesis, data is collected automatically without interrupting the workflow and communication of workers by using agents which retrieve email messages from email servers so as to obtain natural and impersonal data. Automatic data collection speeds up the data collection, avoids the intervention of people and enhances the accuracy of data. In the MAKES, the building of a multi-agent mechanism is based on the work of the investigators (Cheung et al., 2004a, 2004b; Wang et al., 2003, 2004) for the automatic collection and classification of data and information from various sources such as emails, intranet, and internet.

A huge amount of unstructured information is analyzed and classified. Dynamic taxonomy provides an effective method of organizing and managing the unstructured information in order to search and retrieve and browse information.

3.2.2 Phase 2: building knowledge models for elicitation and self-maintenance of useful concepts from unstructured information

Knowledge model is an abstract and intuitionistic representation of information and knowledge. A lot of knowledge models have been put forward by different researchers which represent knowledge using different formats. Different knowledge models reflect different aspects of knowledge objects and their relationships. In this thesis, several knowledge models are proposed to enhance the capability of management of unstructured information.

A Concept Relationship Model (CRM) is built through extracting and identifying the concepts, mapping and uncovering their relationships. A statistical theory based on geographic distance is used to compute the similarity between the category and the document. Concept statistics are used to construct the eigenvector of a document and Euler-distance is the measurement for document classification. Although a number of concept relationship mapping tools are available for the presentation of traditional taxonomy, the construction and interpretation of concept relationship model still heavily rely on human effort. A traditional taxonomy is static and requires human intervention for making any subsequent changes to the maps once they have been developed. The construction of the CRM is time consuming and expensive. Moreover, the elicitation of knowledge and its interpretation relying heavily on humans might be subjective (Wang et al., 2007).

In the MAKES, knowledge elicitation is accomplished by an AI-based knowledge elicitation and mining algorithm named Concept Relationship Exploring Technique (CRET). It can uncover useful concepts as well as relationships among those concepts extracted from the content of the unstructured knowledge assets. It also allows for the automatic eliciting and reasoning useful knowledge from the multi-faceted taxonomy map. It would be useful for simulating human learning activities in which complex and unstructured knowledge or unstructured information is involved.

A conceptual summary of seemingly unrelated islands or fragments of knowledge of a particular concept will be synthesized for the users. The concept relationships will be inferred for the users, and hidden relationships among scattered bits of information will be revealed. For example, a relationship can be inferred between "Asia" and "World Health Organization", if the item about SARS is classified under "Asia" and "World Health Organization". At the same time, a relationship between "China" and "World Health Organization" is inferred in the word "Guangzhou" which is a descendant of "China".

The unstructured text will first be preprocessed to filter out irrelevant data and information (e.g. the stop words, HTML tags). Then, the text will be truncated into a word list and checked against the thesaurus model. Words which do not appear in the thesaurus model will be considered as new concepts. The new concepts will be evaluated using a rule-based analysis. In the present study, several rules will be embedded for the recommendations of any new concepts. They include the popularity and density of the new concept. The popularity of the concept is measured by the number of people in the group who share the same concept; while the density of the concept is measured by the frequency with which the concept appears in the unstructured information over a certain period of time. If the popularity and density of the new concept reaches a certain threshold, a suggestion will be made that it be considered ready for revision and retention in the thesaurus model. Words that already exist in the thesaurus model are considered as old concepts. They will be normalized, based on their relationship to synonyms in the thesaurus model, and then the normalized terms will be used in the indexing of the unstructured information.

The second kind of knowledge models is Dynamic Knowledge Flow Model (DKFM). DKFM builds the dynamic pattern of knowledge flow by acquiring the transmitted information among knowledge workers. The names of the sender and receiver of every single message transmitted in knowledge work can be captured. And then, a network graph of the knowledge flow is constructed. DKFM enables to find out both rational and irrational work patterns by analyzing some sub-structures in the pattern of knowledge flow.

Knowledge Capability Model (KCM) is adopted to measure the capability of every knowledge worker in knowledge work quantitatively. At first, KCM captures all the messages sent and received by a certain knowledge worker, and then, counts the frequency of every knowledge concept occurring in these messages. The frequency denotes the workload of each knowledge domain which the knowledge worker works on. A radar graph is adopted in KCM to express the workload of every knowledge concept of every knowledge worker. Hence, KCM may enable us to find out the advantages and shortcomings in the knowledge domain of every knowledge worker, what he/she is good at and what he/she is not good at. The assessment of every knowledge worker in KCM may improve the usage of human resources in an organization.

In this thesis, multiple forms of knowledge representation of organizational knowledge are built in these knowledge models. Multiple modes are provided for the support of knowledge work and to improve its efficiency.

3.2.3 Phase 3: multi-faceted navigation and analysis of knowledge flow

In the third phase, the multi-faceted navigation is devoted to addressing the need for navigating through unstructured information from different starting points or points of view. After analyzing the concept's relationships, the unstructured information will be browsed in different ways. The unstructured information will be displayed in a ranked list, and the concept relationships of the retrieved information will be presented in a pictorial view. When a search query is entered into the system, the system will retrieve the corresponding knowledge assets from the knowledge repositories, which will be indexed in the concept elicitation and maintenance module. Multi-faceted navigation will enable users to have a deep insight through observing and comparing the various patterns in the knowledge model from different points of view.

Knowledge workflow is an important pattern of work in knowledge intensive organizations. Knowledge work is complex, has no fixed procedure, and is usually non-linear. In a human network or a social network, knowledge is dispersed and scattered among the minds of all the employees. Knowledge work is performed within a social network. As this is the network of functions and activities, knowledge is constantly being exchanged, fused, and created among its participants through some kind of social interaction. Knowledge flow reflects knowledge workflow in nature.

Dynamic Knowledge Flow Model (DKFM) will be constructed to support the analysis of knowledge work qualitatively and quantitatively so that the best patterns of knowledge work are found. The term "dynamic" in DKFM means that the knowledge flow model can be built through the process of automatically extracting and mining the knowledge from knowledge work, and reflect the latest pattern of knowledge work. DKFM analyzes the patterns of knowledge work in real-time and online. The DKFM operates in two phases: text mining, and modelling. The first phase is building the characters of messages by analyzing text messages from knowledge work while the second phase involves modelling and analyzing the DKFM. The latter adopts visualizing technologies to show the patterns of knowledge flow in a diagrammatic way through analyzing the characters of messages. The schema of DKFM is an automatic and intelligent process in which some AI technologies are used, such as some knowledge rules and pattern recognition.

Generally, nodes in DKFM denote knowledge entities while lines denote the relationships among knowledge entities. The knowledge entities are what can provide or produce knowledge and information such as databases, documents, files, knowledge repositories and human. Relationships among knowledge entities will be established through communication and interaction of knowledge work. These relationships include ordinary office work and workflow, querying, finding, searching, answering, discussing and transferring information. Users will display the multiple patterns of knowledge work through multidimensional analysis including slicing up and down. For example, one could choose the knowledge domain of finance and show the pattern of knowledge work related to finance.

3.2.4 Phase 4: generation of reports inferring from knowledge models

The final phase of MAKES will produce some appropriate recommendations and suggestions for managing unstructured information. In knowledge work, a knowledge-based system will be built, based on the previous work of the investigators (Cheung et al., 2005, 2007). This system will synthesize some techniques of artificial intelligence, such as logic inferring, case-based reasoning, to produce some reports about knowledge work automatically. The reports contain some findings from knowledge assets and their relationships of organizations, insights about knowledge work, and etc. These text reports provide readable facts and findings associated with the relevant graphical knowledge models. These reports enable the users of MAKES to understand the relationships of knowledge assets intuitively

3.2.5 Verification testing of MAKES

Verification testing means to run the system in a simulated environment using simulated data. This simulated test is sometimes called alpha testing. The simulated test is primarily looking for errors and omissions regarding end-user and design specifications that were specified in the earlier phases but not fulfilled during construction. In this thesis, two applications are run for verification testing, one in emergency management and one in knowledge audit, to verify the capability of MAKES for managing unstructured information. Emergency management is an emerging realm for governments and organizations to respond to sudden catastrophic events rapidly and effectively during a crisis, e.g. social unrest, natural disaster, unforeseen event, and so on. In all cases, a huge amount of unstructured information is involved. These scenarios stipulate the building of appropriate knowledge models to find out what is known about the event, where the crisis is and how it is represented, from unstructured and seemingly unconnected information. Knowledge audit aims to compile a list of knowledge assets of an organization that is embedded in its business process, and in the know-how of its employees, as well as its distribution and associated risks.

In the trial application of emergency management, a great number of text documents and files are collected into MAKES to be mined and constructed as knowledge models. As soon as an emergency occurs, the knowledge models of emergency management are called upon to form the related knowledge map to help administrators to locate relevant information and decide what action should be taken. Concept map is used to give the clues for decision making about emergency management. Social network analysis identifies the key individuals or teams who are responsible for executing the tasks necessary to deal with the emergency. Knowledge flow provides knowledge support to the emergency work.

The audit work is mainly carried out automatically from the e-mail system. MAKES collects data from email messages and constructs social relationships and knowledge flows. The results of social network analysis will be used at the level of individuals, departments or organizations to identify teams and individuals playing central roles - thought leaders, key knowledge brokers and experts. It will also spot opportunities for knowledge flow improvements to accelerate the flow of knowledge and information across functional and organizational boundaries, and to forecast opportunities for technology development in order to identify business opportunities.

3.3 Data collection and analysis

3.3.1 Automatic data collecting

To manage unstructured information, it is necessary to collect multiple types of data and information from various sources. There are many ways to collect data and information, including document search, desktop search, interviews, questionnaires, focus groups, workshops, etc., and a mixture of all these.

Interview is divided into two types, formal and informal. A formal interview needs a scheduled meeting and has an agenda. In this situation, the interviewee is usually nervous and tense. An informal interview means that the interviewer will go to various work sites to observe the content and flow of work and talk with workers. However, the data and information collected by this kind of method of data collection is not enough for knowledge management. A questionnaire needs a set of purposely designed questions. The accuracy of data collected by this kind of method of data collection depends on the integrity with which the informant fills up the questionnaire, as well as on the design of a set of right questions.

According to NHS (2005), questions will be typically based on factors such as:

i） Who knows who and how well?

ii） How well do people know each other's knowledge and skills?

iii） Who or what gives people information about xyz?

iv） What resources do people use to find information/feedback/ideas/advice about xyz?

v） What resources do people use to share information about xyz?

From another viewpoint, the methods of data collection can be divided into two types also, namely manual collection and automatic collection. Usually, the method of current data collection mainly depends on human labor and usually entails a heavy workload. This is a time-consuming work.

In this thesis, the automatic method of data collection is developed in order to monitor and obtain a great amount of dynamic data and information from multiple sources, such as emails, intranet, internet, and so on. Nowadays, email is a main carrier of data and information which is produced during daily work in many organizations, including office work, communication, transferring documents and files etc. Emails which are written in natural language are one kind of unstructured information resource in that they contain a great amount of knowledge about organization. In fact, the amount of data and information stored in an email system is huge. The content in email includes fresh and dynamic information which usually narrates the latest development of knowledge and is the imperative and important part of knowledge assets of organizations, and even the critical part. Such automatic data collection although limited to the explicit information stored in electronic form, is nevertheless a good way of collecting information without interrupting the workflow and communication of workers. The automatic approach speeds up the data collection stage of the knowledge management, avoids intervention by people, and enhances the accuracy of data.

People often use two kinds of tools to manage emails. One is based on a client software to manage email messages. For example, workers often use Windows Outlook application to receive and send email. Another pattern is called Webmail which enables worker to access email by using a web Browser, such as gmail from

Google. In the Client pattern, email messages are usually stored in the client's software and the email server does not save email messages, in other words, organization's knowledge is dispersed and resides in workers' desktops. In this situation, workers often send email messages to each other through forwarding and copying email for collaborative work. So, the client pattern produces a huge amount of repeated email messages and makes knowledge management more complicated. Another disadvantage of the client pattern is that a worker's demise may result in the loss of some information because email messages in that worker's personal computer may be lost. Nowadays, more and more organizations adopt Web-based email systems to solve the above problems of managing email. Now they use Browsers based on a central email server. In the Browser/Server pattern, organizations can obtain and store all office email messages so that no important information will be lost.

Many enterprises in Hong Kong have been adopting email systems to do routine work. This thesis proposes the setting up of a Web-based email server, and constructing a mechanism of automatic email collection, intercepting all email messages, and analyzing the text in the email messages. The MAKES will enable staff to manage and process unstructured information from emails transparently and automatically based on a central email server.

3.3.2 Multi-faceted data analysis

In knowledge management, knowledge models are constructed through analyzing the data and information collected from various sources in order to represent, store and share knowledge expediently in organizations. The process of data gathering and analysis provides a baseline against which the priorities, and the appropriate changes, and interventions to improve the efficiency of knowledge management, can be planned.

Since most of the information collected from various sources is unstructured information which is usually written in natural languages, Natural Language Processing (NLP) is becoming the state-of-the-art technique in knowledge management. These technologies of NLP can be divided into two types, one is based on semantics and the other is based on statistics. The semantic-based technologies are constructed on the basis of a semantic model of natural language through research into the linguistics of the language. According to the relevant research at Harvard University, a research group with ten experts in English linguistics and computer science took four years to design nearly three thousand semantic and grammatical rules and develop a machine which could analyze English text automatically. The outcome of the Harvard project reached ratio 96% of exactness in understanding English articles. It can be imaged that developing another system to analyze the articles written in other natural language is very complexity and huge workload. In this thesis, we could not follow the lead of this kind of semantic method as our research resources are limited. Instead, the linguistic analysis technique based on statistics is adopted to build the prototype model.

The theory of statistics-based linguistic analysis is based on the faith that the terms or words which appear most frequently in articles represent the main content of the article, to a certain extent. In the statistics-based approaches of NLP, a term database or vocabulary is indispensable. Usually a vocabulary can be obtained from some public sources, or constructed and maintained by the developer. Generally, there are two main characteristics of the statistics-based method of NLP. One is that the workload of research work is manageable, and the other characteristic is that it can be applied in processing the articles written in multiple languages. The researched objects in this thesis are focused on the organizations in Hong Kong and Mainland China. The files and emails in these organizations are usually written in English, Traditional Chinese, Simplified Chinese, and their mix. These multilingual documents form the knowledge assets of organizations. This thesis adopts the statistics-based method to analyze text and build an architecture which can process multi-lingual text to elicit knowledge based on statistics.

Constructing an appropriate model of knowledge storage is critical to acquiring and managing knowledge from unstructured information. Unstructured information is not easy for a machine to read and understand. So it is difficult to support machine learning and intelligent reasoning using unstructured information. An appropriate knowledge model will enable knowledge to be structured in order for people to perform the activities of searching and retrieving. There are many models of knowledge storage. Taxonomy is often used as a method of knowledge storage and representation. In general, a taxonomy is built in advance by humans. Since knowledge and its complexity increase quickly with time, dynamic taxonomies have been proposed as a better tool to describe and classify complex and heterogeneous information and knowledge bases (Sacco, 2000). Classification in a dynamic taxonomy can be adjusted automatically according to the content of the collected information. It can be developed over a period of time to provide an efficient and effective classification management system for unstructured information.

Knowledge needs to be represented in some forms that people can understand easily. A concept relationship map is a simple, direct and intuitive method of knowledge representation which uncovers the knowledge concepts and their relationships by using graphic patterns. Patterns are built up which represent the relevance of knowledge concepts and depict the dependency relationships between concepts. For constructing concept relationship maps, this thesis proposes an automatic method to extract terms and build the dependency relationships among concepts by computing the occurrence frequency between concepts.

The latest development of knowledge management tells us knowledge is not only static but also fluid and dynamic. Therefore, there is a need to depict the patterns of knowledge flow in organizations and to assess the efficiency of workers and their capacity for collaborative work. Social network is a method to map the patterns of knowledge work. In general, different content of knowledge work needs different domain knowledge and needs relevant expertise and professionals to take part in such work. To represent and analyze knowledge work, there is a need to rebuild the social network, to expand the attributes of relationships and record the knowledge domains related to the relationships among the knowledge workers. In this thesis, mapping the pattern of knowledge work is performed by tracing email communications among knowledge workers. The pattern of email communication in organizations contains the pattern of knowledge work. Sending or receiving an email shows the work relationships of knowledge workers. These can differ significantly from the relationships that are shown on a formal organization chart. The attributes of relationships of knowledge workers can be gained from analyzing the email messages exchanged between them. For example, the workers in the financial department will use the terms in financial domain. Such as cash, payment, cheque will be mostly displayed in their email messages. Terms which are more technical will appear in the email messages in an engineering department. Using statistics, we can draw a profile of the knowledge content of every worker and show the characteristics of the relationships between workers. The scope of the knowledge domains which are owned by knowledge workers can be illustrated using a radar graph to make up for shortcomings of the social network.

One of the main objectives of knowledge representation is to support knowledge application and to enable knowledge creation. There are many methods of knowledge representation which have different usages. Referring to the idea of the multi-dimensional model based on a data warehouse in the field of data processing, this thesis provides a multi-faceted knowledge navigation platform to provide multidimensional knowledge views for organizations. This tool extracts content and patterns of knowledge work from the view of concept association, social relationships and knowledge work flow. These representations form a synthesis and are integrated in a knowledge map. The knowledge map portrays the sources, flows, constraints and sinks (losses or stopping points) of knowledge within an organization. It uses explicit and codified knowledge, and tacit knowledge to uncover the relationships among knowledge entities.

From the taxonomy, the knowledge inventory is summarized. The report of knowledge inventory includes the names of the people who possess knowledge and the names of the places where knowledge resources are stored. The degree of importance of every knowledge resource is reported after it has been measured by counting the frequency with which the knowledge resource is used in knowledge work. The concepts and their relationships are evaluated through analyzing the concept map. The key concepts are identified and automatically ranked according to some rules. The tendency of technology development can be inferred and forecasted through comparing the changes in key concepts in different periods of time.

DKFM uncovers the pattern of the flow of knowledge and information in knowledge work. The pattern of DKFM associated with a certain knowledge domain is constructed through analyzing the properties of the relationships of knowledge sources. The key knowledge sources can be recognized and assessed quantitatively. Social network can identify the teams and individuals playing central roles - thought leaders, key knowledge brokers and experts, and find out which teams or individuals are isolated. The strengths of relationships between knowledge actors can be derived from the social network analysis, then measured and evaluated. The results of a social network analysis will be used at the level of individuals, departments or organizations.

KCM displays the competence of every knowledge worker. The important knowledge domains associated with the work of knowledge worker are identified and measured. The ability of a knowledge worker to learn can be inferred by analyzing the transformation of the patterns of knowledge capability of the knowledge worker. This can then be reported on.

3.4 System design

3.4.1 System architecture

The framework of MAKES is composed of three tiers shown in Figure 3.2. The input tier is the data input of the system. It includes the unstructured information of the company (e.g. emails), the information outside the company (e.g. the WebPages of competitors), and the public thesaurus of the domain industry. MAKES is basically composed of four components which are multi-agent modules: thesaurus model, concept elicitation and maintenance, concept relationship, and multi-faceted navigation. The output tier includes: the knowledge inventory, social network analysis (among the staff and client of the company), and the relationship analysis among the concepts. The output of MAKES provides a convenient way to examine knowledge among different people and different concepts which supports a multi-faceted taxonomy. The middle tier is a process tier and the processes of knowledge modeling and multi-faceted knowledge navigation are developed in this tier.

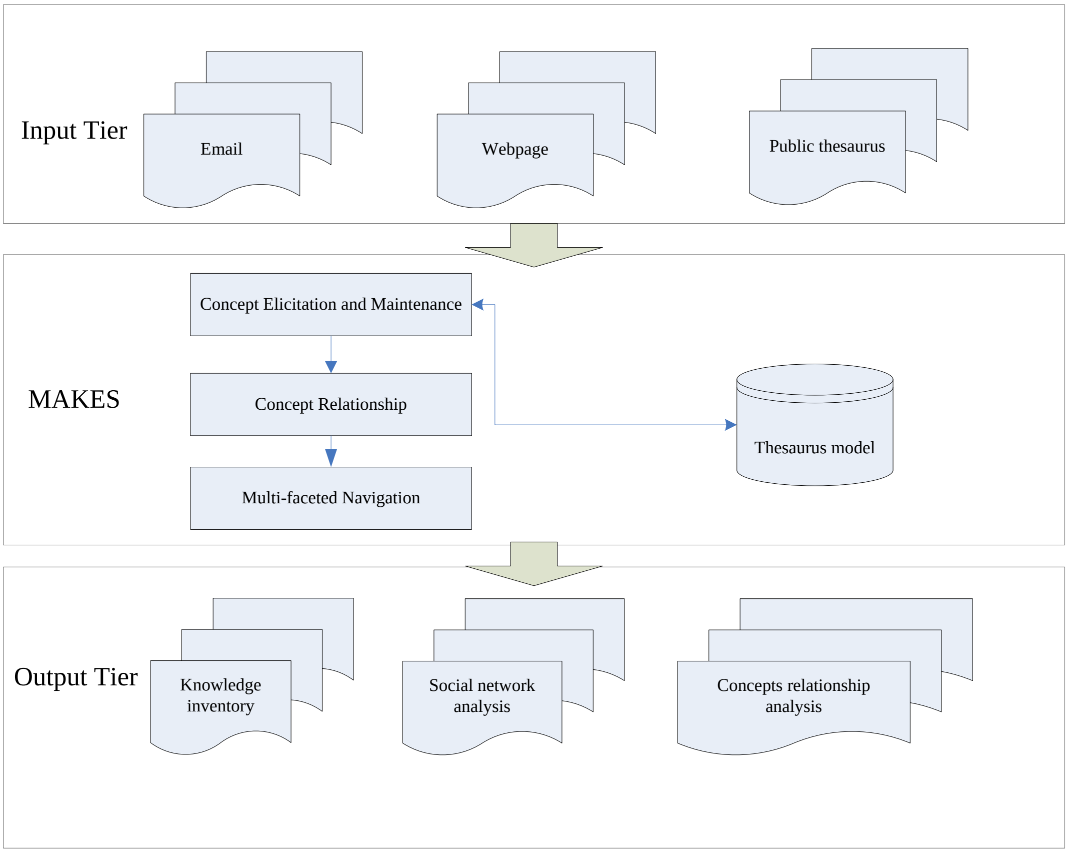


Figure 3.2 The system framework of MAKES.

The MAKES adopts a modular architecture based on multi-agents, namely: a Thesaurus Model Module, a Concept Elicitation and Maintenance Module, a Concept Relationship Module, and a Multi-faceted Navigation Module.

The thesaurus model (Cheung et al., 2005) contains the controlled vocabularies, synonyms of concepts, stop words, and the hierarchies and relationships among the words. The thesaurus model is continuously updated by two processes. Firstly, there are a number of thesauruses of the application domain in the public realm. Most of them allows for Really Simple Syndication (RSS). An automatic engine will be built to extract the thesauruses from the websites of the public thesauruses. When something is updated in the public thesauruses, the updated details will be brought to the attention of authorized personnel of the company who may wish to revise the company's thesaurus. New terms will be retained in the thesaurus model of the company for future use. The thesaurus model will be updated through the analysis of new incoming information (e.g. emails or documents) and analysis of the WebPages of competitors (Cheung et al., 2004a).

The Concept Elicitation and Maintenance Module works with the input tier and the thesaurus model. Its aim is to construct and maintain the thesaurus model automatically by analyzing unstructured information from various text sources.

The Concept Relationship Module depicts the relationships of concepts relevant to a given domain. The relationships between concepts are represented in such a way as to show the different degrees of importance of the concepts. The Multi-faceted Navigation Module navigates unstructured information in different dimensions, after the analysis of concept associations. Moreover, the concept relationships and the social network among the retrieved information are presented in a graphical view. When a search query is entered into the system, the system retrieves the corresponding knowledge assets, which are indexed in the concept elicitation module, from the knowledge assets database.

All of the three modules are interconnected. Generally, the front module is the base of the later system. The Concept Elicitation and Maintenance Module provides the structural data and information to the Concept Relationship Module, which enables the model of concept associations to be constructed easily. The Multi-faceted Navigation Module makes use of the models from Concept Elicitation and

Maintenance Module and the Concept Relationship Module to undertake the navigation operations in knowledge assets among the multi-dimensions. Through a collaborative work pattern based on multi-agents in these modules, agents can be supported as they work automatically and collaboratively to serve the MAKES with high efficiency.

3.4.2 Multi-agent collaborative mechanism

A multi-agent mechanism is a hierarchical collaborative mechanism with multiple intelligent agents. Basically, the agents in modules can be classified into three types and organized into three sub-tiers (Figure 3.3):

i） The coordination agents which are responsible for scheduling and monitoring the collaborative work among agents.

ii） The function agents which perform the concrete business processes.

iii） The common agents which are the basic and public function agents, e.g. web search engines.

All autonomous agents are controlled by the coordinating agent. A global blackboard is built to store and transfer the messages among the agents. All agents are built on the basis of a common agent model (Figure 3.4) which is composed of three functional blocks, Intelligent Process Block, Communication Block and Knowledge Block.

The Knowledge Block stores two types of data and knowledge rules which are used to guide the reasoning for the intelligent process block and data/case base. The

Communication Block is responsible for communicating with other agents through reading/writing messages on the global blackboard.

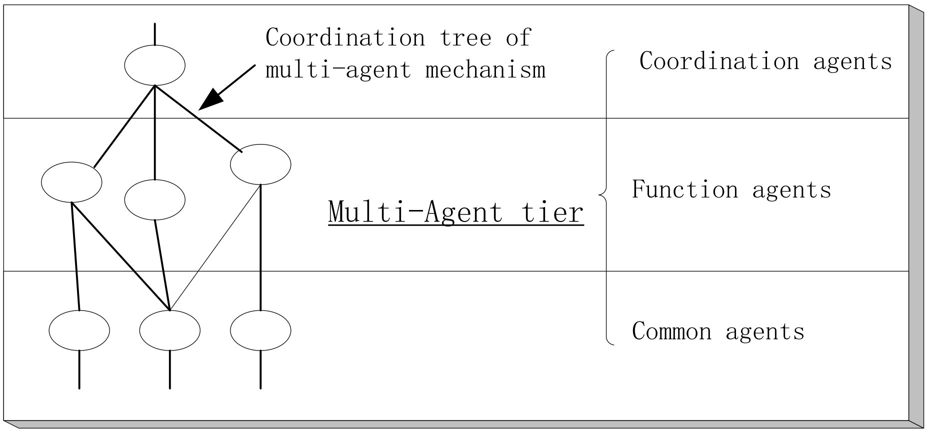


Figure 3.3 The hierarchical multi-agent mechanism of MAKES.

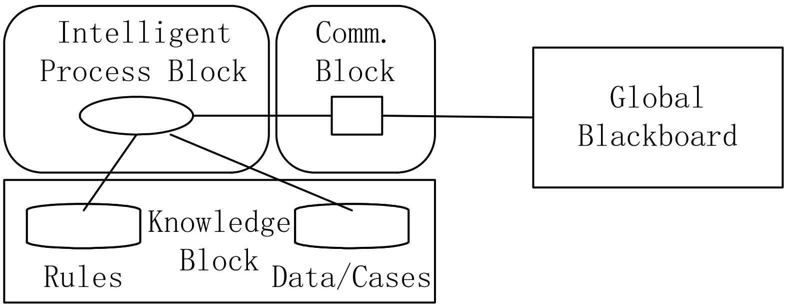


Figure 3.4 The framework of agent.

3.4.3 Knowledge processing patterns of MAKES

Basically, MAKES is divided into two phases for processing knowledge: the learning phase and the application phase. Figure 3.5 gives a schematic diagram of the learning phase in MAKES while a schematic diagram of the application phase of the MAKES is shown in Figure 3.6.

From Figure 3.5 and 3.6, the entities in MAKES denote the various knowledge sources. They include human beings, intelligent agents, and even other Knowledge Management Systems (KMSs). Entities are the sources of knowledge for MAKES, and knowledge supports the collaborative work among the entities.

In the learning phase, when the messages are transferred through a knowledge portal, they are decomposed into message fields and word strings. The receiver field in a message provides the information for locating the address of the receiver so that the message is forwarded to the correct receiver. Meanwhile, the terms extracted from the fields of the message, and the message is then classified into the taxonomy system according to how the message matches the terms of the taxonomy.

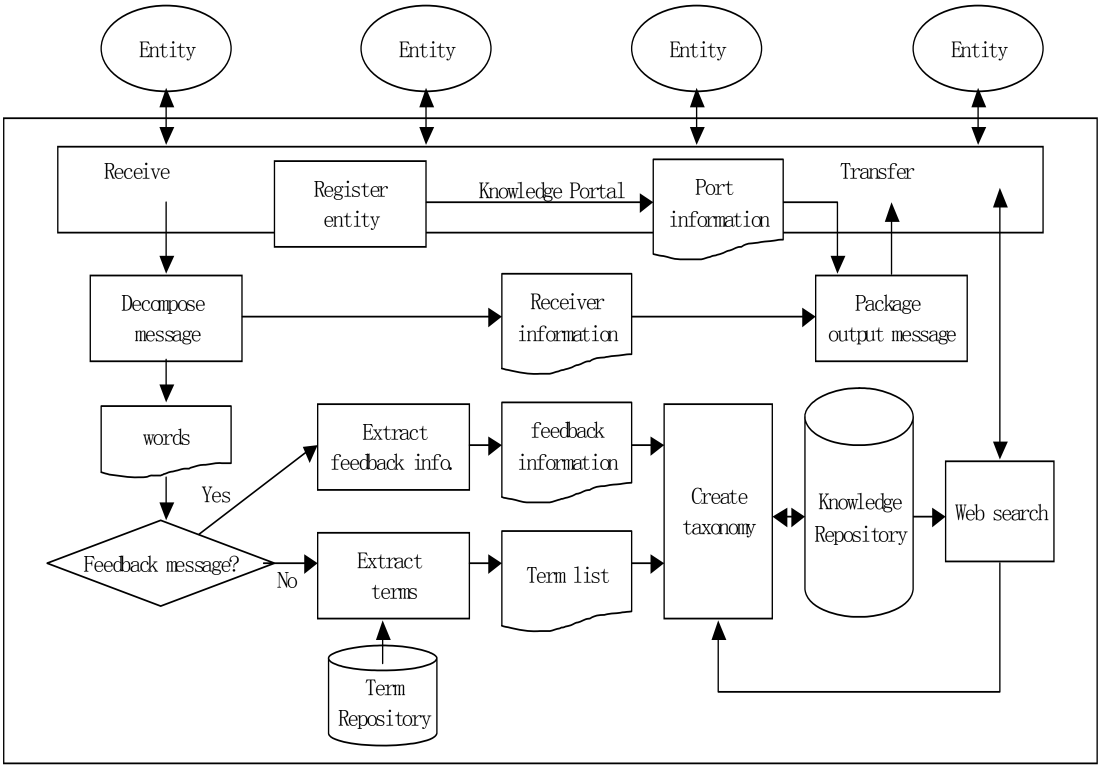


Figure 3.5 A schematic diagram of the learning phase of MAKES.

The feedback information is about the measurement and the assessment of the performance of the knowledge resources and knowledge work. This feedback includes several performance indicators such as the indicator for the degree of satisfaction with the knowledge resources or with the work. The assessment of knowledge resources and work is useful for optimizing the knowledge workflow.

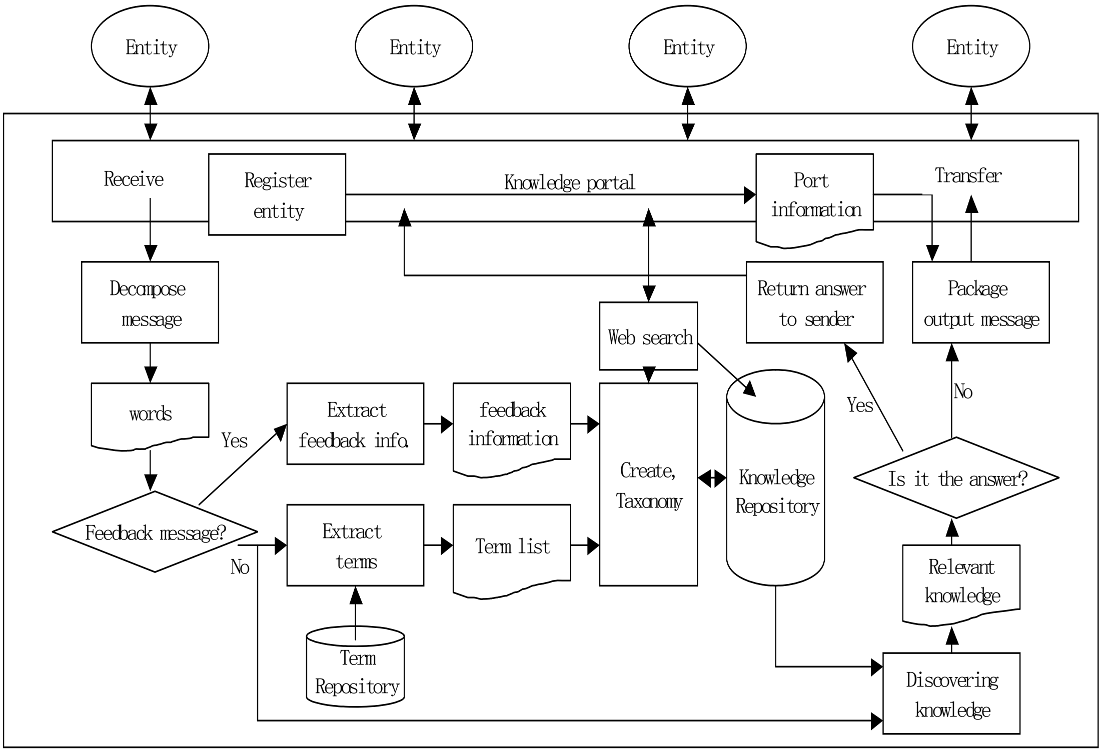


Figure 3.6 A schematic diagram of the application phase of MAKES.

In the application phase, the messages cannot be forwarded to the receiver directly as in the learning phase. The agent for discovering knowledge is responsible for finding out the appropriate knowledge or locating expertise in the taxonomy, and then, knowledge workflows are constructed automatically and intelligently, based on mining knowledge from the knowledge repository.

In the multi-agent architecture of MAKES, the blocks in the schemas of knowledge processing patterns are responsible for one agent or for a cluster of agents which work collaboratively to perform the specific functions. The design of key agents with their algorithms is presented in the following chapter.

3.5 Development platform

The specification of the development platform is shown below:

(i) Programming language

Java is an effective object-oriented programming language which is widely used. JavaBean provides the component-based mechanism to support the high level runtime performance. In collaborative work, it is essential that the system is able to operate in different platforms. The main character of Java language is cross-platform.

It is a rational choice that the MAKES is developed by using JAVA because the MAKES needs to be adaptable to various runtime platforms and different organizations some of which maybe using different operating systems, such as Windows, Linux, etc.

(ii) Network environment

The MAKES must collect data and information from multiple sources on the Internet. So, the web–based technologies are adopted in the MAKES. The system of MAKES uses the application pattern of Browser/Server, and users can use standard internet Browser to do knowledge work. The back-ward data collection must adopt the technologies of search engines based on the Internet protocol.

(iii) Database management system

Concerning the building cost of developing an information system for an organization, a valuable strategy for reducing the cost of the MAKES is to use free software products. MySQL is free software with high performance which can be run in multiple platforms on different Operating Systems, such as Windows and Linux. So MySQL can cope with both free and cross-platforms. It is useful for the MAKES to be applied widely using MySQL because MySQL is free.

The following chapters will describe the designs and key technologies of every phase for developing MAKES. The cases for verification testing are narrated after the technique chapters.

Chapter 4. Text Mining and Dynamic Taxonomy

A taxonomy is a key structure for storing knowledge assets and building knowledge models in MAKES. Text mining provides an important approach to building a dynamic taxonomy automatically.

4.1 Architecture of text mining

The concept elicitation and maintenance module works with the input tier and with the thesaurus model as shown in Figure 4.1.The output of the module is a dynamic taxonomy. The thesaurus model contains the controlled vocabularies, synonyms of concepts, stop words, and the hierarchies and relationships among the words. The thesaurus model is continuously updated by two processes. Firstly, there are some thesauri of the application domain existing in the public realm. Most of them enable Really Simple Syndication (RSS). An automatic engine is used to extract the thesauri from the websites of the public thesauri. When anything is updated in the public thesauri, the authorized personnel of the company are prompted to use the updated details for revising the company thesaurus. Hence, new terms are retained in the thesaurus model of the company for future use.

Secondly, the thesaurus model is updated through the analysis of new incoming emails or documents of the company and from the WebPages of the competitors. The unstructured text is firstly preprocessed to filter out irrelevant data and information (e.g. the stop words, HTML tags). Then, the text is truncated into a word list and checked against the thesaurus model. Words which do not appear in the thesaurus model are considered as new concepts. The new concepts are evaluated based on a rule-based analysis. In the present study, several rules are embedded for the recommendation of new concepts. They include the popularity and intensity of the new concept.

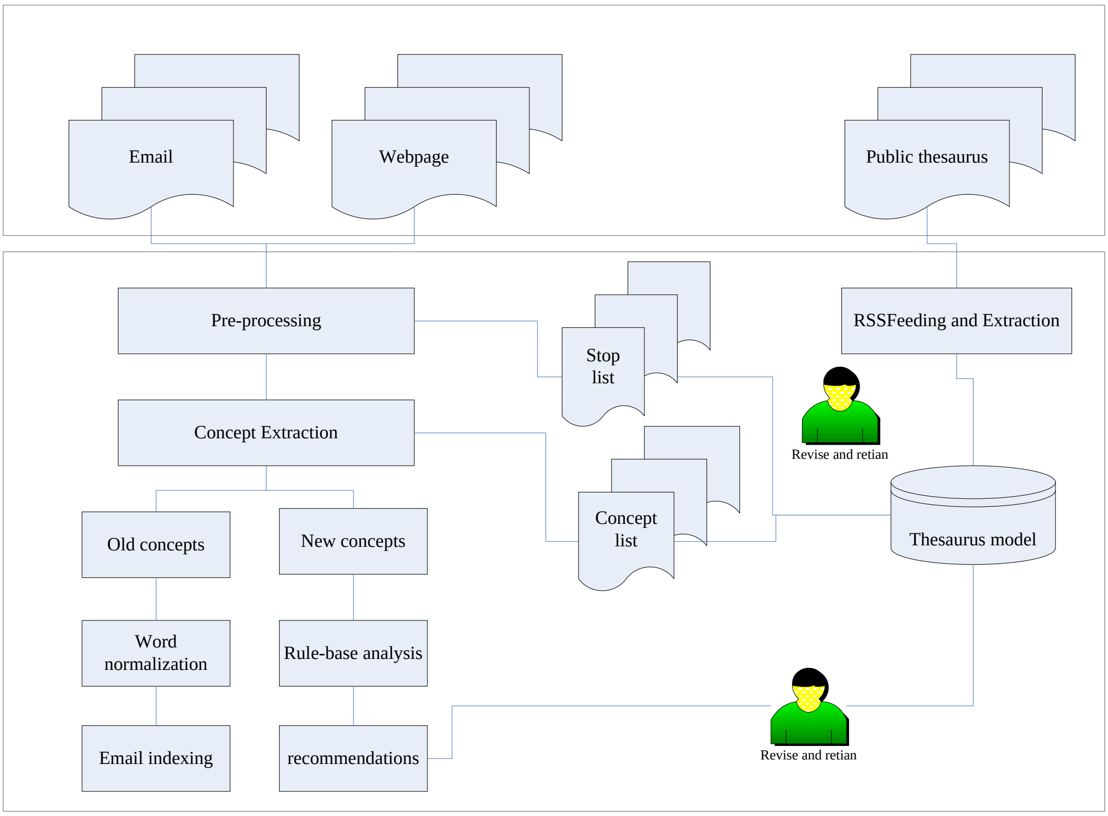


Figure 4.1 The schema of concept elicitation and maintenance module.

The popularity of the concept is measured by the number of people who are sharing the same concept within the group, while the intensity of the concept is measured by the frequency with which the concept appears in the unstructured knowledge assets over a certain period. If the popularity and intensity of the new concept achieve a certain threshold, it is suggested for revision and retention in the thesaurus model. On the other hand, they are considered as old concepts if the words already exist in the thesaurus model. They are normalized based on the synonym relationships in the thesaurus model, and then the normalized terms are used for email indexing.

The module of concept elicitation and maintenance consists of four agents, Information Pre-processing Agent, Concept Extraction Agent, Dynamic Taxonomy Agent, and Thesaurus Model Maintenance Agent. The Information Pre-processing Agent is responsible for capturing information and decomposing the text according to linguistic rules, and then the segment of text is fed to the Concept Extraction Agent to extract the concepts according to on the thesaurus model. As soon as the concepts are extracted, the Thesaurus Model Maintenance Agent decides whether to modify the thesaurus model or not. The Dynamic Taxonomy Agent is responsible for automatic classification of the unstructured information.

4.2 Information pre-processing

4.2.1 Pre-processing of unstructured information

The first task of MAKES is capturing information from various sources. Because email has become one of the most important and widely used communication media, some companies consider the email accounts of their staff as the company’s assets since a great amount of business information and corporate knowledge is embedded in them. Email clients such as Outlook and others are client-side applications which receive email messages from email servers on the Internet and store them in the hard disk of personal computers. A large number of dispersed email messages is not convenient and is difficult for staff to share and search, and information contained in them cannot be effectively organized for re-use. A web-based email system is one of the solutions to managing business information and the knowledge of the company as it stores all email messages in a centralized database server.

The Information Pre-processing Agent manages the Information Capture Agent (ICAgent) which allows staff to use a commercial browser to send and receive email messages, and save email messages into a database. It should be noted that the ICAgent itself is not an email server, instead it contains two email agents, the Send Agent and the Receive Agent as shown in Figure 4.2.

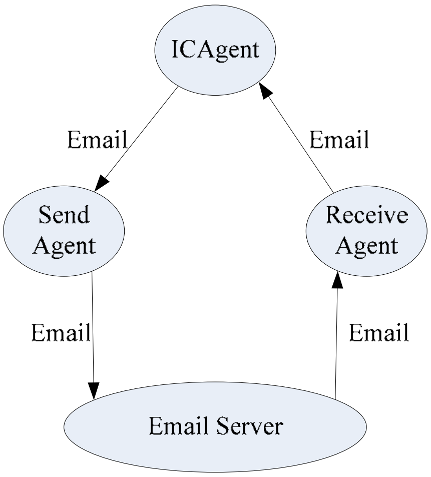


Figure 4.2 A coordination schema of multiple agents for capturing email.

Send Agent sends email messages through connecting SMTP servers, while Receive Agent receives email messages from POP3 servers on the Internet. This scalable and flexible structure enables MAKES to transfer email messages and collect information and knowledge while it is working on the Intranet, Extranet & Internet.

4.2.2 Decomposing information

As soon as the information is captured, the Information Pre-processing Agent begins to process the messages. In practice, email is unstructured information and viewed as a string or bag of words. The importance of different fields in email is different. For example, the content of the subject is more abstract and important than the content of the body. Different fields of email should be treated differently according to the importance of the field.

Mathematically, one email is represented as a 4-tuple $(\mathbf{Sd}, \mathbf{Rv}, \mathbf{T}, \mathbf{B})$, where

$\mathbf{Sd} = < s_{s1} s_{s2} ... s_{sn} >$ is a sequence of symbols in the sender field and $s_{si} \in \mathbf{S}$, where $\mathbf{S}$ is a set of letters of the alphabet of English, other letters of the alphabet, or numbers. In fact, any symbols could be transferred, $\mathbf{Rv} = < s_{r1} s_{r2} ... s_{rn} >$ is a sequence of symbols in the receiver field and $s_{ri} \in \mathbf{S}$, $\mathbf{T} = < s_{t1} s_{t2} ... s_{tn} >$ is a sequence of symbols in the subject field and $s_{ti} \in \mathbf{S}$, and $\mathbf{B} = < s_{b1} s_{b2} ... s_{bn} >$ is a sequence of symbols in the body and $s_{bi} \in \mathbf{S}$.

The task of the Information Pre-processing Agent is to decompose the email messages into some text segments according to the rules of linguistic sentences and/or paragraphs. These segments are fed into Concept Extracting Agent for further analysis. The rules are stored in the thesaurus model which directs how to decompose text. In MAKES, a message is divided into segments by stop words and the relevant rules. For example, the symbol "." is generally thought as the end of sentence, namely, a stop word. Other stop words include "(", ")", "," etc. The stop symbol is usually not used in any concept or term. The objective of using stop words is to cut message into small segments and thus decrease the complexity of analyzing the message.

After pre-processing, the 4-tuple $(\mathbf{Sd}', \mathbf{Rv}', \mathbf{T}', \mathbf{B}')$, the message is transferred into the following format:

$\mathbf{Sd}' = < sg_{11} >$ is a segment of the sender field,

$\mathbf{Rv}' = < sg_{21} sg_{22} ... sg_{2m} >$ is a sequence of segments in the receiver field,

$\mathbf{T}' = < sg_{31} sg_{32} ... sg_{3m} >$ is a sequence of segments in the subject, and

$\mathbf{B}' = < sg_{41} sg_{42} ... sg_{4m} >$ is a sequence of segments in the body,

where any sg_{ij} in $\boldsymbol{Sd}'$, $\boldsymbol{Rv}'$, $\boldsymbol{T}'$, $\boldsymbol{B}'$ is a sequence of the symbol, $sg_{ij} = < t_{i1} t_{i2} ... t_{ij} >$, $t_{ij} \in \mathbf{S}$.

As soon as a message is decomposed into some segments, the task of the Information Pre-processing Agent has finished and these segments are stored into the database. Then, the Concept Extracting Agent is invoked.

4.3 Concept extracting and elicitation

4.3.1 Concept matching algorithm

The Concept Extracting Agent (CEA) is responsible for extracting concepts from messages through analyzing the small segments of the message. When a segment is handled, CEA matches concepts from the thesaurus model to the segment. As soon as a concept has been matched, the concept and its occurrence number are recorded.

The theoretical basis of a concept matching algorithm is that the longer the length of the concept, the more consolidated the meaning of the concept. For example, the concept of “The Hong Kong Polytechnic University” is more concrete than the concept of “Hong Kong”. Of course, the concept matched from a message should be as concrete as possible. The concept matching algorithm adopts the method of reverse matching. The pseudocode of the concept matching algorithm is described as below:

```
begin
    define $T =< t_1, t_2, ..., t_n >$ is an array of symbols;
    input $S_{ij}$; // $S_{ij}$ is a segment from message
    set $T = S_{ij}$;
    set $i = 1$;
    set $j = n$;
    while $i \neq n$ do
```

```
            while $i \neq j$ do
                set $Tp = < t_i, t_{i+1}, ..., t_j >$;
                match $Tp$ to the concepts of the thesaurus model;
                if matching is successful then
                record the sequence number of the concept in the thesaurus
                and increase its occurrence number;
                    $i = j$;
                    exit; (go to the external loop)
                    or
                    set $j = n - 1$;
                endif
        enddo
          set $i = i + 1$;
          set $j = n$;
      enddo
  end.
```

For example, here is a sentence “The Hong Kong Polytechnic University is a University in Hong Kong”. The concept matching algorithm can find three concepts at least which include “The Hong Kong Polytechnic University”, “Hong Kong”, and “University”. The meanings of these three concepts are different. The first one means the name of an organization, the second one is a location while the third one is a property of an organization. The relationships of these concepts are built in the thesaurus model.

4.3.2 Eigenvector of the message

After extracting concepts from messages through analyzing and matching concepts, the eigenvector of the message is built and stored. To construct the eigenvector of the message, some basic structures are defined as follows:

Definition 1. $\mathbf{U} = \{u_1, u_2, ..., u_m\}$ is a whole set of users and knowledge entities, and u_i denotes the i^{th} user or knowledge entity. Furthermore, define $u_i =< na_i, po_i, ad_i >$, where $\boldsymbol{na}_i$ is the i^{th} user's name, $\boldsymbol{po}_i$ is the position in organization, e.g. staff, manager, or CEO, and $\boldsymbol{ad}_i$ is the user's email address.

Definition 2. $\mathbf{C}$ is a set of distinct concepts and terms in the thesaurus model of MAKES, and c_i denotes the i^{th} distinct concept. Furthermore, $c_i =< co_i, sy_i, kd_i >$ is defined, where co_i is the i^{th} distinct concept, sy_i is the list of synonym terms of the i^{th} distinct concept, and kd_i is a list of knowledge domains which this distinct concept belongs to.

It is interesting to note that a concept or term belongs only to one distinct connotation which is denoted by a designated word. For example, “bike” and “bicycle” are the same thing. Also, multiple languages can be used in MAKES, the Chinese phrases of “单车” and “脚踏车” are the same thing as “bike”. All of these words/phrases are synonyms, and the word “bike” is chosen as the controlled vocabulary of these synonyms.

In the database of the thesaurus model for MAKES, every distinct concept is given a serial number by which it is entered. The serial number of distinct concepts is very important in the construction of a multi-dimensional eigenvector of messages, as shown below:

Definition 3. $\mathbf{M_i} = (\mathbf{Sd_i}'', \mathbf{Rv_i}'', \mathbf{F_i})$ is the eigenvector of the ith piece of message, where

$\mathbf{Sd_i}'' = \{sd_i'' | sd_i'' \in \mathbf{U}\}$,

$\mathbf{Rv_i}'' = \{rv_{ij}'' \mid rv_{ij}'' \in \mathbf{U}\}$, where j is a natural number, and

$\boldsymbol{F_i} =< f_{i\boldsymbol{1}}, f_{i2}, ..., f_{iN} >$, where f_{ij} is the weighting value of occurrence number of the j^{th} distinct concept of the i^{th} piece of message, and N is the amount of distinct concepts in the thesaurus model.

It is interesting to note that f_{ij} is not a simple occurrence number of the j^{th} distinct concept in the i^{th} piece of message. The value of f_{ij} is the weighting sum of the number of occurrence of the j^{th} distinct concept according to the location of the appearance of the j^{th} distinct concept in the i^{th} piece of message, namely,

$$f_{ij} = ft_{ij} \times w_T + fb_{ij} \times w_B \tag{4.1}$$

where

ft_{ij} is the number of occurrences of the j^{th} distinct concept in the field of the subject of the i^{th} message, and fb_{ij} is the number of occurrences of the j^{th} distinct concept in the field of body of the i^{th} message, and w_T, w_B are the weightings of fields of subject and body in the message. Generally, the content of subject is more important than the body, so w_T is bigger than w_B. Typically, $w_T = 4$ and $w_B = 1$.

The value of the frequency of occurrence of concepts must be normalized. The eigenvector of the message is changed to $\mathbf{M_i} = (\mathbf{Sd_i}, \mathbf{Rv_i}, \mathbf{E_i})$, where $\mathbf{E_i} = < \mathbf{e}_{i1}, \mathbf{e}_{i2}, ..., \mathbf{e}_{iN} >$, and

$$\mathbf{e}_{ij} = \frac{\mathbf{f_{ij}}}{\sum_{\mathbf{k=1}}^{\mathbf{N}} \mathbf{f_{ik}}} \tag{4.2}$$

In the eigenvector $\boldsymbol{M_i}$, the part of $\mathbf{Sd_i''}$ denotes the sender of message, $\mathbf{Rv_{ij}''}$ contains the list of receivers, and $\boldsymbol{E_i}$ is the main content part of the message, so the $\boldsymbol{N}$ dimensional vector $\mathbf{E_i} = < \mathbf{e}_{i1}, \mathbf{e}_{i2}, ..., \mathbf{e}_{iN} >$ could be thought of as the eigenvector of the message in most application circumstances of MAKES. In the present study, $\mathbf{m_i} = < \mathbf{e_{i1}}, \mathbf{e_{i2}}, ..., \mathbf{e_{iN}} >$ is used to denote the eigenvector of the i^{th} message if there is no special statement.

4.4 Automatic classification

Automatic classification of messages is one of main tasks of the concept elicitation and maintenance module. Dynamic classification enables users to view all possible categories of information, and gives them the capability and tools to view, cross-correlate, and match the categories. Any email messages could be tagged and attached to multiple categories.

Classifying a message means a message should be put into the appropriate category or categories. MAKES classifies a message into a certain category through computing the similarity function between the two eigenvectors of the message and the category. The eigenvector of a category in a taxonomy is usually constructed and stored in the thesaurus model.

To compute the similarity between the i^{th} message and the j^{th} category, let $\mathbf{d_j} =< d_{j1}, d_{j2}, ..., d_{jN} >$ as the eigenvector of the j^{th} category in taxonomy, where d_{jk} is the normalization value of occurrence number of the i^{th} distinct concept in the category.

To decide whether the i^{th} message belongs to the j^{th} category or not, the similarity between the message and category should be computed. Let $m_i =< e_{i1}, e_{i2}, ..., e_{iN} >$ be the eigenvector of the i^{th} message. The formula of Cosine similarity computing is given as follow,

$$sim(m_i, d_j) = \frac{\sum_{k=1}^{N} e_{ik} * d_{jk}}{| m_i | * | d_j |} = (e_{i1} * \frac{d_{j1}}{|d|} + ... + e_{iN} * \frac{d_{jN}}{|d|}) / | m_i | \qquad (4.3)$$

where, | d_j | is the length of the vector d_j, namely, $| d_j |= \sqrt{d_{j1}^2 + ... + d_{jN}^2}$, and | m_i |is the length of the vector m_i, namely, $| m_i |= \sqrt{e_{i1}^2 + ... + e_{iN}^2}$. If the value of $sim(m_i, d_j)$ is bigger than the threshold which is pre-set according to experience and testing, the i^{th} message could be classified into the j^{th} category in taxonomy.

4.5 Building thesaurus model

4.5.1 Learning cycle of thesaurus model

A thesaurus model is the knowledge base of MAKES which contains the controlled vocabularies, synonyms of concepts, stop words, and the hierarchies and relationships among the words. MAKES develops some functions to input the controlled vocabularies, synonyms of concepts, stop words, and to build the hierarchies and relationships among the words.

First, the taxonomy is built by human effort. To automate the indexing and classifying of messages, a supervised learning method is adopted. The word "learning" indicates that the taxonomy should be built with the kind of features which could describe the property and content of the taxonomy so that messages could be classified automatically according to the features. The word "supervised" means the learning process is supervised by a human being.

To match the similarity computing algorithm, every category of a taxonomy should be built with an eigenvector. There are three main steps for building eigenvectors of a taxonomy. Authorized personnel must:

i) select some messages as learning samples for every category,

ii) build an eigenvector for every sample message using the method in Section 4.5 and

iii) synthesize these eigenvectors of sample messages and construct a single eigenvector for every category.

Definition 4. The eigenvector of the i^{th} category is defined as $d_i = < d_{i1}, d_{i2}, ..., d_{in} >$, where

$$d_{ij} = \frac{\sum_{k=1}^{M} f_{i,j,k}}{\sum_{k=1}^{M} \sum_{j=1}^{N} f_{i,j,k}} \tag{4.4}$$

Where $f_{i,j,k}$ is the number of occurrences of the j^{th} concept in the k^{th} sample message for the i^{th} category of taxonomy, M is the total amount of sample messages of i^{th} category, and N is the total number of distinct concepts.

The supervised learning process helps to enhance the accuracy and the efficiency of the extracted concepts.

4.5.2 Automatic maintenance of the thesaurus model

The world is developing continuously. New concepts occur as time goes by. The thesaurus model should be updated continuously. Adding new concepts by human effort is one approach to maintain the thesaurus model. MAKES, however, adopts a rule-based approach using artificial intelligence to update the model.

The aim of maintaining the thesaurus model is to find new concepts and place them into an appropriate location of the thesaurus model. In the present study, new words are added to the controlled vocabulary not only by human beings but also by several rules which are embedded in the thesaurus model. The popularity and intensity of the new concept are the indictors which decide whether the concept should be added into the thesaurus model or not.

The popularity of the concept is measured by the number of people who are sharing the same concept among the group, while the intensity of the concept is measured by the frequency with which the concept appears in unstructured information over a certain period of time. So every concept in the thesaurus model is maintained by a 2-tuple $< p_i, f_i >$, where p_i is the number of people who are using the i^{th} distinct concept, and f_i is the frequency with which the concept appeared in the messages during the last 2 months. If the popularity p_i and intensity f_i of the new concept achieves a certain threshold, it is suggested for revision and retention in the thesaurus model.

4.6 Producing a dynamic taxonomy

A dynamic taxonomy is part of the knowledge repository in MAKES. There are three patterns of dynamic taxonomies. One is based on the user's role while the second one is based on the properties of the search keyword. These two patterns focus mainly on how to classify logically a large amount of information. The third pattern is based on mining and analyzing information by using Artificial Intelligence (AI) technologies to produce a new taxonomy when the amount of information increases with time.

4.6.1 Pattern 1: a dynamic taxonomy based on user identification

In any information system, the management of user access rights is an important function to allow valid access and prohibit any invalid access according to the users' role, their positions, department, and other identifications. The users' viewing of the information differs according to their role.

For example, Figure 4.3 gives three kinds of taxonomies according to the position of the user in a company, staff (Figure 4.3a), manager (Figure 4.3b), or CEO (Figure 4.3c). As managers are responsible for supervising staff, there are staff categories in the manager's taxonomy. Similarly, the CEO's taxonomy has departmental categories.

A dynamic taxonomy configured according to a user's identification is considered as a personal interface. After the user logs in, the identification information of the user is retrieved from the staff list in the database. While he/she browses the email messages which are classified into corresponding categories such as, staff, manager, or CEO. This kind of classification only provides those categories which are useful for the role of the user.

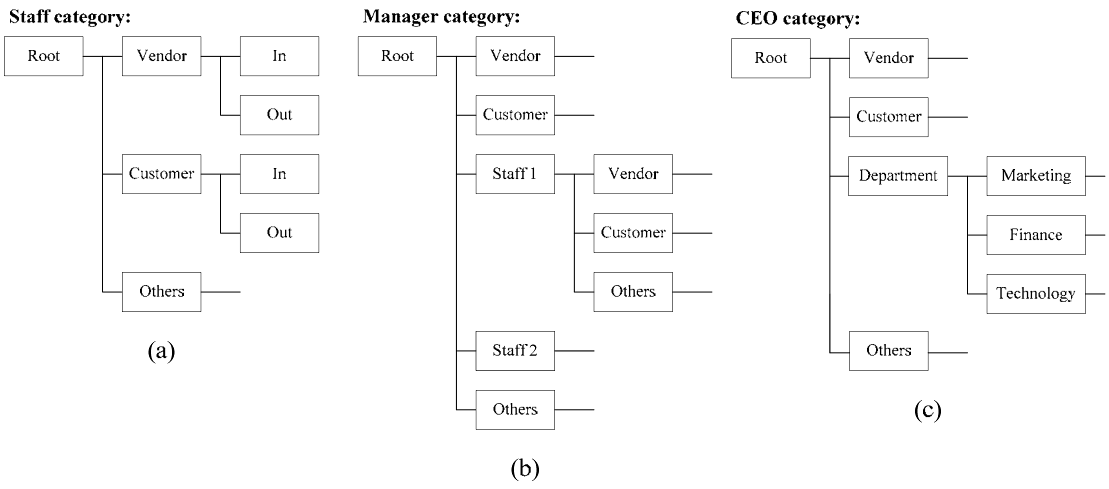
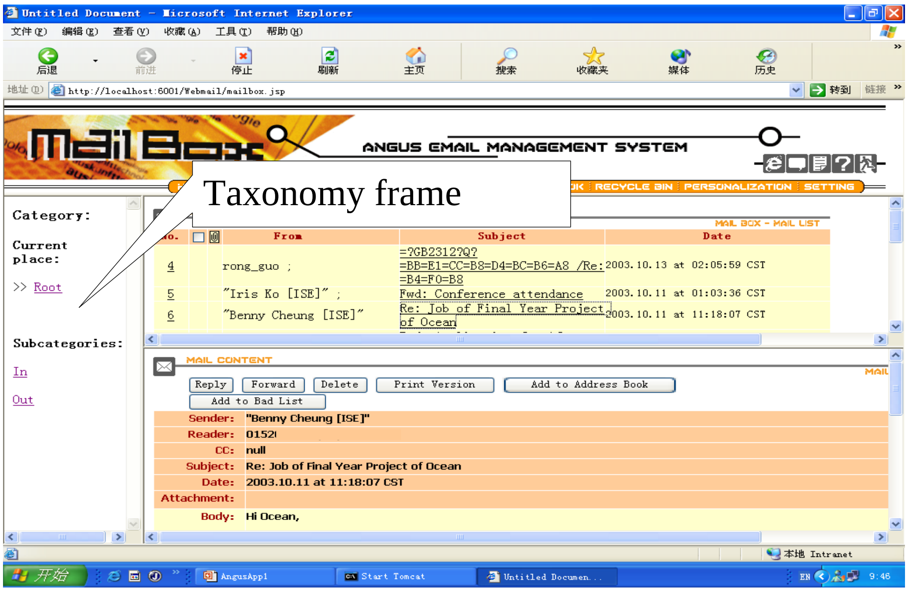


Figure 4.3 Examples of taxonomies associated with three kinds of users.

In this kind of dynamic taxonomy pattern, the messages are filtered by the values of the fields of the sender and the receiver of the message. In the eigenvector of the message, $\mathbf{M}=(\mathbf{Sd},\mathbf{Rv},\mathbf{E})$, $\mathbf{Sd}$ and $\mathbf{Rv}$ are the fields to be filtered. For example, the role of user in the field $\mathbf{Sd}$ can be found.

MAKES is designed as a web-based system to be plugged into the original email system. It receives emails from the email server and stores them in the centralized dynamic taxonomies. Staff members can then browse the entire dynamic taxonomy by using the browsers from their personal computers. It is easy to share information and to quickly search, and acquire useful information. Meanwhile, the sending email function is available so that staff can use MAKES to execute their email-based knowledge work instead of Outlook. Figure 4.4 provides the interface with a personal category according to the user's role.



Category:

Current place:

>> Root

Subcategories:

In

Out

No.	□	From	Subject	Date
4		rong_guo ;	=?GB2312?Q? =BB=E1=CC=B8=D4=BC=B6=A8 /Re: =B4=F0=B8	2003.10.13 at 02:05:59 CST
5		"Iris Ko [ISE]" ;	Fwd: Conference attendance	2003.10.11 at 01:03:36 CST
6		"Benny Cheung [ISE]"	Re: Job of Final Year Project of Ocean	2003.10.11 at 11:18:07 CST

Sender:	"Benny Cheung [ISE]"
Reader:	0152(
CC:	null
Subject:	Re: Job of Final Year Project of Ocean
Date:	2003.10.11 at 11:18:07 CST
Attachment:	
Body:	Hi Ocean,

Figure 4.4 An interface for browsing email messages.

4.6.2 Pattern 2: a dynamic taxonomy based on property of keyword

Another pattern of dynamic taxonomy is produced by the property of search keywords. Search results are displayed in a set of folders that are semantically correlated to the keywords inputted by the user. This also means that more taxonomies are less seen as static entities but more as highly dynamic and evolving structures driven by the user's needs and behavior. For example, to search for the word “quotation” means the user wants to get information related to marketing. The marketing category is then used to classify the search result. Similarly, “whitepaper” of product may mean the user wants to get some technical information about a product, etc. Figure 4.5 provides two examples of taxonomies for the two fields of marketing, and technology.

This is a useful pattern of classification by logically organizing information and providing appropriate categories according to search keywords. It differs from the search function in Google.com (www.google.com) which lists numerous matched results to find out useful information. Some tables of controlled vocabulary and lexicons are dependent on this pattern of dynamic taxonomy. When the search keywords are entered, the system retrieves the meanings and properties of them from some tables, deduces the user’s purpose and the keywords’ field, and then, forms the corresponding categories.

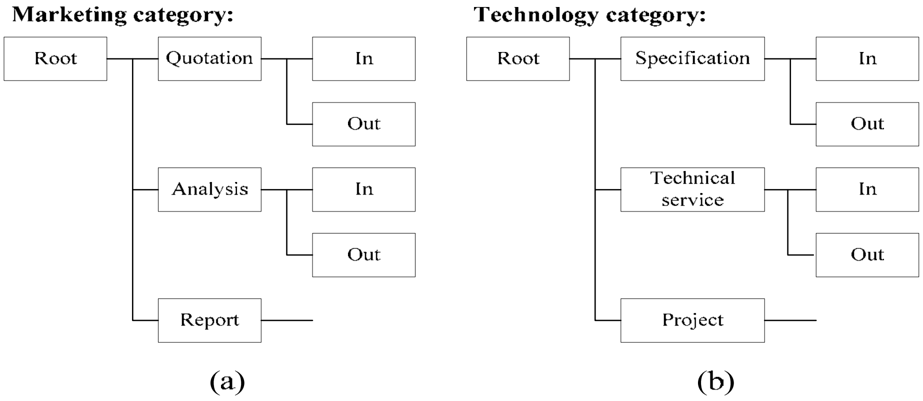
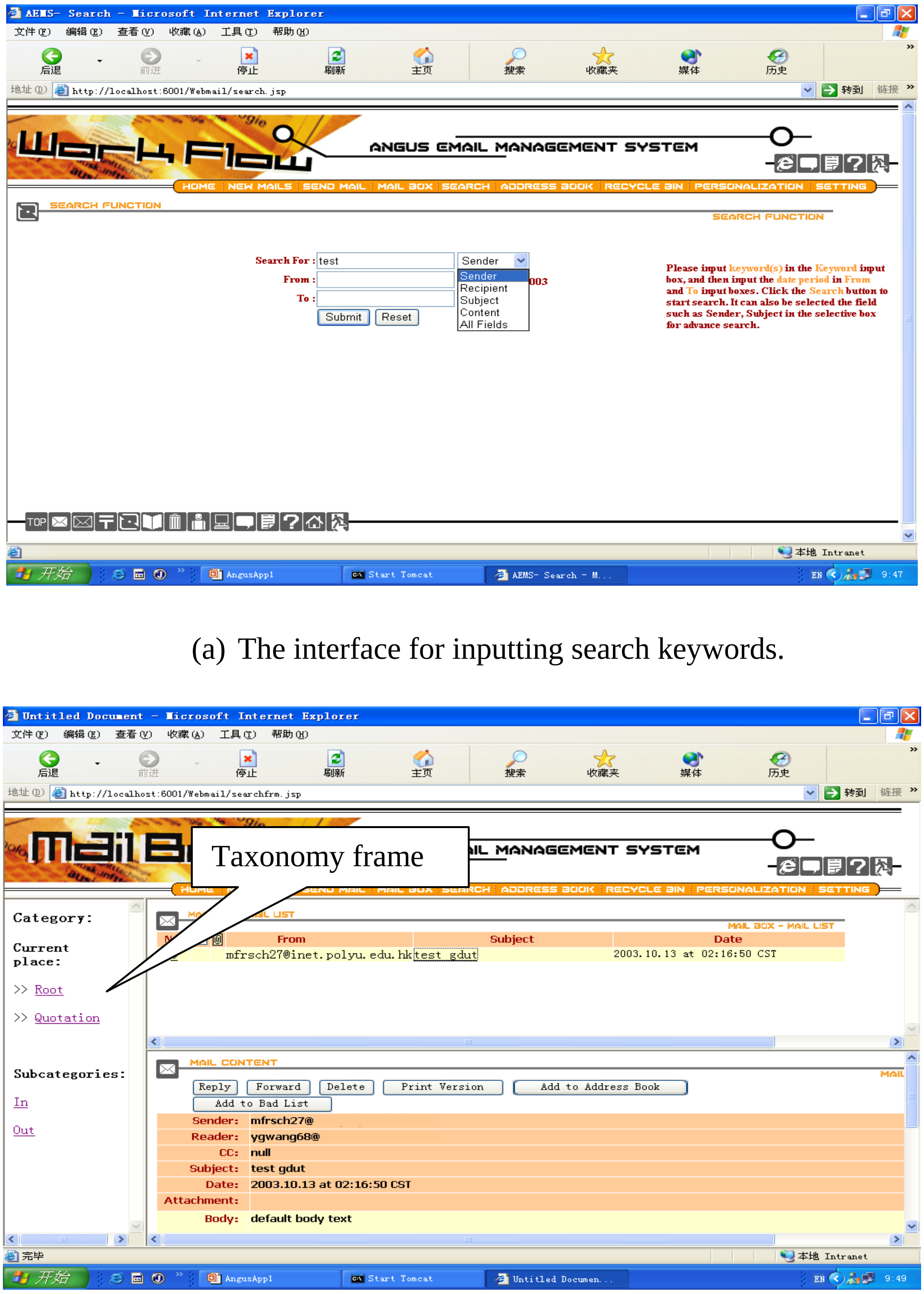


Figure 4.5 Examples of taxonomies in marketing and technical domains.

Figure 4.6a is the interface for inputting searching keywords while Figure 4.6b is the search results interface with a category according to the field associated with the search keywords.

4.6.3 Pattern 3: a dynamic taxonomy based on machine learning

With the increase in knowledge, new terms and relationships between terms are emerging, and new categories should be added and even the taxonomies should be revised. One approach to constructing a dynamic taxonomy is based on machine learning. Figure 4.7 describes the result of a learning cycle for building a dynamic taxonomy.



Search For : test Sender

From :

To :

Please input keyword(s) in the Keyword input box, and then input the date period in From and To input boxes. Click the Search button to start search. It can also be selected the field such as Sender, Subject in the selective box for advance search.

(a) The interface for inputting search keywords.

Category:

Current place:

>> Root

>> Quotation

Subcategories:

In

Out

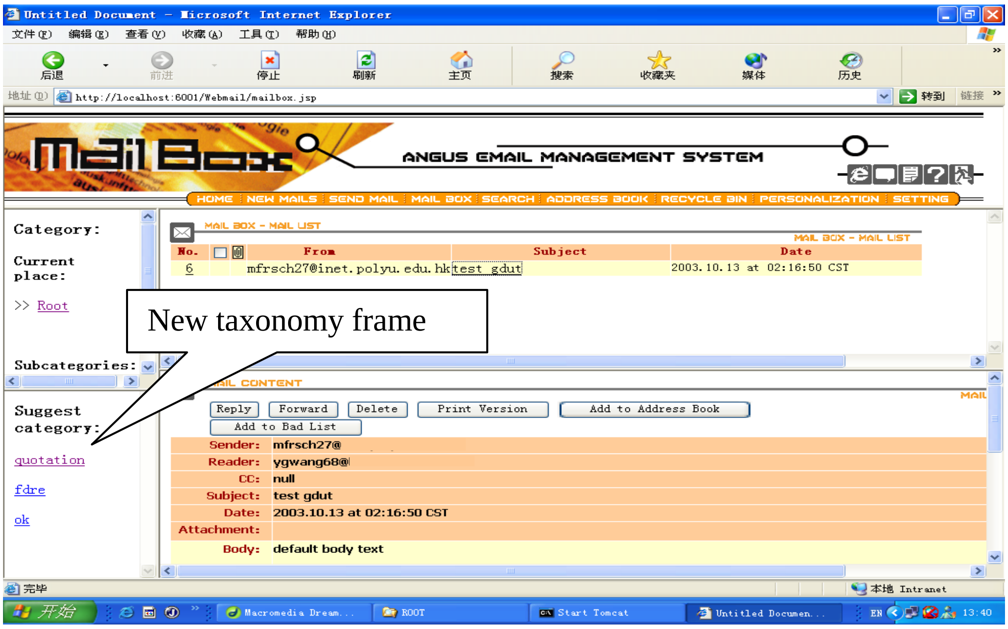
From	Subject	Date
mfrsch27@inet.polyu.edu.hk	test gdut	2003.10.13 at 02:16:50 CST

Sender:	mfrsch27@
Reader:	ygwang68@
CC:	null
Subject:	test gdut
Date:	2003.10.13 at 02:16:50 CST
Attachment:	
Body:	default body text

(b) The search results interface.

Figure 4.6 The dynamic taxonomy in the search function.

Now, the staff of a company can send and receive emails in MAKES. All email messages are stored in a centralized web server which is constructed based on Linux and MySQL. Requested information can be sought and retrieved from the Intranet and the Internet easily and quickly. Dynamic taxonomies are becoming an effective and efficient tool for knowledge management. Based on three patterns of dynamic taxonomy, MAKES provides some logic categories for knowledge workers to browse and search for useful information from email messages logically and conveniently. Based on text mining, a dynamic taxonomy will enable users to construct multiple knowledge models which will provide insights about knowledge assets and knowledge work.



Category:

Current place:

>> Root

Subcategories:

Suggest category:

quotation

fdre

ok

No.		From	Subject	Date
6		mfrsch27@inet.polyu.edu.hk	test gdut	2003.10.13 at 02:16:50 CST

Sender: mfrsch27@

Reader: ygwang68@

CC: null

Subject: test gdut

Date: 2003.10.13 at 02:16:50 CST

Attachment:

Body: default body text

New taxonomy frame

Figure 4.7 The interface for constructing dynamic taxonomy.

Chapter 5. Knowledge Modeling and Multi-faceted Navigation

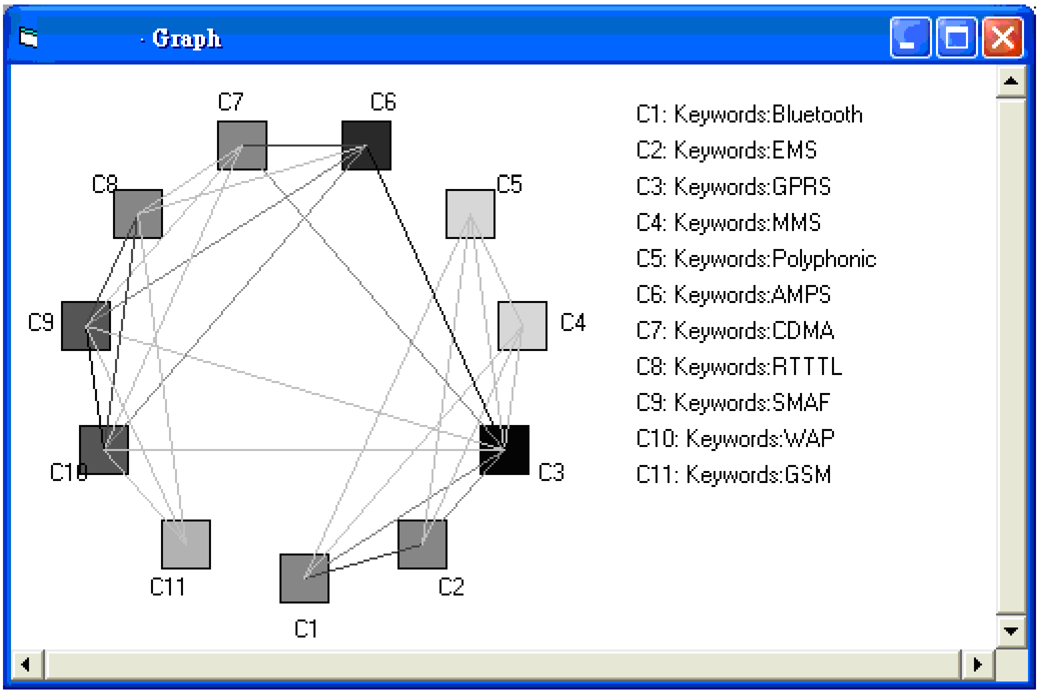
A knowledge entity means the carrier or source of knowledge, including knowledge workers, databases, documents, files, knowledge repository, etc. To manage unstructured information, the relationships among various knowledge entities are represented in various forms of knowledge models. Concept Relationship Model (CRM), Dynamic Knowledge Flow Model (DKFM), and Knowledge Capability Model (KCM) are used in the present study to represent and retain knowledge. Multi-faceted navigation platform enables users to access knowledge from multi-dimensional views. These knowledge models provide an important means for supporting the effective application of unstructured information.

5.1 Concept Relationship Model (CRM)

5.1.1 Pattern of concept relationship

There are several relationships among various concepts. Every relationship between two concepts uncovers how the two concepts are associated and elicits the knowledge about their mutual or interactive influence. In the MAKES, the relationships of concepts are based on indexed unstructured information such as emails and text files. The concept relationships are represented by a Concept

Relationship Model (CRM). A CRM is shown by a simple graph with nodes and edges. The nodes represent concepts which are relevant to a given domain of knowledge and the relationships between them are depicted by the direction of the edges. The importance of the concepts and the relationships between different concepts are indicated by the depth of the coloured background i.e. the deeper the color the greater the importance of the concept. An example of CRM is shown in Figure 5.1.



C1: Keywords:Bluetooth
C2: Keywords:EMS
C3: Keywords:GPRS
C4: Keywords:MMS
C5: Keywords:Polyphonic
C6: Keywords:AMPS
C7: Keywords:CDMA
C8: Keywords:RTTTL
C9: Keywords:SMAF
C10: Keywords:WAP
C11: Keywords:GSM

Figure 5.1 An example of concept relationship model.

As shown in Figure 5.1, Ci denotes the concept in the communications field. For example, $C1$ means bluetooth, $C2$ denotes EMS, etc. The edge shows there is a relationship between the two concepts. There are four edges connecting with $C1$.

They represent that $\boldsymbol{C1}$ is related to $\boldsymbol{C2}$, $\boldsymbol{C3}$, $\boldsymbol{C4}$ and $\boldsymbol{C5}$ which means that the concept of bluetooth is related to the concepts, EMS, GPRS, MMS, and to Polyphonic.

As regards a dynamic taxonomy, the CRM can be used to classify the new explicit knowledge of organizations automatically under the appropriate concepts according to some predefined rules. Its advantage is that it dramatically reduces the time and cost as well as the human effort that is spent in the classification of a huge amount of knowledge assets.

5.1.2 Algorithm for discovering and mapping concept relationships

In the CRM of MAKES, every relationship of two concepts is measured by a set of numeric values, namely support and confidence. Support means the co-occurrence probability of two concepts in a single message, and the confidence is the occurrence probability of the second concept if, and only if, the first concept appears in a single message, i.e. the probability concept depends on another concept. In mathematical terms, a CRM is defined as shown below.

Definition 5.1 *a CRM is represented as 3-tuple* $(\mathbf{C},\mathbf{S},\mathbf{R})$, *where*

$\boldsymbol{C} = (c_1, c_2, ..., c_n)$ *is a set of* n *distinct concepts forming the nodes of a CRM;*

$\boldsymbol{S} = \{(c_i, c_j) \mid \textbf{support}(c_i, c_j) > \lambda, \textbf{and } c_i, c_j \in C, i \neq j\}$ *is a support set of relationships of concepts,* $\textbf{support}(c_i, c_j)$ *is a function that computes the co-occurrence frequency of the concepts of* c_i ***and*** c_j, λ *is a threshold in order to* *make sure that there is a relationship between two concepts when the support value is bigger than* λ *;*

$\boldsymbol{R} = \{ < c_i, c_j, \ r > | \ r = \boldsymbol{confidence}(c_i, c_j) \ \boldsymbol{and} \ (c_i, c_j) \in S, i \neq j \}$ *is a confidence set of relationships of concepts,* ***c****onfidence(*c_i, c_j*) is a function to compute the occurrence probability of* c_j *when* c_i *occurs,* $\boldsymbol{r} = \boldsymbol{c}onfidence(c_i, c_j)$ *denotes the degree of tightness of the relationships of the concepts.*

It is interesting to note that the confidence in CRM is directed. In other words the confidence from concept A to concept B may not be equal to the confidence from concept C to concept A. This situation is described in the following algorithm.

According to the definition of the CRM, this thesis develops an algorithm of Concept Relationship Exploring Technique (CRET). CRET consists of two steps for computing both the support value and the confidence value.

Step 1. Compute the support value of any two concepts to measure the association degree of two concepts.

According to the previous definition given in this thesis, the eigenvector of a message is in the form of a vector as follows,

$$e_i = < e_{i1}, e_{i2}, \ldots, e_{in} > \qquad (5.1)$$

where i means the i^{th} message, e_{ij} is the value of the normal frequency of occurrence of the j^{th} distinct concept in the i^{th} message, and n is the total number of concepts.

In order to decrease the cost of computing the support value, the eigenvector of a message can be simplified as follows,

$$\boldsymbol{e_i'} = < \boldsymbol{e_{i1}'}, \boldsymbol{e_{i2}'}, \ldots, \boldsymbol{e_{in}'} > \tag{5.2}$$

where $e'_{ij} = \boldsymbol{1}$ if $e_{ij} \geq \phi$, and ϕ is a threshold value between 0 and 1, or $e'_{ij} = \boldsymbol{0}$ if $e_{ij} < \phi$, and n is the number of concepts.

The ϕ is an indicator to denote the boundary to measure the importance of a concept in a message. Hence, all of the eigenvectors of messages are re-organized as shown below,

$$\left.\begin{array}{l} \boldsymbol{e_1'} = < \boldsymbol{e_{11}'}, \boldsymbol{e_{12}'}, \ldots, \boldsymbol{e_{1n}'} > \\ \boldsymbol{e_2'} = < \boldsymbol{e_{21}'}, \boldsymbol{e_{22}'}, \ldots, \boldsymbol{e_{2n}'} > \\ \ldots \\ \boldsymbol{e_m'} = < \boldsymbol{e_{m1}'}, \boldsymbol{e_{m2}'}, \ldots, \boldsymbol{e_{mn}'} > \end{array}\right\} \tag{5.3}$$

where, m is the total number of messages and n is the total number of distinct concepts.

The order for computing the support value between any two concepts in the set of concepts is: first compute the support value between the first concept and the second concept, and then compute the support value between the second concept and the third one which follows the second concept in the list of concepts in the thesaurus, and so on.

The formula for computing the support value of the i^{th} concept c_i and j^{th} concept c_j is expressed by:

$$\boldsymbol{support(}c_i,c_j\boldsymbol{)}=\frac{\sum_{k=1}^{m}e'_{ik}e'_{jk}}{m} \tag{5.4}$$

where, m is the number of messages, c_i and c_j belong to the set of concepts and $i \neq j$.

Step 2. The confidence value between the two concepts determines which are to be depended with each other according to the results of step one.

Based on the Equation (5.3), the confidence value from the i^{th} concept to the j^{th} concept is given as follows:

$$\boldsymbol{confidence(}c_i,c_j\boldsymbol{)}=\frac{\sum_{k=1}^{m}e'_{ik}e'_{jk}}{m_i} \tag{5.5}$$

where, m_i is the number of messages containing the i^{th} concept c_i,

and

$$\boldsymbol{confidence(}c_j,c_i\boldsymbol{)}=\frac{\sum_{k=1}^{m}e'_{ik}e'_{jk}}{m_j} \tag{5.6}$$

where, m_j is the number of messages containing the j^{th} concept c_j.

With the use of Equation (5.5) together with Equation (5.6), the two confidence values may not be the same because the denominators of the two formulas are different. All of the support values and confidence values of the relationship between any two concepts are recorded into the database and are used to construct the graph of CRM.

As shown in Figure 5.2, a CRM graph denotes the relationships among the three Concepts, $c\boldsymbol{1}$, $c\boldsymbol{2}$, and $c\boldsymbol{3}$. The edge denotes that there is a relationship between the two concepts. The values of $p\boldsymbol{1}$, $p\boldsymbol{2}$ and $p\boldsymbol{3}$ denote the confidence between the two concepts.

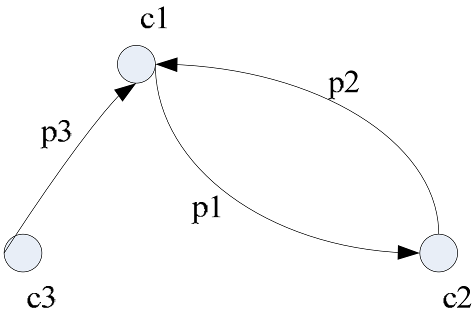


Figure 5.2 A pattern of concept relationships.

5.1.3 Findings from concept relationship

In a graph of CRM, the nodes denote concepts of knowledge. Knowledge and information is collected, represented, organized, and stored in the form of a graph showing the network of concept association. The main discovery from CRM is given below:

i) The importance of concepts of knowledge is determined and the key concepts are identified. In the CRM, the occurrence frequency of every concept is computed and the importance of every concept is measured. As a result, the key concepts of knowledge in knowledge repository are revealed. The tendency of development of this knowledge may be inferred and predicted.

ii) The key concept of knowledge is identified in order to find out the other knowledge that is related to the key concept. In some sense, a graph of CRM can be viewed as the formation of a dynamic taxonomy. Knowledge and information is classified and attached to the nodes in CRM. To handle a problem, the problem can be first analyzed and profiled with some key concepts of knowledge. As soon as the key concepts of knowledge are related to a problem, the corresponding knowledge and information can be retrieved by the CRM model.

iii) The relationships of the concepts are determined in order to enable associational thinking. Associational thinking is the natural pattern of thinking of human beings. During the processing of a problem, a group and cluster of relevant and associated concepts can be retrieved according to the key concepts of the problem. The concepts and their complex association relationships may enlighten the associational thinking and produce a possible solution to resolve the problem.

iv) The associated knowledge is used to automate knowledge discovery and problem-solving. In the CRM, the network pattern of concepts may enable automatic searching and matching of the concepts according to some rules. This is similar to crawling over the Web. This may lead to the development of some intelligent software which can automatically discover the knowledge related to a problem domain, and produce some rational solutions.

In fact, the concepts in CRM include various knowledge entities, such as concepts in the knowledge domain, and sources of knowledge, even persons, teams, and organizations. So, a social network can be produced from the CRM through defining the attribute of nodes as persons, teams, and organizations. In a social network, some key actors and the relationships of communication of actors can be revealed and found out. The key actors and their teams will play an important role in solving problems.

5.2 Dynamic Knowledge Flow Model (DKFM)

A model of knowledge work can support the analysis of knowledge work and enable knowledge auditing. Knowledge workflow is the most important work form in knowledge intensive companies. The effectiveness and efficiency of knowledge flow directly influences the efficiency of all work, the competency of knowledge creation, and the competitive capability of companies.

5.2.1 Pattern of knowledge flow

As knowledge work is complex and has no fixed procedure, knowledge work is usually non-linear. A typical topology of non-linear knowledge work network is shown in Figure 5.3. In a social network, knowledge is dispersed and resides in the minds of the employees. Knowledge work is performed by the members within the social network. It is necessary to exchange, fuse, and innovate knowledge among the participants of a social network through their interaction with each other during knowledge work. This is one of the key research topics of the present study.

A Dynamic Knowledge Flow Model (DKFM) is constructed as a framework to support the analysis of knowledge work both qualitatively and quantitatively so that the best patterns of knowledge work can be found. The term “dynamic” in DKFM means that the dynamic knowledge flow model can be built through the process of automatically extracting and mining useful knowledge from knowledge work. DKFM also helps to analyze the patterns of knowledge work in real-time and online.

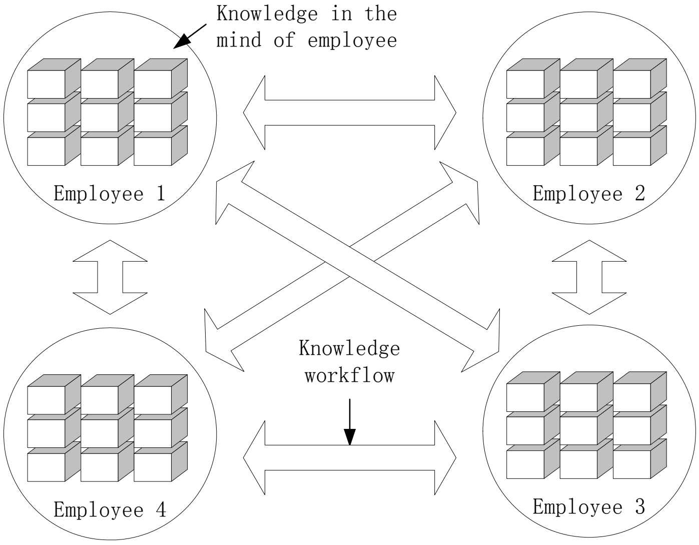


Figure 5.3 The topology of a non-linear knowledge work network.

The DKFM is composed of two parts which are nodes and lines. Generally, nodes in DKFM denote knowledge entities while lines denote the relationships among knowledge resources. Knowledge entity in this thesis means the carrier or source of knowledge, including knowledge workers, databases, documents, files, knowledge repositories, etc. It is interesting to note that employees are a very important kind of knowledge resource because they can offer tacit knowledge which is retained in their minds. They also have the ability to create new knowledge. Relationships among knowledge resources are established through communication and interaction. These relationships denote the communication that takes place during ordinary office work, querying, finding, searching, answering, discussing, transferring information, etc. Figure 5.4 gives an example of DKFM, where the nodes denote employees and are given numbers, and the relationship edges are denoted by numbers too. For example, the edge e[i] possesses some properties, e.g. knowledge domains about which two employees discuss through correspondence.

Basically, the DKFM provides a framework for the analysis of knowledge work patterns. Users can display the patterns of knowledge work and execute multidimensional analysis including slicing, up and down, etc., through operating the control panel in a DKFM system. For example, one could choose the financial knowledge domain and show the pattern of knowledge work related to finance. The term of “subject” denotes a domain of knowledge and is usually featured as a set of terms which are semantically related to the subject. Terms are words or phrases which have meaning and are not ambiguous. Here the subject of finance includes some key terms such as sales, order, and marketing etc. Since the relationship element e[i] in DKFM contains one property of knowledge domains, a new pattern of knowledge work about finance could be obtained as shown in Figure 5.4 by hiding these relationship edges which do not contain the knowledge domain related to finance.

By comparing Figures 5.4 and 5.5, it is interesting to note that there is a difference between them in that some nodes and lines shown in Figure 5.4 do not appear in Figure 5.5, such as Staff 5 and Staff 7 and their relationship lines. This means these two staff members do not discuss the subject of finance. This implies that they might be technicians who do not handle any transactions related to finance, such as order processing.

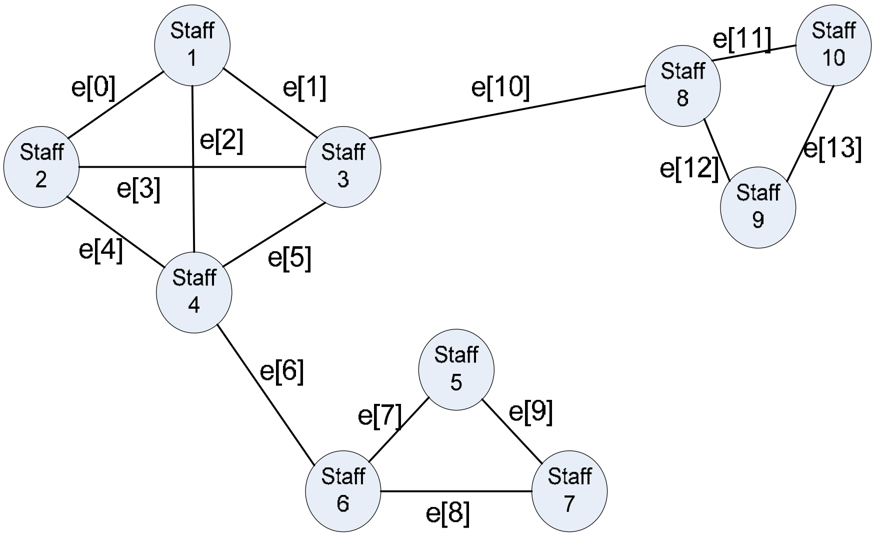


Figure 5.4 An example of dynamic knowledge flow model.

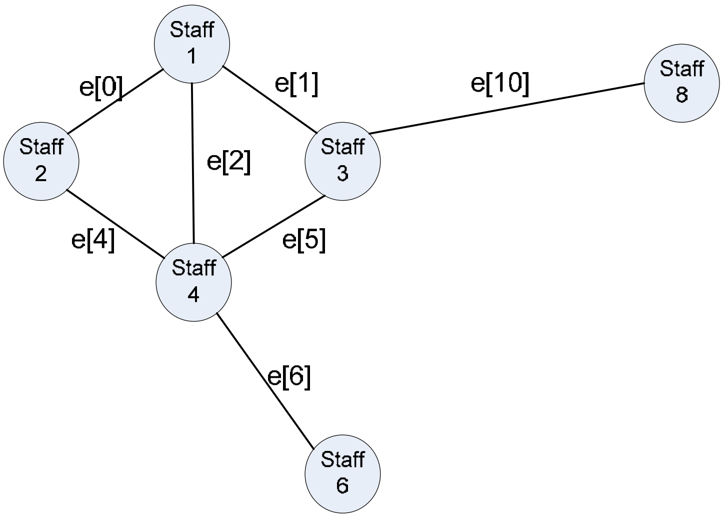


Figure 5.5 A pattern of DKFM about two subjects of sales and marketing.

5.2.2 Schema of constructing DKFM

Currently, the DKFM is composed of two phases: text mining, and DKFM modeling and analyzing. The first phase is text mining, through mining text messages from knowledge work. The second phase of modeling and analyzing of the DKFM adopts visualized technologies to construct the DKFM and shows the diagram of patterns of knowledge flow through statistics, and analyzing based on the knowledge taxonomy as shown in Figure 5.5. The schema of DKFM is an automatic and intellectual process in which some AI technologies are used, such as product rules, a database of knowledge rules, pattern recognition, similarity computing, etc.

In the phase of constructing and analyzing the DKFM, the eigenvectors of text documents are reformed and stored in the special database, or data warehouse, in order to operate the DKFM efficiently and effectively. The control panel in the DKFM system is the main operating interface for users to set conditions for the analysis of the DKFM. These conditions are fed into the module of statistics which make up of the feature vectors. The module for constructing and visualizing the DKFM is then executed and the pattern of the DKFM is generated. The module of list text documents is used for showing the list of the text documents which are attached to a line in the pattern of DKFM. In general, the objective of DKFM is to enable analyzing and improving knowledge work effectively and efficiently.

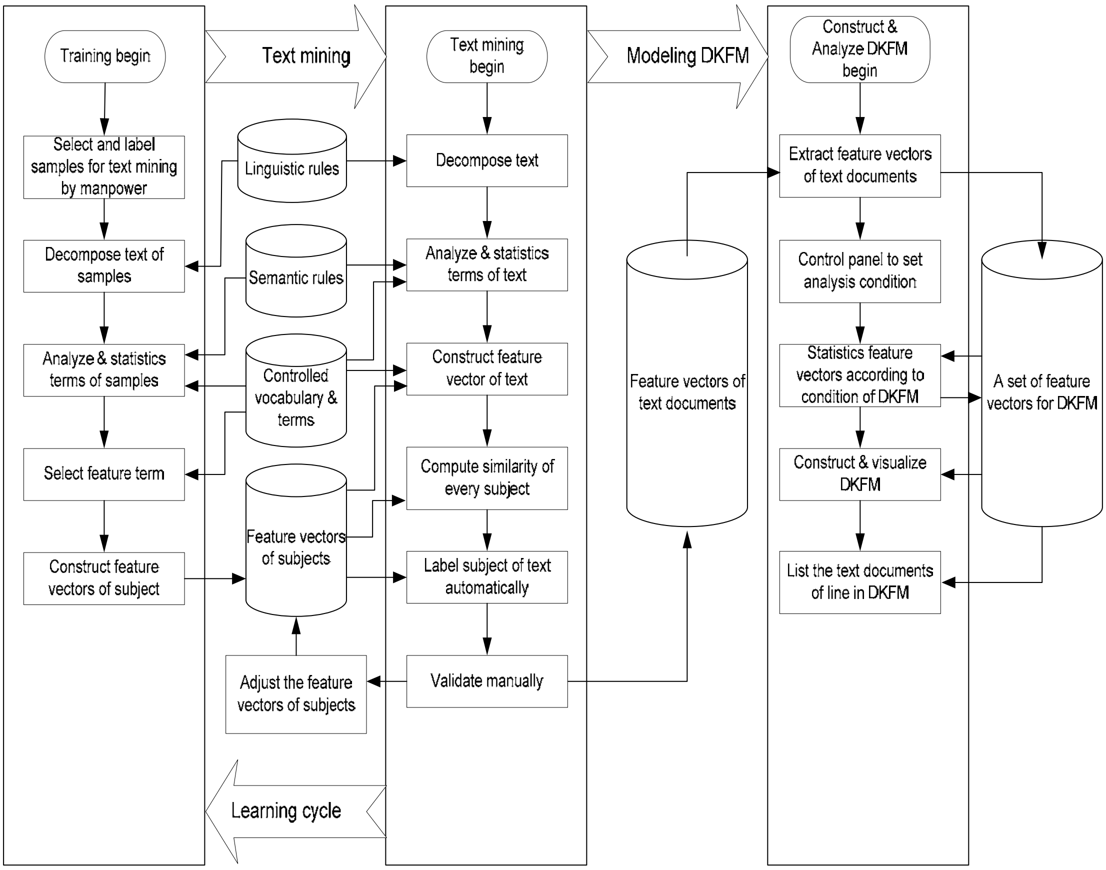

Training begin
Text mining
Text mining begin
Modeling DKFM
Construct & Analyze DKFM begin
Select and label samples for text mining by manpower
Linguistic rules
Decompose text
Extract feature vectors of text documents
Decompose text of samples
Semantic rules
Analyze & statistics terms of text
Control panel to set analysis condition
Analyze & statistics terms of samples
Controlled vocabulary & terms
Construct feature vector of text
Feature vectors of text documents
Statistics feature vectors according to condition of DKFM
A set of feature vectors for DKFM
Select feature term
Feature vectors of subjects
Compute similarity of every subject
Construct & visualize DKFM
Construct feature vectors of subject
Label subject of text automatically
List the text documents of line in DKFM
Adjust the feature vectors of subjects
Validate manually
Learning cycle

Figure 5.6 The schema for constructing DKFM.

5.2.3 Algorithm of DKFM

To construct a model for knowledge flow, it is necessary to trace the flow of a message and record the sender who sends the message and the receivers who receive the message. An eigenvector of a message can be denoted a 3-tuple as below，

$$\boldsymbol{M}_i = (\boldsymbol{Sd}_i, \boldsymbol{Rv}_i, \boldsymbol{F}_i) \qquad （5.7）$$

where, i means the i^{th} message,

Sd_i is the sender,

Rv_i is a set of the receivers, and

$\boldsymbol{F_i} =< \boldsymbol{f_{i1}}, \boldsymbol{f_{i2}},..., \boldsymbol{f_{in}} >$ is the eigenvector of the i^{th} message and f_{ij} is the frequency of occurrence of the j^{th} distinct entity in this message.

Sender and receivers are called knowledge entities in the present study, since they are the sources of knowledge. The knowledge entities mean persons, teams, even any organizations. They are the nodes in the model of knowledge flow.

In a DKFM, the knowledge entities also focus on the relationships about the knowledge domains which the messages belong to. In MAKES, every distinct entity is belonged to one knowledge domain. As a result, the eigenvector of the message is re-mapped to a new 3-tuple

$$\boldsymbol{M_i'} = (\boldsymbol{Sd_i}, \boldsymbol{Rv_i}, \boldsymbol{Kd_i}) \tag{5.8}$$

where, i means the i^{th} message,

Sd_i is the sender of the message,

Rv_i is a set of receivers of the message,

$\boldsymbol{Kd}_i =< \boldsymbol{kd_{i1}}, \boldsymbol{kd_{i2}},..., \boldsymbol{kd_{im}} >$ is an eigenvector about knowledge domains, where m is the total number of knowledge domains, kd_{ij} is the sum of frequency of occurrence of the distinct concepts which belong to the j^{th} knowledge domains.

According to Formula 5.7, the relationships of knowledge entities are described by the set of concepts. A graph of DKFM based on the relationship of entities can be constructed. Based on Formula 5.8, the relationships of knowledge entities are abstracted by knowledge domains. A pattern of DKFM based on the relationships of knowledge domains is built.

5.2.4 Analysis of knowledge flow model

In the DKFM, the nodes denote knowledge entities. Information and knowledge is acquired and analyzed in order to generate relationships among knowledge entities. DKFM uncovers the pattern of knowledge flow among knowledge entities. The main output of DKFM includes the capability:

i) To identify the knowledge entities. There are a variety of knowledge entities in DKFM such as personnel, teams, and organization. The role every knowledge entity plays and his/her work can be observed and revealed from DKFM.

ii) To study the communication between knowledge entities. A DKFM reveals what the pattern of knowledge flow is, how knowledge entities communicate with each other, and what knowledge concepts knowledge entities deal with. It can enable people to find out and better understand the role which every knowledge entity plays in knowledge work. This can be seen from the position where every knowledge entity is in the pattern of the structure of the knowledge flow.

iii) To find out the pattern of knowledge flow according to the criteria of knowledge domains. The pattern of the knowledge domain in the knowledge flow enables people to observe a model of knowledge flow on the upper level since a knowledge domain is more general than a knowledge concept. In this way, more conceptual knowledge may be found in this kind of model of knowledge flow. For example, more knowledge entities are found in the department of finance if the knowledge transferred among them is mostly in the financial domain.

iv) To find out some rational and irrational patterns of knowledge work by analyzing the structure of knowledge flow. The structure of the graph of the network of DKFM contains more or less the main characteristics of knowledge work. Some specific sub-structures in the DKFM may uncover some advantages and shortcomings of knowledge work. For example, a longer linear structure of knowledge flow may denote there are many knowledge entities taking part in the same knowledge work, and the knowledge entities are working one by one sequentially. This situation may mean there is some redundancy in this knowledge work.

v) To uncover some teams which work collaboratively well. A graph of DKFM shows there are many interactions among some knowledge entities. It may mean these knowledge entities work well together and rely on each other. These knowledge entities may constitute a good work group.

vi) Automate re-engineering of knowledge work. It is possible to analyze the structure of the network of DKFM automatically because many network structures can be recognized using the mathematical methods of graph theory and reasoning machines and technologies of artificial intelligence. For example, a linear structure or circular structure can be discovered using a reasoning algorithm.

5.3 Knowledge Capability Model (KCM)

5.3.1 A radar graph of knowledge capability

Based on the model of knowledge flow, this thesis proposes a Knowledge

Capability Model (KCM) to evaluate the capability of knowledge work of every knowledge entity. The KCM mainly gathers the information and knowledge from knowledge work, analyzes the information which belongs to knowledge domains, and measures the workload of every knowledge entity, where the workload means the amount of correspondence the knowledge entity handles in every distinct knowledge domain. Finally, the knowledge domains are ranked according to the measures of workload of every knowledge domain of the knowledge entity, and the knowledge capability model of the knowledge entity is constructed.

Knowledge domain is a relative concept, namely, it can mean a bigger knowledge area but can also mean a smaller professional range. For example, the knowledge domain “display technology” includes CRT , LED, and etc, but the knowledge domain can be defined as a concrete technical range such as CDMA, BLUETOOTH, and so on. In practice, knowledge domain is a general level above the knowledge concept.

This thesis adopts a radar graph to express multidimensional measurements. Figure 5.7 shows an example of a knowledge capability model of a knowledge entity. KD_i denotes the i^{th} knowledge domain. The black dot on a certain axis of a knowledge domain represents the capability measurement with which the knowledge entity works on the related knowledge domain.

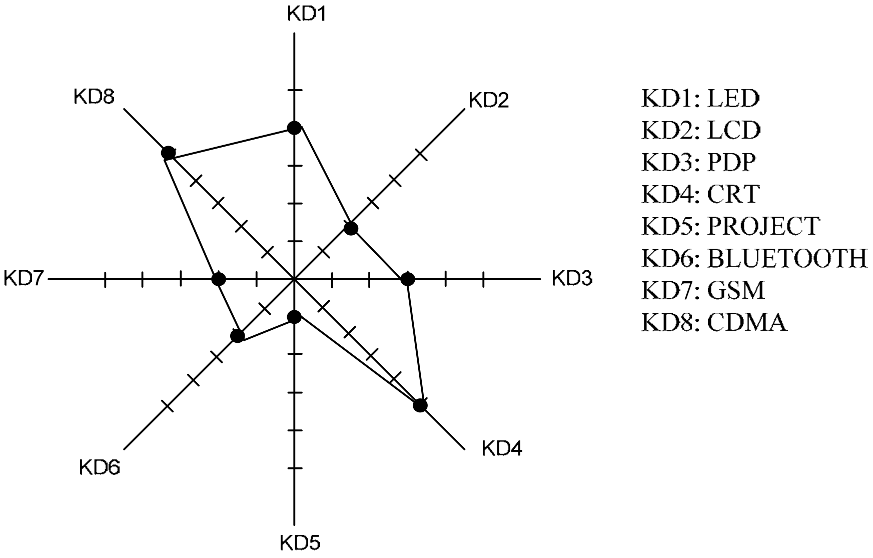


Figure 5.7 A radar graph of knowledge capability model of a knowledge entity.

This radar graph shows a capability pattern of knowledge domains which a certain knowledge entity uses and applies in knowledge work. As shown in Figure 5.7, the domain which is focused on by staff is CDMA. It means the user may be a specialist in communicating about CDMA. A radar graph may uncover the knowledge distribution which this knowledge entity is engaged in and can help to find the specialists in knowledge domains.

5.3.2 Measurement algorithm of knowledge capability

To analyze the knowledge domains which a knowledge entity works on, the knowledge domains must be extracted from the information obtained from the knowledge entity. With the use of the thesaurus model of MAKES, the entities captured from the information can be mapped into related knowledge domains. The measurement algorithm of a knowledge capability model contains the following steps:

Step 1. Classifying the unstructured information according to a knowledge entity.

The messages of the unstructured information of a knowledge entity include the messages sent or received by the knowledge entity. The eigenvector of the message of knowledge entity is defined as $\boldsymbol{M_{i,j}} =< \boldsymbol{c_{i,j1}}, \boldsymbol{c_{i,j2}}, ..., \boldsymbol{c_{i,jn}} >$, where, i means the i^{th} knowledge entity, j denotes the j^{th} message, n is the number of knowledge concepts, and $c_{i,jk}$ is the frequency of the j^{th} knowledge entity.

Step 2. Building the eigenvector of the message of knowledge entity based on knowledge domains.

Transform and construct the eigenvector of the message with knowledge domains as $\boldsymbol{M'_{i,j}} =< \boldsymbol{kd_{i,j1}}, \boldsymbol{kd_{i,j2}}, ..., \boldsymbol{kd_{i,jm}} >$, where, i means the i^{th} knowledge entity, j denotes the j^{th} message, $kd_{i,jk}$ is the frequency of the k^{th} knowledge domain, and m is the number of knowledge domains.

Step 3. Normalizing the eigenvector of the message.

Rebuild the eigenvector of the message based on normalization in the form $\boldsymbol{M''_{i,j}} =< \boldsymbol{kd'_{i,j1}}, \boldsymbol{kd'_{i,j2}}, ..., \boldsymbol{kd'_{i,jm}} >$, where,

$$\boldsymbol{kd'_{i,j}} = \frac{\boldsymbol{kd_{i,j}}}{\sum_{k=1}^{m} \boldsymbol{kd_{i,jk}}} \tag{5.9}$$

Step 4. Computing the frequency of every knowledge domain and draw the radar graph of the KCM.

The frequency of messages about every knowledge domain of a knowledge entity is computed first. Then, the ranking of the knowledge domains is determined according to the frequency of interaction with the knowledge domain. Next, select the top five knowledge domains with the highest frequencies to constitute a radar graph of the knowledge capability of this knowledge entity.

5.3.3 Explanation of knowledge capability

A knowledge capability model reflects the distribution of knowledge domains in the knowledge entity. The main output of KCM is the capability to:

i) Identify the highlighted knowledge domains. A KCM can find out which knowledge domain is most up-to-date and hot. The model can display the changes that take place in the highlighted knowledge domains and can forecast the development of knowledge and technology by comparing the different KCM in different periods.

ii) Identify the knowledge domains which a knowledge entity focuses on and find out an appropriate knowledge entity. The knowledge domains which a knowledge entity puts emphasis on may mean that the knowledge domains are the most familiar and most focused on areas which the knowledge entity works on. According to the knowledge domains, an appropriate knowledge entity could be selected to perform related knowledge work.

iii) Identify the knowledge domains which a knowledge entity needs to reinforce. Some knowledge domains are evaluated as being of lower rank. This may mean that they may be the weak spots in the knowledge work of the knowledge entity. It is necessary for the knowledge entity to study more knowledge about these knowledge domains.

iv) Evaluate the learning capability of knowledge entities. Through comparing the different knowledge capability models at different periods, the learning capability of every knowledge entity may be inferred.

5.4 Multi-faceted navigation for unstructured information

5.4.1 Multidimensional framework of unstructured information

For managing and applying information, different users need different knowledge since different users apply knowledge to achieve different objectives. For example, the Department of Research & Development (R&D) needs to acquire technical information and knowledge. In contrast, the department of Human Resource Management focuses on collecting information about the performance of staff, and so on. It is necessary to build a flexible and dynamic mechanism which offers multidimensional views for knowledge classification in order to manage a great amount of information and knowledge to satisfy the demands of various applications.

In the present study, a multi-faceted navigation system is developed for the management of unstructured information. It supports various processes for UIM such as classifying, organizing, searching, and retrieving unstructured information from a multi-dimensional perspective. As shown in Figure 5.8, MAKES displays a concept relationship model and a social network model. In the knowledge repository, the concept relationship model uncovers the strength of association relationships of the knowledge concepts, and the social network model expresses the degree of communication presented in the relationships of knowledge entities.

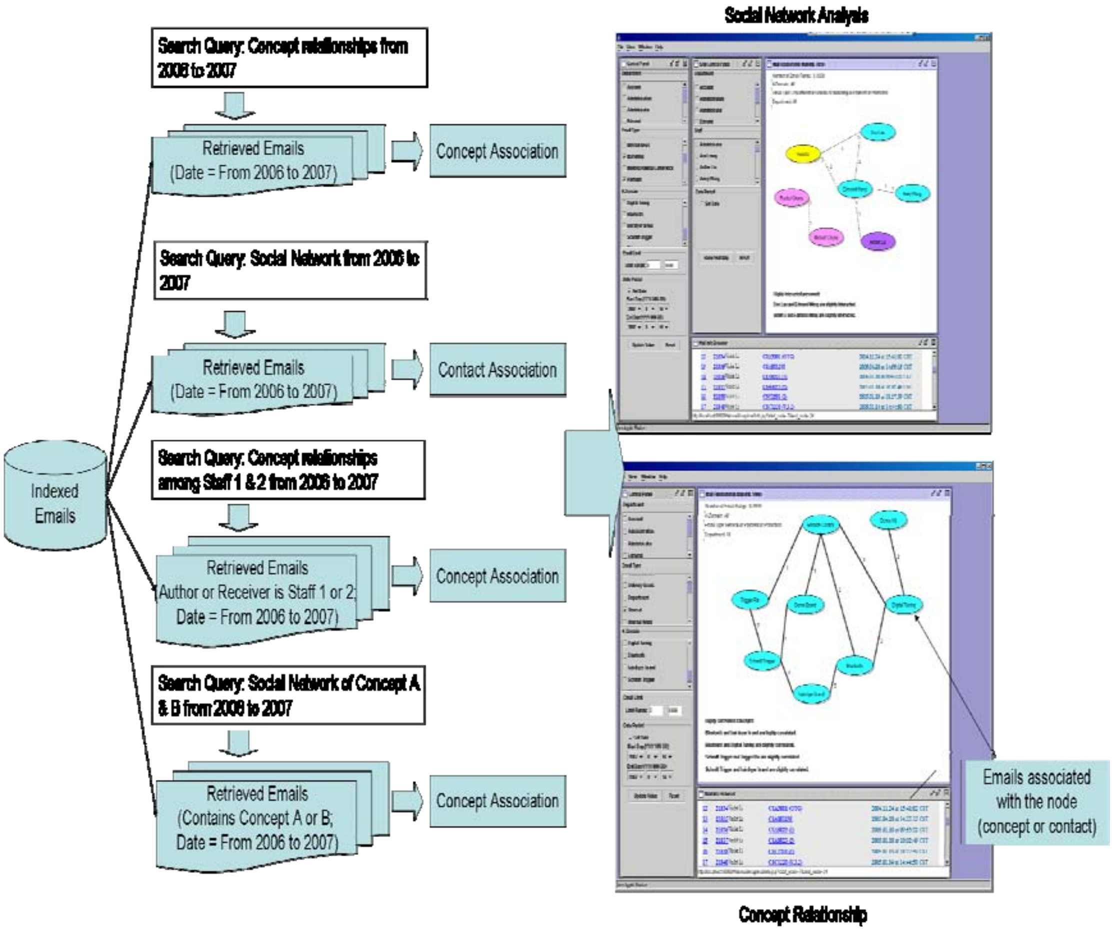

Social Network Analysis
Search Query: Concept relationships from 2006 to 2007
Retrieved Emails (Date = From 2006 to 2007)
Concept Association
Search Query: Social Network from 2006 to 2007
Retrieved Emails (Date = From 2006 to 2007)
Contact Association
Indexed Emails
Search Query: Concept relationships among Staff 1 & 2 from 2006 to 2007
Retrieved Emails Author or Receiver is Staff 1 or 2; Date = From 2006 to 2007)
Concept Association
Search Query: Social Network of Concept A & B from 2006 to 2007
Retrieved Emails (Contains Concept A or B; Date = From 2006 to 2007)
Concept Association
Emails associated with the node (concept or contact)
Concept Relationship

Figure 5.8 A pattern of multi-faceted navigation of unstructured information.

Multi-faceted navigation includes building a multidimensional space of information and providing a pattern for information management. The multidimensional space for information provides the classification and measurement of information which is scaled using different dimensions according to the characteristics of the information. The definition of multiple dimensions can be pre-defined according to what is necessary for the application and management of information. Figure 5.9 expresses a 3-dimensional space, including the dimension of knowledge entity, knowledge domain, and time. In the dimension of knowledge entity, several scales are used: individual, workgroup, department, organization, etc.

The dimension of knowledge domain consists of various knowledge domains, such as CDMA, Bluetooth, and so on. The dimension of time provides the scale of time, e.g. 2006, 2007, and 2008. In fact, the example of Figure 5.9 is a technical view because the knowledge domains concern technology.

Multi-faceted navigation enables a user to navigate in the multidimensional space of information management and select the viewpoint to browse the knowledge models. At first, the viewpoint is mapped into every dimension and the related scales of various dimensions are combined to construct the search conditions in the information space. Then, a flexible multi-faceted information browsing and retrieving function is performed according to the search criteria.

A control panel is designed in the MAKES for multi-faceted navigation. A searching condition is composed of the user input for the options in the multiple choice frames of the control panel. Then, the related knowledge models are built. The multi-faceted navigation provides useful data, information, models, and reasoning in order to analyze the knowledge work in organizations.

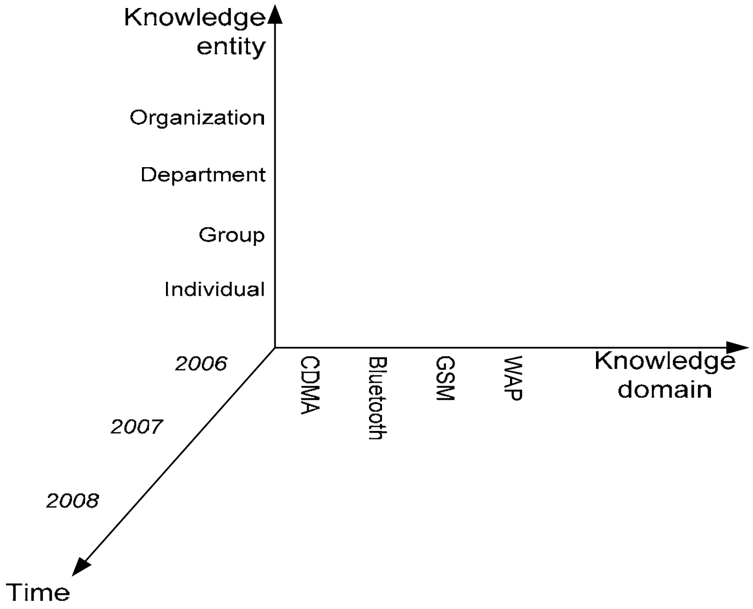


Figure 5.9 A multidimensional coordinate for unstructured information management.

5.4.2 Assessment of knowledge elicitation

After social network and concept relationships are elicited, they are evaluated and measured based on a rule-based analysis. Figure 5.10 shows the criteria for the classification of the concepts elicited from the concept relationships and social network. Basically, the elicited concepts are classified according to two attributes “Importance” and “Permeability” into four categories which are: critical concept, focus concept, normal concept and abandon concept. Importance refers to the number of pieces of unstructured information which contain the elicited concept. The greater the number of pieces of unstructured information is, the greater the importance of the elicited concept is. Permeability refers to the number of knowledge workers who used the concept elicited from the unstructured information. The higher the number of knowledge workers, the greater the permeability. The results are automatically determined by a rule-based inference engine and are reported in a knowledge inventory report. The average number of pieces of unstructured information and the average number of knowledge workers who used the unstructured information are used as indicative criteria for the classification of the elicited concepts.

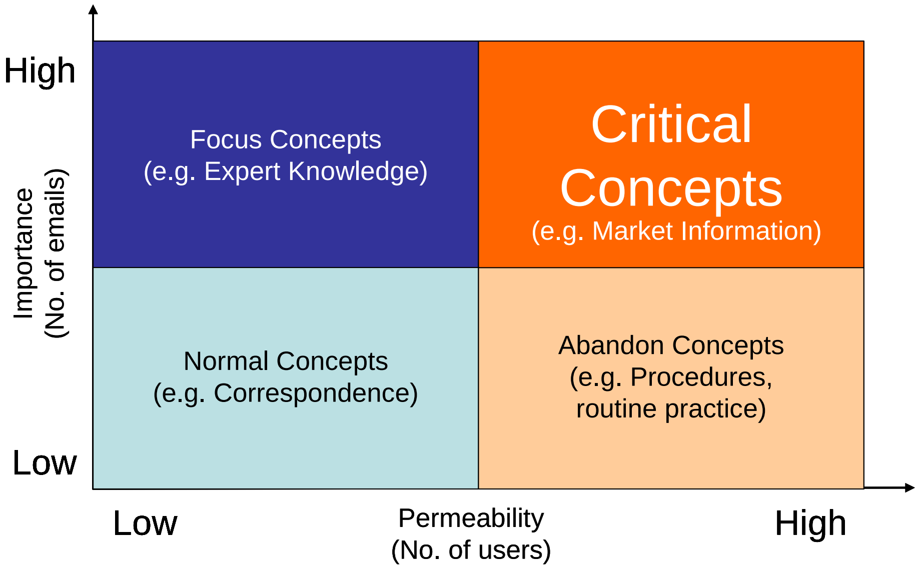


Figure 5.10 A 2-dimensional space for classification of elicited concepts.

On the other hand, the results of a social network analysis are used at the level of individuals, departments or organizations to identify what roles the teams and individuals play in the organizations. As shown in Figure 5.11, the knowledge workers are classified according to the two attributes such as “Knowledgeable” and “Impact” into four categories which are: normal user, focus user, critical user, and abandon user, respectively. Knowledgeable refers to the number of identified concepts provided by the knowledge worker. The higher the number of identified concepts is, the more knowledgeable the worker is. Impact refers to the number of users who cite the identified concept provided by the knowledge worker. The higher the number of citations is, the greater the impact is. The results are automatically determined and reported in a critical user report. The average number of concepts and the average number of citations are used as indicative criteria for the classification of the knowledge workers.

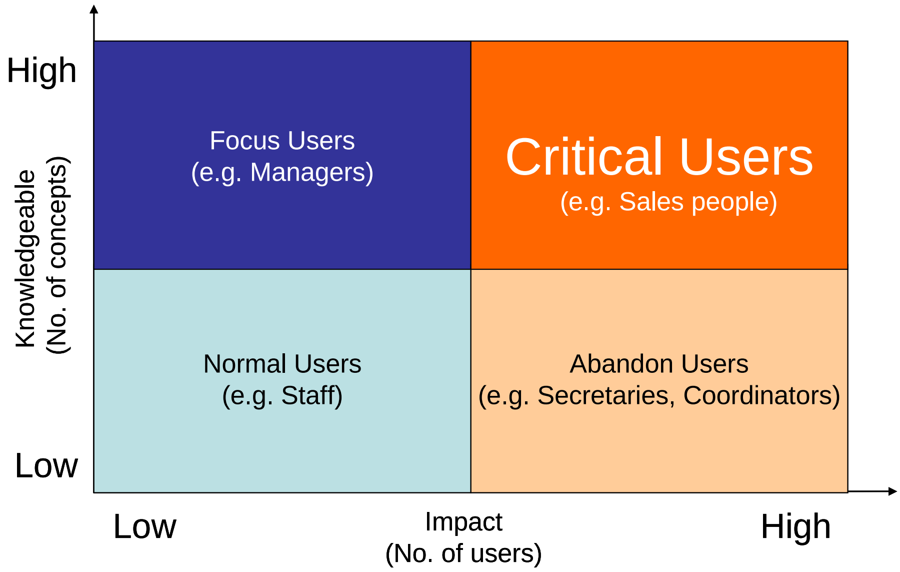


Figure 5.11 A 2-dimensional space for classification of knowledge users.

5.5 Generation of reports for knowledge work

Although the structure of the graph is simple, the graphical knowledge models are not very definite or easy to understand, based only on the construction of knowledge models. Hence, text reports are needed to explain and clearly narrate the knowledge assets and their relationships from the knowledge models in natural language. The main output reports derived from MAKES are described in the following sections.

5.5.1 Knowledge concept report

CRM is made up of concepts and their relationships. The key concepts are identified and ranked automatically according to some pre-defined rules from CRM. Then, the report of critical concepts is formulated. In the critical concept report, the concept field records, the term of the concept, the field of the number of users denotes the average number of occurrences of the concept which occur in every user's emails, and the field of the number of emails which the concept occurs in. The importance of every concept is measured and classified in the field of concept type in the concept list. A sample structure of a critical concept report is shown in Table 5.1.

Table 5.1 Critical concept report.

No.	Concept	Number of user	Number of message	Concept type
1	c_1	u_1	e_1	t_1
2	c_2	u_2	e_2	t_2
…				
n	c_n	u_n	e_n	t_n

In the critical concept report, n denotes the total number of concepts, c_i means the term of the i^{th} concept, u_i is the average number which is the total number of occurrences of the i^{th} concept in all messages divided by the number of users, e_i is the average number of messages which is the total number of occurrences of the i^{th} concept in all messages divided by the total number of messages, and t_i denotes the rating of importance of the i^{th} concept.

To rate the importance of the i^{th} concept, some rules are pre-defined. The importance of the i^{th} concept is assessed and measured by both values of the number of users and number of messages. The concept types are classified by focus, critical, normal, and abandon, respectively. The concrete rating rules are:

- Focus – the value of the number of users is lower than ζ, where ζ is a threshold while the number of messages of the i^{th} concept is higher than υ, where υ is a threshold;
- Critical – both values of the number of users and the number of messages containing the i^{th} concept are higher than υ, where υ is a threshold;
- Abandon –both values of the number of users and the number of messages containing the i^{th} concept are smaller than τ, where τ is a threshold;
- Normal – both values of the number of users and the number of messages containing the i^{th} concept are middle, namely, the concepts belong to the type of normal when they are neither ‘critical’ nor ‘abandon’.

Based on the critical concept reports, two reports of critical concepts in two different periods can be compared. The differences and changes between the two lists of critical concepts in the two critical concept reports indicate the evolving critical knowledge in the technology development. The change of importance of a critical concept may mean that the concept is attracting more attention or being abandoned.

An agent is developed to compare two lists of critical concepts in two years in the recent past, e.g. 2006 and 2008. The change of sorting order of every concept in the list of critical concept reports is tagged with up or down arrows which denote ascending or descending order of the concepts. The main rules for inferring the trends in technology development are defined as follows:

Rule 1: if the sorted position of a concept in the list of current critical concepts is higher than that in the previous year, the concept is tagged with an up arrow in the tendency list of technology development;

Rule 2: if the sorted position of a concept in the list of current critical concepts is lower than in the previous year, the concept is tagged with a down arrow in the tendency list of technology development.

The report of tendency of technology development is shown below.

Table 5.2 Report of tendency of technology development.

No.	Concept	Current position	Position last year	Change type
1	c_1			
2	c_2			
…				

N	c_n			

The field of current position means the ranked sequence of number of the i^{th} concept in the critical concept report of the current year. The field of position last year denotes the ranked sequence number of the i^{th} concept in the critical concept report of the previous year. The comparison between the two values decides the trend of development of the i^{th} concept. The up or down arrows are used in the field of change type to tag the ascending or descending tendency in which more or less attention needs to be paid to the i^{th} concept.

For example, LED has gained more and more attention in the last two or three years as shown by the increase of occurrence frequency of this concept in messages in knowledge work.

5.5.2 Knowledge inventory report

DKFM uncovers the pattern of flow of knowledge and information in knowledge work. The components of DKFM are knowledge entities and their relationships. Knowledge entities include knowledge workers and other artifacts which are the knowledge sources. The knowledge inventory can be derived from the knowledge flow in knowledge work. Key knowledge sources and actors can be identified and assessed quantitatively according to the analysis of the structures of DKFM.

The knowledge inventory report has four main fields. The knowledge entity field registers the name and address of the knowledge entity. Number of messages sent records the total number of messages which are sent by the knowledge entity, and the number of messages received is the total number of messages which are received by the knowledge entity. The importance of every knowledge entity is measured and classified. The structure of knowledge inventory report is shown in Table 5.3.

Table 5.3 Knowledge inventory report.

No	Knowledge entity	Number of messages sent	Number of messages received	Type
1	en_1	sd_1	rv_1	t_1
2	en_2	sd_2	rv_2	t_2
…				
n	en_n	sd_n	rv_n	t_n

In the knowledge inventory report, n denotes the total number of knowledge entities, en_i means the term of the i^{th} knowledge entity, sd_i is the total number of messages which are sent by the i^{th} knowledge entity, rv_i is the total number of messages which are received by the i^{th} knowledge entity , and t_i denotes the rating of importance of the i^{th} knowledge entity.

To rate the importance of the i^{th} knowledge entity, some rules are pre-defined. The importance of the i^{th} knowledge entity is assessed and measured by both values of the number of messages sent and received. The grades of types of knowledge entity are classified as focus, critical, normal, and abandon. The concrete rating rules are:

- Focus – the value of the number of messages sent by the i^{th} knowledge entity is lower than ζ, where ζ is a threshold while the number of messages received by the i^{th} knowledge entity is higher than υ, where υ is a threshold, or the other way round;
- Critical – both values of the number of sent and received messages of the i^{th} knowledge entity are higher than υ, where υ is a threshold;
- Abandon –both values of the number of sent and received messages of the i^{th} knowledge entity are smaller than τ, where τ is a threshold;
- Normal – both values of the number of sent and received messages of the i^{th} knowledge entity are moderate.

From the knowledge inventory report, the key knowledge sources and knowledge entity are identified.

5.5.3 Critical user report

Since knowledge workers or organizations are regarded as one of the many kinds of knowledge entities, the critical user report can be derived from the DKFM. The key users can be identified. Meanwhile, the relationships among users can be extracted from DKFM as shown in Table 5.4.

Table 5.4 Critical user report.

No	User	Number of sent messages	Number of received message	Type
1	ur_1	sd_1	rv_1	t_1
2	ur_2	sd_2	rv_2	t_2
…				
n	ur_n	sd_n	rv_n	t_n

In the critical user report, n denotes the total number of users, ur_i is the name of the i^{th} user, sd_i is the total number of messages which are sent by the i^{th} user, rv_i is the total number of messages which are received by the i^{th} user , and t_i denotes the rating of the importance of the i^{th} user.

To rate the importance of the i^{th} user, some rules are pre-defined. The importance of the i^{th} user is assessed and measured by both values of the number of sent and received messages. The types of knowledge entity can be classified as focus, critical, normal, and abandon. The concrete rating rules are:

- Focus – the value of the number of sent messages of the i^{th} user is lower than ζ, where ζ is a threshold while the number of received messages of the i^{th} user is higher than υ, where υ is a threshold, or the other way round;
- Critical – both values of the number of sent and received messages of the i^{th} user are higher than υ, where υ is a threshold;

- Abandon –both values of the number of sent and received messages of the i^{th} user are smaller than τ, where τ is a threshold;
- Normal – both values of the number of sent and received messages of the i^{th} user are moderate.

The key users are identified and the bottleneck can be recognized. The report also enables the department of human resource management to focus on improving on the recruitment and deployment of employees.

5.5.4 User's knowledge capability report

For assessing a user more precisely, KCM displays the competence of every knowledge worker. A user's knowledge capability report, which extracts and assesses the knowledge domains to which the user needs to pay attention, finds out the pattern of knowledge capability and learning capability of the user.

The user's knowledge capability report has four main fields. The knowledge domain field registers the name of the knowledge domain. Number of sent messages records the total number of messages which are about the relevant knowledge domain and are sent by the user, and the number of received messages is the total number of messages which are about the relevant knowledge domain and are received by the knowledge entity. The importance of every knowledge domain of the user is measured and classified. The structure of a user's knowledge capability report is shown in Table 5.5.

In the user's knowledge capability report, n denotes the total number of knowledge domains, do_i is the name of the i^{th} knowledge domain, sd_i is the total number of messages which are about the i^{th} knowledge domain and are sent by the user, rv_i is the total number of messages which are about the i^{th} knowledge domain and are received by the user, and t_i denotes the rating of the importance of the i^{th} knowledge domain in the user's knowledge capability.

Table 5.5 A user's knowledge capability report.

No	Knowledge domain	Number of sent messages	Number of received messages	Type
1	do_1	sd_1	rv_1	t_1
2	do_2	sd_2	rv_2	t_2
…				
n	do_n	sd_n	rv_n	t_n

To rate the importance of the i^{th} knowledge domain, some rules are pre-defined. The importance of the i^{th} knowledge domain is assessed and measured by both values of the number of sent and received message. The types of knowledge domains are classified by focus, critical, normal, and abandon. The concrete rating rules are:

- Focus – the value of the number of sent messages about the i^{th} knowledge domain is lower than ζ, where ζ is a threshold while the number of received messages about the i^{th} knowledge domain is higher than υ, where υ is a threshold; or contrariwise;

- Critical – both values of the number of sent and received messages about the i^{th} knowledge domain are higher than υ, where υ is a threshold;
- Abandon –both values of the number of sent and received messages about the i^{th} knowledge domain are smaller than τ, where τ is a threshold;
- Normal – both values of the number of sent and received messages about the i^{th} knowledge domain are of medium frequency.

From the user's knowledge capability report, the pattern of knowledge domains which the user possesses is described. The preponderant knowledge domains of the user are tagged in the report.

The above multiple reports are about knowledge concepts, knowledge inventory, and users. The text reports provide other kinds of forms to elicit useful knowledge and explain the knowledge models. Some intelligent inferring functions can be developed based on these reports.

Chapter 6. Verification Testing in Emergency Management (Case One)

Emergency management has drawn the attention of governments worldwide. There is a lot of information and knowledge involved in emergency management in. In this chapter, MAKES is embedded as an unstructured information management system for trial implementation in a city emergency management system in Guang Zhou and its performance in terms of scalability and compatibility are discussed.

6.1 Emergency management architecture

Natural disasters and accidents occur in our daily life, and the severity of crises grows in scale such as the “9.11” disaster, SARS, avian influenza and the financial tsunami etc. The world is not only suffering from natural disasters and accidents such as earthquakes, floods, typhoons, fires, chemical hazards and epidemics, but also from social-economical and ecological crises: from food safety, pollution, social unrest to terrorist attacks. These risks and crises would not only give rise to huge economic losses, but also losses of people's lives and properties and challenges to the stability of governments. Emergency management has become one of the most important tasks in the administration of every government. Figure 6.1 describes the emergency platform of the State Department of China.

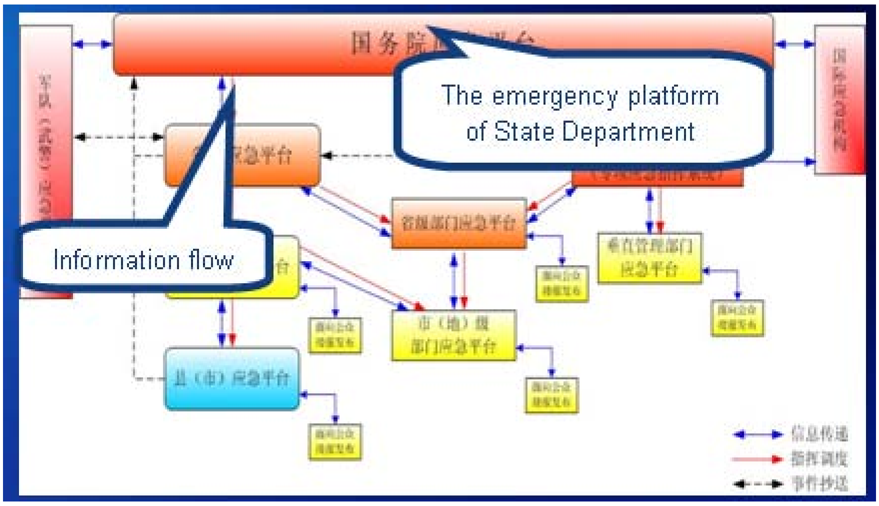


Figure 6.1 The Emergency Platform of State Department of China.

Emergency management includes a series of complex and complicated processes (Figure 6.2). The Emergency Management System is a dynamic system which includes four stages: prevention and emergency preparation, monitoring and early warning, correction and rescue and recovery and reconstruction. Emergency management is also a kind of “system engineering” that can be summarized as “one plan, three mechanisms”. In other words, this means the emergency management involves a response plan and the mechanisms of emergency response, organization and laws. Through the establishment of an early warning signal, emergency response and social mobilization mechanism in the community, a good emergency management system is able to minimize the occurrences of incidents and the corresponding damage. Figure 6.3 gives a model of information flow of emergency management.

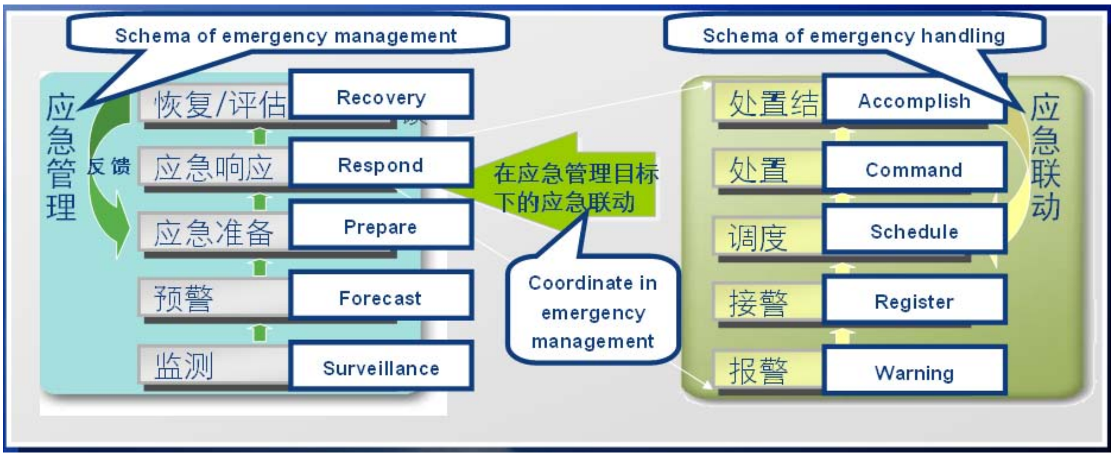


Figure 6.2 The Schema of Emergency Management.

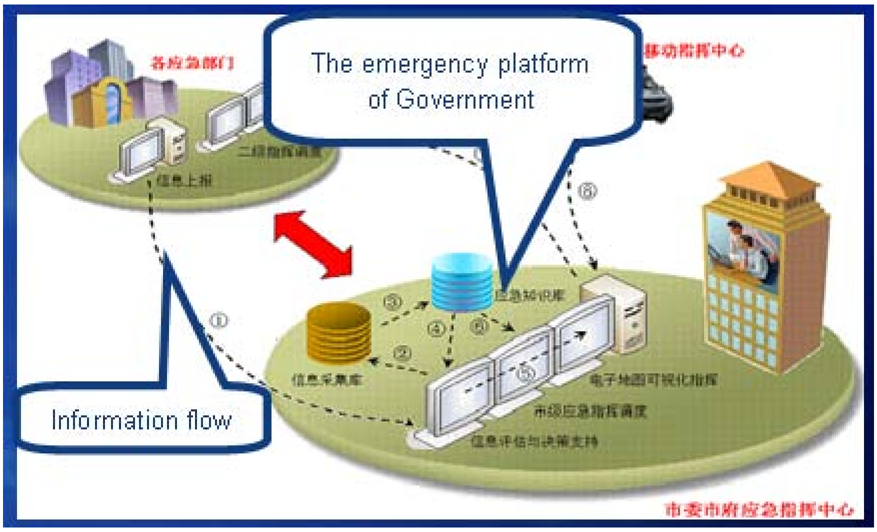


Figure 6.3 The Model of Information flow of Emergency Management.

6.2 Knowledge management in emergency management

An effective emergency management requires the collection, analysis and dissemination of a large amount of information and corresponding knowledge at all stages which include the collection of information monitored, early warning analysis, the simulation and research of trends of the emergency development, as well as intelligent decision-making in the action plan. The problems encountered in emergency management covers a broad area of expertise. Collecting the existing relevant information would create a massive amount of data. The key question is how to search the appropriate information and knowledge quickly and accurately from the massive information and knowledge based on the nature, characteristics, status and situations of the unexpected emergency in order to support the intelligent decision-making process in handling emergencies.

6.3 An emergency management system

The author of this dissertation is the project leader and chief architect, in charge of the organization and implementation of the Guang Zhou City Emergency Management System (GZ-CEMS). The project started in July 2006 and the system was launched in June 2008, and it is applied in The Committee of City Management of Guang Zhou Municipal Government which plays an important positive role in the daily monitoring and disposal of emergencies

The "Guang Zhou City Emergency Management System" is an integrated monitoring, commanding and scheduling system. The system has a city-wide input of on-line and real-time data from various sources, which include remote video monitor, Global Position System (GPS), vehicle tracking, Personal Digital Assistant (PDA) and handheld wireless data terminals. Emergency instructions could also be sent via the voice integration calling system, and various wired and wireless networks connected to all government departments and emergency handling units. In dealing with emergency incidents, the breakout of an incident often leads to another related incidence. For example, in the case of a chemical leakage, this could lead to air pollution. If the pollutants are highly toxic, then an evacuation plan for the occupants in the vicinity will be required. A complex chain of decisions are needed to assess the situations and make appropriate decisions.

6.4 Application of MAKES for emergency management

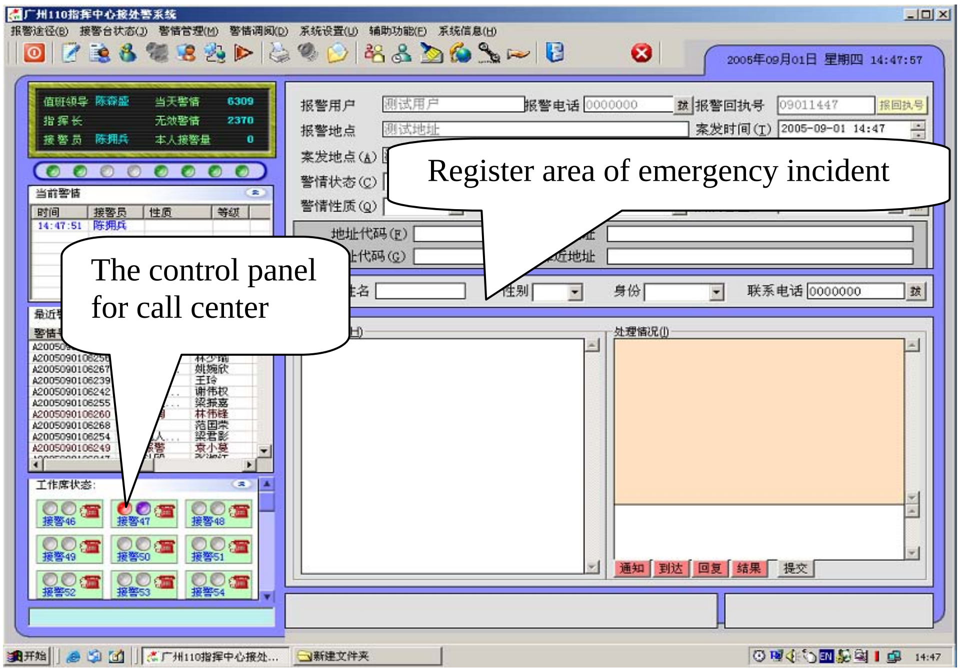
This chapter describes how the author introduced knowledge management to the development of the Guang Zhou City Emergency Management System in which the author was in charge. During the development, an automated classification system of the unstructured information and emergency knowledge based on MAKES has been established. Moreover, the knowledge models for emergency management systems have also been developed.

6.4.1 Background of an emergency incident case

There was an emergency incident of chemical plant explosion. A chemical warehouse chemical explosion occurred in the suburbs in Guang Zhou city. The fire brigade arrived at the scene firstly to execute the rescue task. They first confirmed the explosion combustion types of the chemicals, and then looked for related information through the Guang Zhou City Emergency Management System, and liaise with experts in the field to discuss and determine the chemical characteristics of the explosion and combustion and hence the expert advice was sought for extinguishing the fires etc.

At the same time, environmental experts were identified. They were asked for assessing the effect of the chemical explosion on the air and water quality in the vicinity and their likely spread. A program for prioritizing the steps and measures to be taken was then considered together with the provision plan for the variety of resources, equipment and personnel that are needed.

Figure 6.4 shows a snapshot interface of the GZ-CEMS for registering and handling the emergency incident. Figure 6.5 shows a geographical map which is built based on the electronic monitoring of live video images, with scene showing the dynamic deployment of vehicles and persons.


广州110指挥中心接处警系统
报警途径(B) 接警台状态(J) 警情管理(M) 警情调阅(D) 系统设置(U) 辅助功能(F) 系统信息(H)
Register area of emergency incident
The control panel for call center
当前警情
最近
警情
工作席状态:
接警46
接警47
接警48
接警49
接警50
接警51
接警52
接警53
接警54
处理情况(I)
通知
到达
回复
结果
提交
开始
广州110指挥中心接处...
新建文件夹
14:47

2005年09月01日 星期四 14:47:57

值班领导 陈森盛 当天警情 6309
指挥长 无效警情 2370
接警员 陈拥兵 本人接警量 0

时间	接警员	性质	等级
14:47:51	陈拥兵		

A20050
A2005090106256 林少瑞
A2005090106267 姚婉欣
A2005090106239 王玲
A2005090106242 谢伟权
A2005090106255 梁振嘉
A2005090106260 林伟峰
A2005090106268 范国荣
A2005090106254 梁君影
A2005090106249 袁小慧

报警用户 测试用户 报警电话 0000000 拨 报警回执号 09011447 报回执号
报警地点 测试地址 案发时间(T) 2005-09-01 14:47
案发地点(A)
警情状态(C)
警情性质(Q)
地址代码(F)
址代码(G)
近地址
姓名 性别 身份 联系电话 0000000 拨

Figure 6.4 The interface of registering an emergency incident.

The explosive combustion of chemicals is a very complicated event. It is necessary to collect multiple sources of relevant information and expertise knowledge, as well as the possible consequences of the environmental and social impact. It is worth noting that the program for the fire extinguishing itself may cause environmental pollution, and this should also be taken into account.

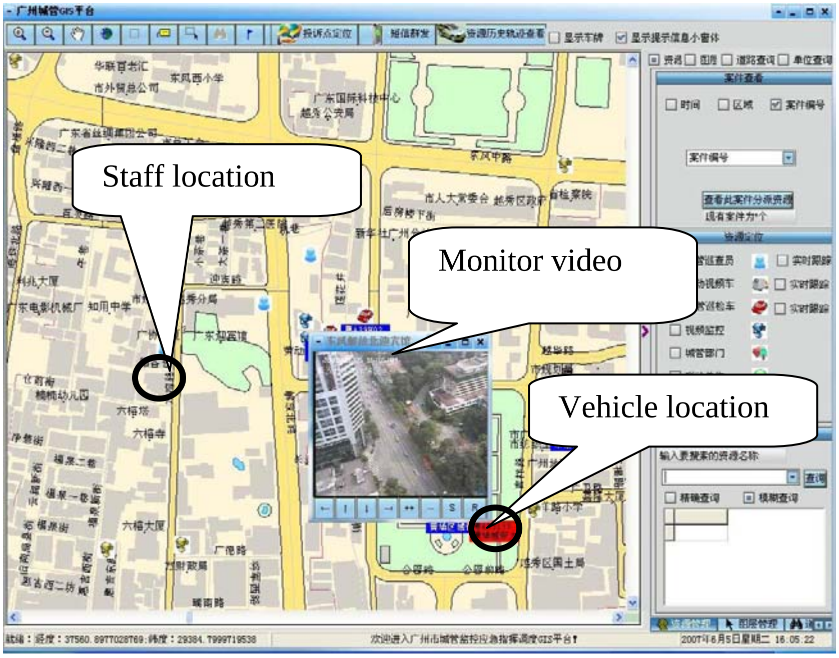


Figure 6.5 A snapshot of the geographic map with various resources for handling emergency incident.

The effectiveness of the emergency management system depends on the useful information the system can gather within a short period of time from multiple sources and the comprehensiveness of the decision-making process. If there is no effective integration of information and expertise knowledge from various functional departments in dealing with the emergency, mistakes in decision-making can easily be made which can lead to serious consequences.

Very often, different kinds of information are stored in different units in different documents and databases, and a lot of information is saved in the text files. It is not easy to locate and retrieve the information and expertise knowledge quickly and accurately because of the lack of effective organization of the information and knowledge that supports the decision-making process. This adversely affects the efficiency and response time in managing the emergency. As a result, it is important to develop and integrate an auxiliary system which processes text information and provide knowledge modeling features. In the development of Guang Zhou City Emergency Management System (GZ-CEMS), the MAKES is integrated into the entire emergency management system as an independent supporting subsystem as shown in Figure 6.6. The main function of MAKES is processing information and knowledge for contingency analysis and modeling. Knowledge models are built to provide different aspects of navigation, and to generate some necessary textual reports which contain useful information and knowledge for decision support. Relevant information and knowledge can be accessed through the MAKES' user operation interface. Since MAKES is based on Bowser / Server architecture, it is relatively easy to be "integrated" or "embedded" to the application system.

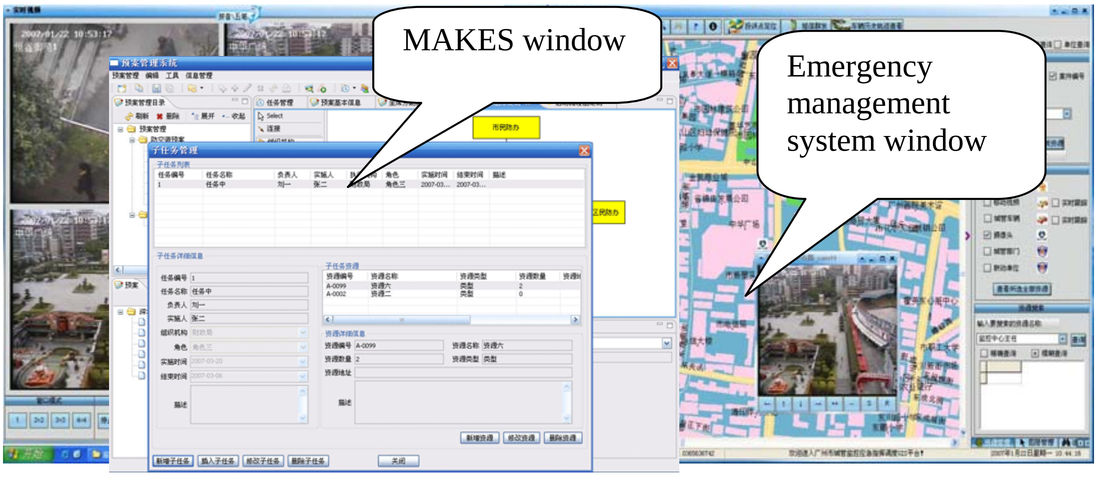


Figure 6.6 The snapshot of MAKES window in GZ-CEMS.

6.4.2 Dynamic taxonomy for emergency information

In an emergency management system, a large amount of emergency related information and knowledge is needed. It is interesting to note that most emergency information and knowledge includes a variety of documents, reports and papers are highly unstructured. In the GZ-CEMS, MAKES is used to build up the dynamic classification mechanism for storing emergency information and knowledge. The processes for building dynamic taxonomy for emergency information are described as below.

Step 1: Collecting emergency information.

During the development of GZ-CEMS, a large amount of information is collected, which includes information about city management, public safety, fire fighting, environmental protection, chemical hazards, etc. The information is mostly unstructured and is mainly stored in text format. Table 6.1 gives a filename list for the information about the emergency management.

Currently, GZ-CEMS has collected about ten thousand files related emergency management. New information is now being collected continuously. A large amount of information and knowledge about emergency management were dispersed in these files. For example, the fire fighting information includes safety knowledge for fire-extinction, as well as some fire-fighting methods using chemicals. A sample of the information is shown in Figure 6.7. Environmental protection knowledge includes environmental pollution caused by various factors as shown in Figure 6.8. In Figure 6.9, chemical product safety protection includes the protection and the safe disposal methods which cover a large number of chemical products.

Step 2: Performing dynamic classification.

Taxonomy is a general approach to manage and classify information. In the traditional way, emergency-related information is categorized manually and a lot of directories are setup to preserve them. However, such a traditional mode of managing information needs a lot of labor work, and is not convenient. Since the amount of emergency information is huge and the relationships of emergency information are complex, dynamic taxonomy is used as a technology that supports the classification, storage of emergency information and the maintenance of the taxonomy of emergency information in GZ-CEMS.

Table 6.1 The list of files about emergency management.

No.	File content	File name
1	Fire safety-knowledge-Fire and burning	xf1.txt
2	Environment safety-knowledge-leakage emergency disposal	hb1.txt
3	Chemical safety-case-Chemical leakage in Sanmenxia city	hx1.txt
…	etc.	…
…	…	…

…

Combustion products refer to all substances generated from the burning or pyrolysis. Combustion products include: combustion-generated gas, energy, tobacco that can be seen, etc. Combustion-generated gas generally refers to: carbon monoxide, hydrogenation of hydrogen, carbon dioxide, acrolein, hydrogen chloride, sulfur dioxide and so on.

Fire statistics show that around 80 percent fire deaths are due to the fire inhalation of toxic products of combustion flue gas. The fire generated a large number of flue gas containing toxic components such as carbon dioxide, HCH, sulfur dioxide, nitrogen dioxide, etc.. Carbon dioxide is one of the main combustion products, and the fire died of carbon monoxide are the main combustion products, and its toxicity is the blood hemoglobin of the high-affinity, its affinity for hemoglobin is 250 times higher than that of oxygen.

…

Figure 6.7 A text segment about fire-fighting.

…

Withdraw the people from the leak contaminated areas to the windward side quickly, and a 150 metres isolation area is built immediately with limited access strictly and the fire source is cut off. The emergency dealer wear positive pressure self-contained breathing apparatus and fire protective clothing are suggested. As far as possible, the leakage source is cut off. It is important to provide reasonable ventilation to accelerate proliferation. Atomized water is sprayed to dilute and

> dissolve the chemical. A dike or trench is built to accept the mass wastewater generated. If it is possible, the leakage gas is sent to the open place by exhaust machine, or sprinklers are installed to burn it. The pipeline leading can also lead to the furnace, the valley to burn it. The containers are treated with leaks properly which can only use it after repair and testing.
>
> …

Figure 6.8 A text segment about environmental protection.

> …
>
> Formaldehyde is a strong pungent odor of gas. After inhalation of high concentrations of formaldehyde, there will be serious respiratory stimulation and edema, eye irritation, headache, gas sticks may also occur Officer asthma. Butyl acrylate is the high flash point flammable liquid, inhaled or absorbed through the skin is hazardous to health, its smell or fog on the eyes, mucous membrane and respiratory tract have a stimulating effect.
>
> …

Figure 6.9 A text segment about chemical product safety protection.

(i) Collecting information

During the construction of GZ-CEMS, a large amount of emergency information is collected from various government departments and sent to the city emergency management system. Figure 6.10 shows the folder storing the files about emergency management.

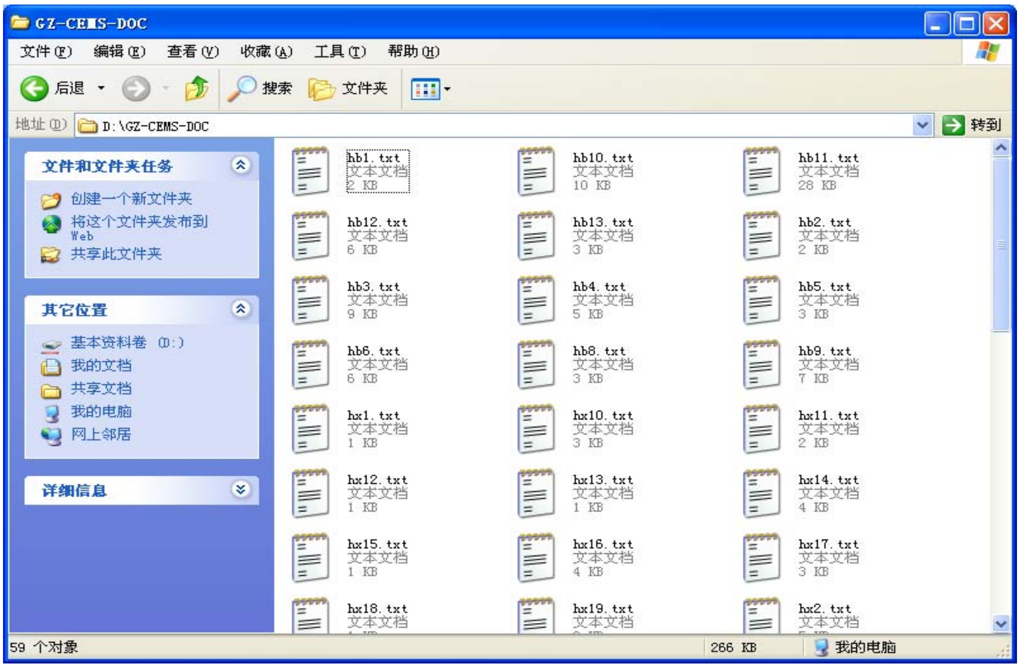


Figure 6.10 A folder storing the files about emergency management.

(ii) Predefining taxonomy

To adapt for the use of customary of users, MAKES defines a traditional taxonomy including several knowledge domains, such as fire safety, environment protection and chemical safety. Figure 6.11 displays a part of hierarchical taxonomy about emergency knowledge. In MAKES, users can browse emergency information in the taxonomy like using the Explore to browse files in Windows.

However, it is worth to point that the traditional hierarchical taxonomy is not enough for managing massive information. It costs more time for seeking information through browsing the categories in a hierarchical taxonomy. Automatic classification is becoming the approach to maintain the traditional taxonomy.

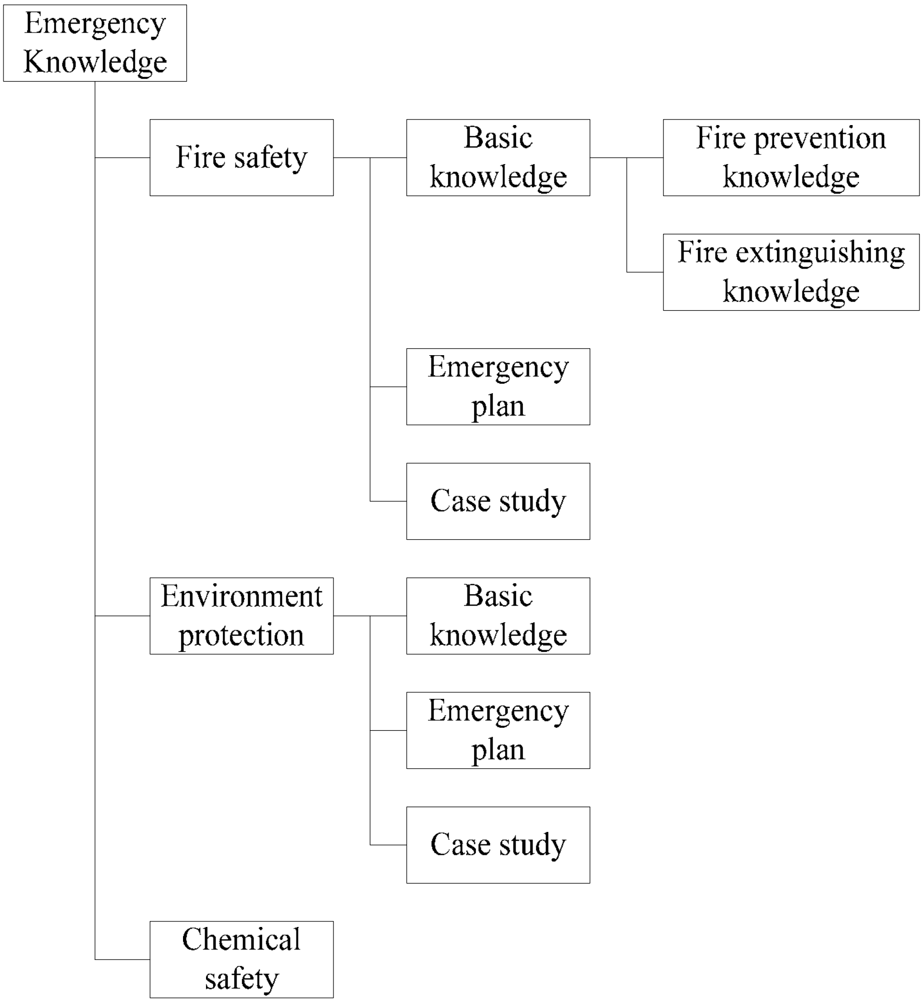


Figure 6.11 A part of taxonomy about emergency management.

(iii)Text analysis

For the traditional taxonomy, MAKES builds up an automatic classification mechanism. As soon as the emergency information is obtained, MAKES is responsible for performing the text analysis and text mining in order to build the eigenvectors of the files. Figure 6.12 shows the process of analyzing a file.

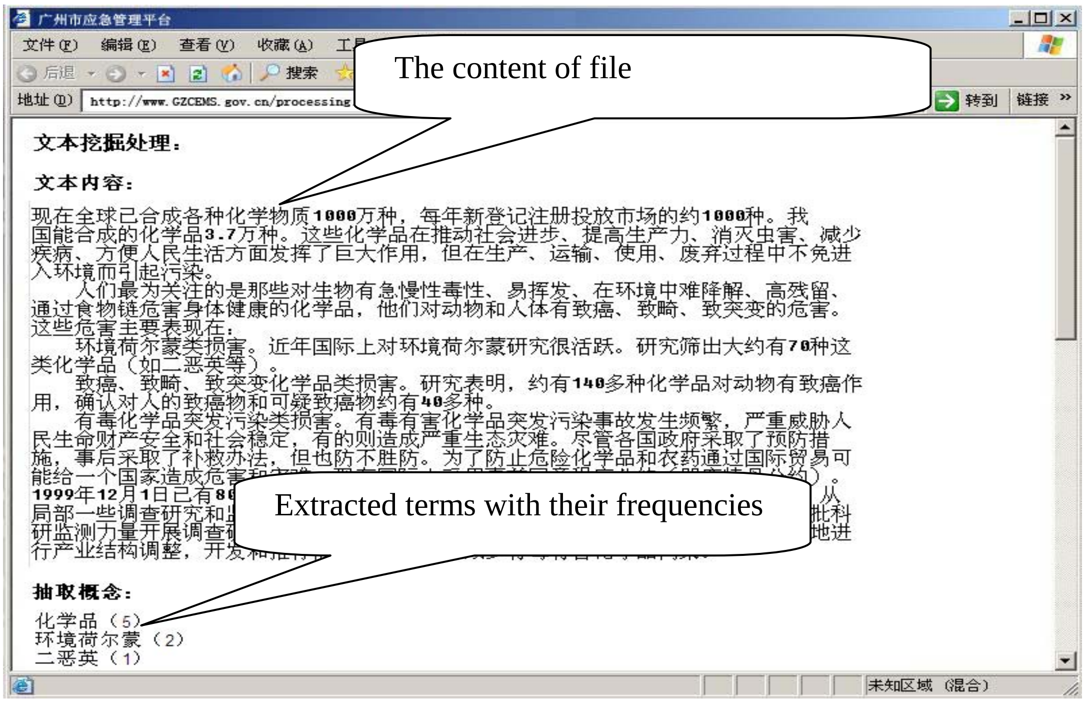
(iv) Automatic classification

Before automatic classification, some files are selected as the species which are used to train the MAKES. Every specimen is classified into one category by human. Every category is built as one eigenvector based on analyzing the specimens which have been classified into this category. Through computing the similarity between two eigenvectors of the file and every category, the file is classified automatically. Figure 6.13 shows one category and its files.

(v) Emergency information retrieval

Emergency information retrieval is an important function in emergency management. Keywords are inputted through the interface in Figure 6.14(a) and the search results are displayed in Figure 6.14(b). The difference in the search function compared to the general search engines, such as www.google.com and others, is that the search results in MAKES have been classified in the taxonomy of emergency management.

In nature, the method for listing search results is only a linear approach to organize information, and cannot be efficient for users to find out useful information from a lot of files quickly. Meanwhile, the associated relationships in massive information cannot be represented in the traditional taxonomy easily. The relationships in information are dissevered. It blocks the capability of “association thinking”.


广州市应急管理平台
文件(F) 编辑(E) 查看(V) 收藏(A) 工具
后退 搜索
地址(D) http://www.GZCEMS.gov.cn/processing
转到 链接
The content of file
Extracted terms with their frequencies
未知区域（混合）

文本挖掘处理：

文本内容：

现在全球已合成各种化学物质1000万种，每年新登记注册投放市场的约1000种。我国能合成的化学品3.7万种。这些化学品在推动社会进步、提高生产力、消灭虫害、减少疾病、方便人民生活方面发挥了巨大作用，但在生产、运输、使用、废弃过程中不免进入环境而引起污染。

人们最为关注的是那些对生物有急慢性毒性、易挥发、在环境中难降解、高残留、通过食物链危害身体健康的化学品，他们对动物和人体有致癌、致畸、致突变的危害。这些危害主要表现在：

环境荷尔蒙类损害。近年国际上对环境荷尔蒙研究很活跃。研究筛出大约有70种这类化学品（如二恶英等）。

致癌、致畸、致突变化学品类损害。研究表明，约有140多种化学品对动物有致癌作用，确认对人的致癌物和可疑致癌物约有40多种。

有毒化学品突发污染类损害。有毒有害化学品突发污染事故发生频繁，严重威胁人民生命财产安全和社会稳定，有的则造成严重生态灾难。尽管各国政府采取了预防措施，事后采取了补救办法，但也防不胜防。为了防止危险化学品和农药通过国际贸易可能给一个国家造成危害

）。

1999年12月1日已有8

从

局部一些调查研究和

批科

研监测力量开展调查

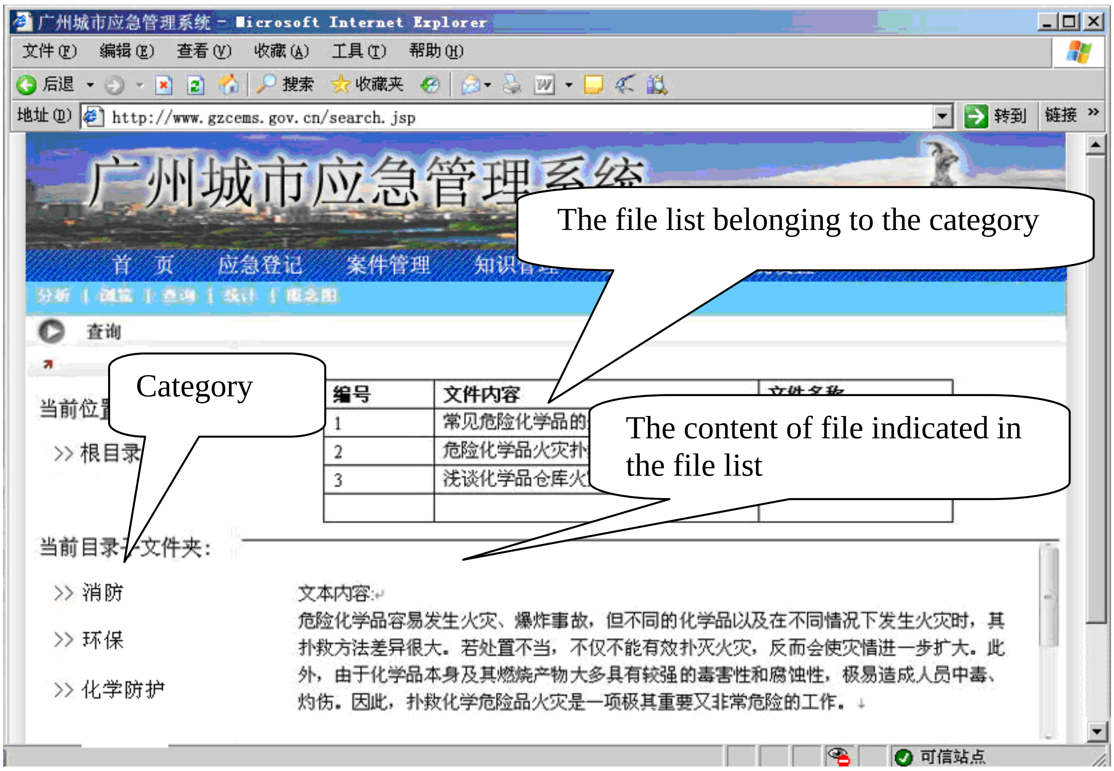
地进

行产业结构调整，开发和

抽取概念：

化学品（5）
环境荷尔蒙（2）
二恶英（1）

Figure 6.12 The process of text analysis.


广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp
转到 链接
广州城市应急管理系统
首 页
应急登记
案件管理
知识管理
分析 | 浏览 | 查询 | 统计 | 概念图
查询
The file list belonging to the category
Category
The content of file indicated in the file list
可信站点

当前位置

>> 根目录

编号	文件内容	文件名称
1	常见危险化学品的	
2	危险化学品火灾扑	
3	浅谈化学品仓库火	

当前目录子文件夹：

>> 消防

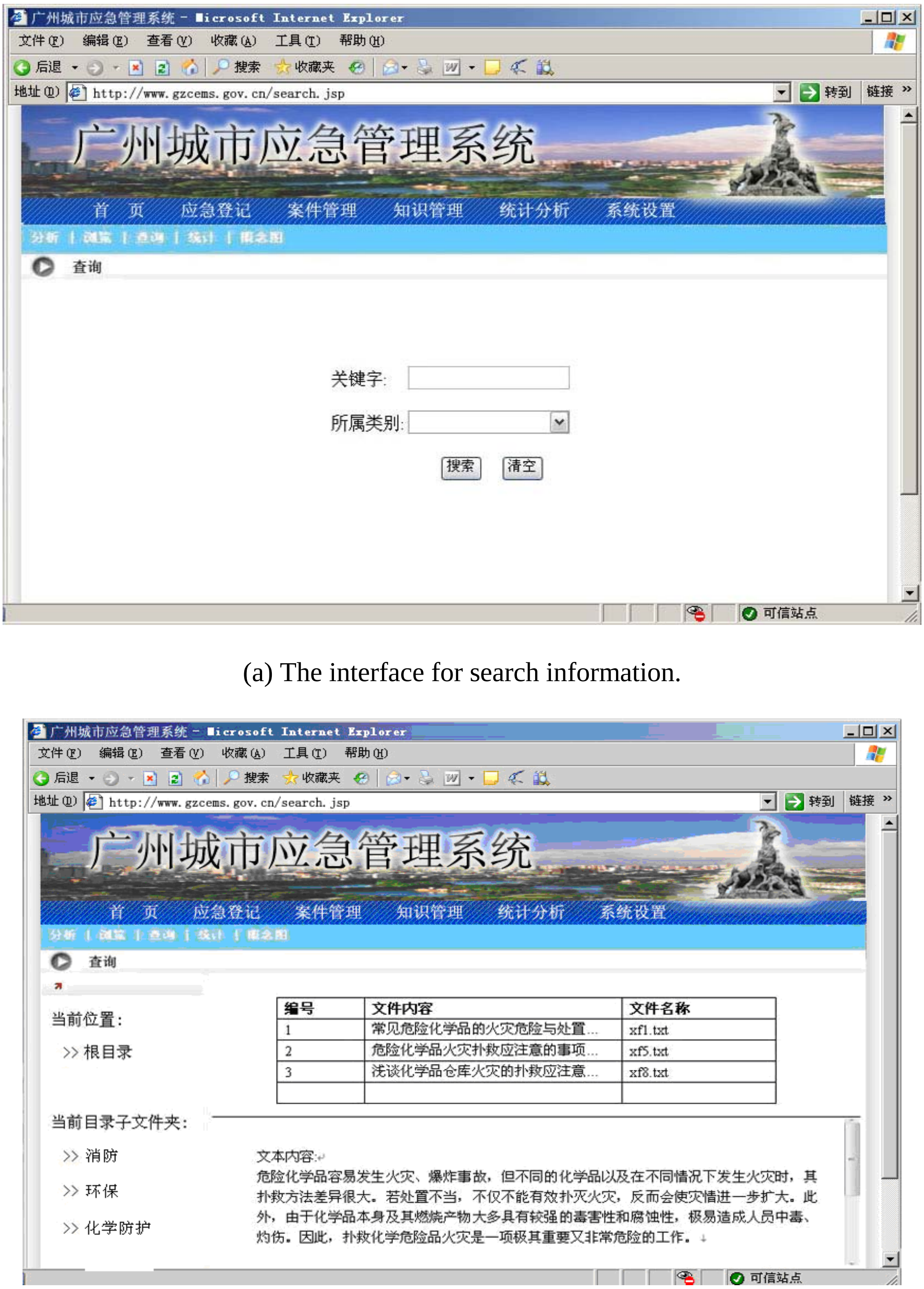
>> 环保

>> 化学防护

文本内容：

危险化学品容易发生火灾、爆炸事故，但不同的化学品以及在不同情况下发生火灾时，其扑救方法差异很大。若处置不当，不仅不能有效扑灭火灾，反而会使灾情进一步扩大。此外，由于化学品本身及其燃烧产物大多具有较强的毒害性和腐蚀性，极易造成人员中毒、灼伤。因此，扑救化学危险品火灾是一项极其重要又非常危险的工作。

Figure 6.13 A category for automatic classification.


广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp 转到 链接
广州城市应急管理系统
首 页 应急登记 案件管理 知识管理 统计分析 系统设置
分析 | 浏览 | 查询 | 统计 | 概念图
查询
关键字:
所属类别:
搜索 清空
可信站点

(a) The interface for search information.

广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp 转到 链接
广州城市应急管理系统
首 页 应急登记 案件管理 知识管理 统计分析 系统设置
分析 | 浏览 | 查询 | 统计 | 概念图
查询
当前位置:
>> 根目录
当前目录子文件夹:
>> 消防
>> 环保
>> 化学防护
可信站点

编号	文件内容	文件名称
1	常见危险化学品的火灾危险与处置…	xf1.txt
2	危险化学品火灾扑救应注意的事项…	xf5.txt
3	浅谈化学品仓库火灾的扑救应注意…	xf8.txt

文本内容:

危险化学品容易发生火灾、爆炸事故，但不同的化学品以及在不同情况下发生火灾时，其扑救方法差异很大。若处置不当，不仅不能有效扑灭火灾，反而会使灾情进一步扩大。此外，由于化学品本身及其燃烧产物大多具有较强的毒害性和腐蚀性，极易造成人员中毒、灼伤。因此，扑救化学危险品火灾是一项极其重要又非常危险的工作。

(b) The interface of search results.

Figure 6.14 The operation of emergency information retrieval.

6.4.3 Knowledge models for emergency handling process

In the traditional commanding systems and emergency management systems, the massive information and knowledge is basically stored as text files. Information and knowledge is dispersed in various documents, reports and papers. MAKES builds up several knowledge models to organize and represent emergency knowledge, Knowledge models enhance the capability of “association thinking” of human brain in emergency management.

Concept Relationship Model (CRM) of the emergency incidence is built based on extracting the concept from the unstructured information that contains emergency information and knowledge, and counts the co-occurrence frequency among the concepts. CRM can support decision-makers to understand better the dependence and degree of correlation between different concepts about the emergency in order to determine its cause and effect. The steps for emergency handling process are depicted as follows.

Step 1: Preprocessing data.

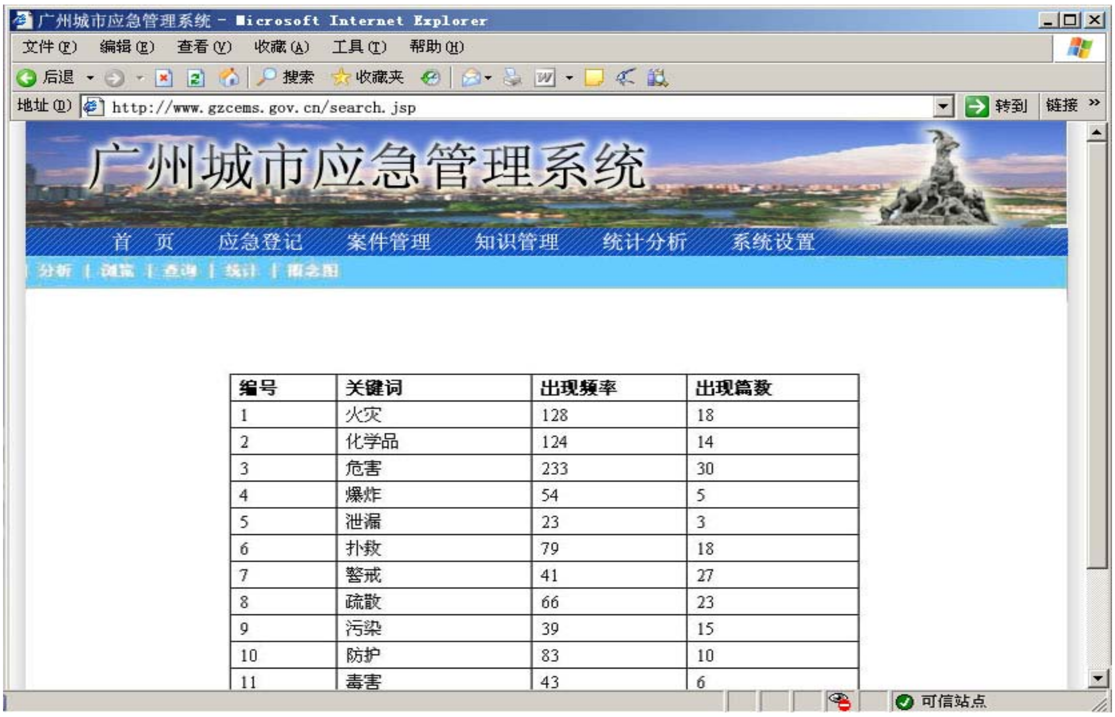
Firstly, a series of terms are extracted through mining the many files about emergency management. Table 6.2 lists the occurrence frequency of every term in the controlled vocabulary of thesaurus model.

Step 2: Concept relationship modeling.

The concept relationship modeling is done based on the statistics of the characteristics of the terms. The occurrence frequency between two terms is also calculated, and the results are shown in Table 6.2.

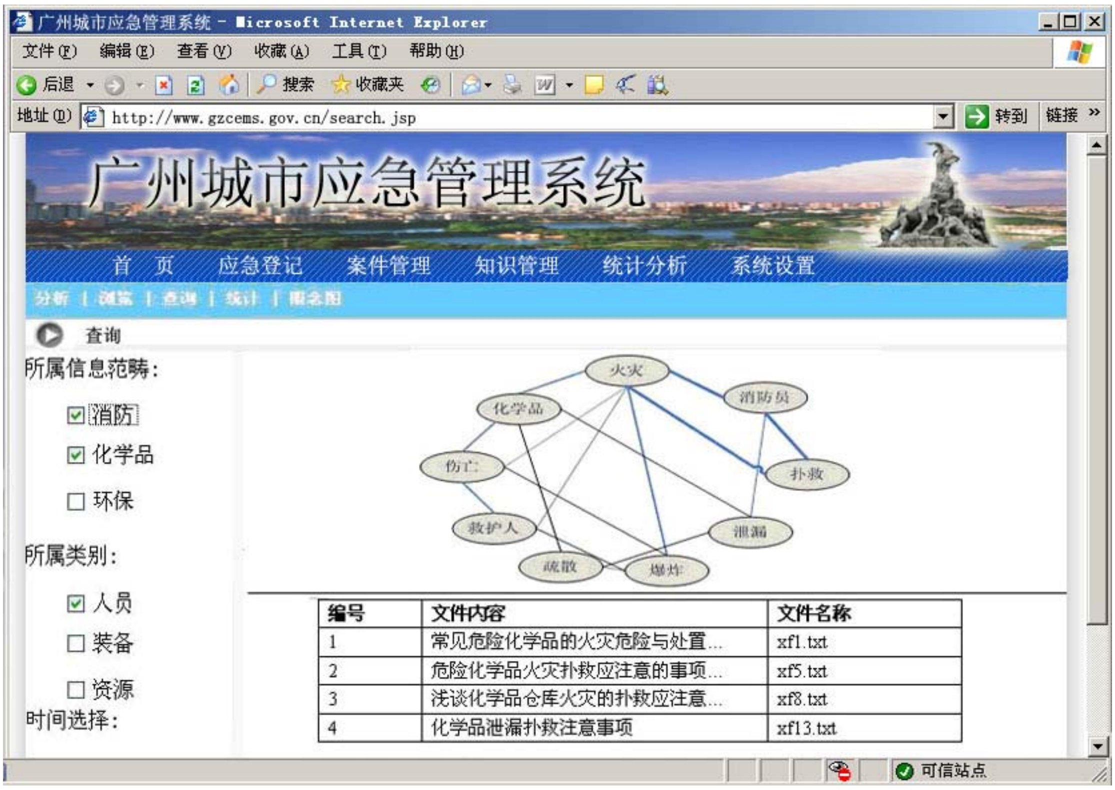
According to the frequency of the terms and the frequency of occurrence between the two terms, the association relationship between two terms can be constructed (Figure 6.15.) Through the concept relationship model of emergency process, people can see the concept association relationship in many fields of knowledge such as fire fighting, environmental protection, chemical products protection, etc. When the emergency incidents occurs, people can browse the multi-faceted view of the knowledge concept relationship model to search out a variety of interrelated knowledge of emergency in order to support the decision-making for the contingency plan.

Table 6.2 A list of occurrence frequency of terms.


广州城市应急管理系统 - ■icrosoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp
转到 链接
广州城市应急管理系统
首 页 应急登记 案件管理 知识管理 统计分析 系统设置
分析 | 浏览 | 查询 | 统计 | 概念图
可信站点

编号	关键词	出现频率	出现篇数
1	火灾	128	18
2	化学品	124	14
3	危害	233	30
4	爆炸	54	5
5	泄漏	23	3
6	扑救	79	18
7	警戒	41	27
8	疏散	66	23
9	污染	39	15
10	防护	83	10
11	毒害	43	6

A threshold is set to filter the term with the lower frequency of occurrence.


广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp
转到 链接
广州城市应急管理系统
首 页 应急登记 案件管理 知识管理 统计分析 系统设置
分析 | 浏览 | 查询 | 统计 | 概念图
查询
火灾
消防员
化学品
伤亡
扑救
救护人
泄漏
疏散
爆炸
可信站点

所属信息范畴：

☑ 消防

☑ 化学品

□ 环保

所属类别：

☑ 人员

□ 装备

□ 资源

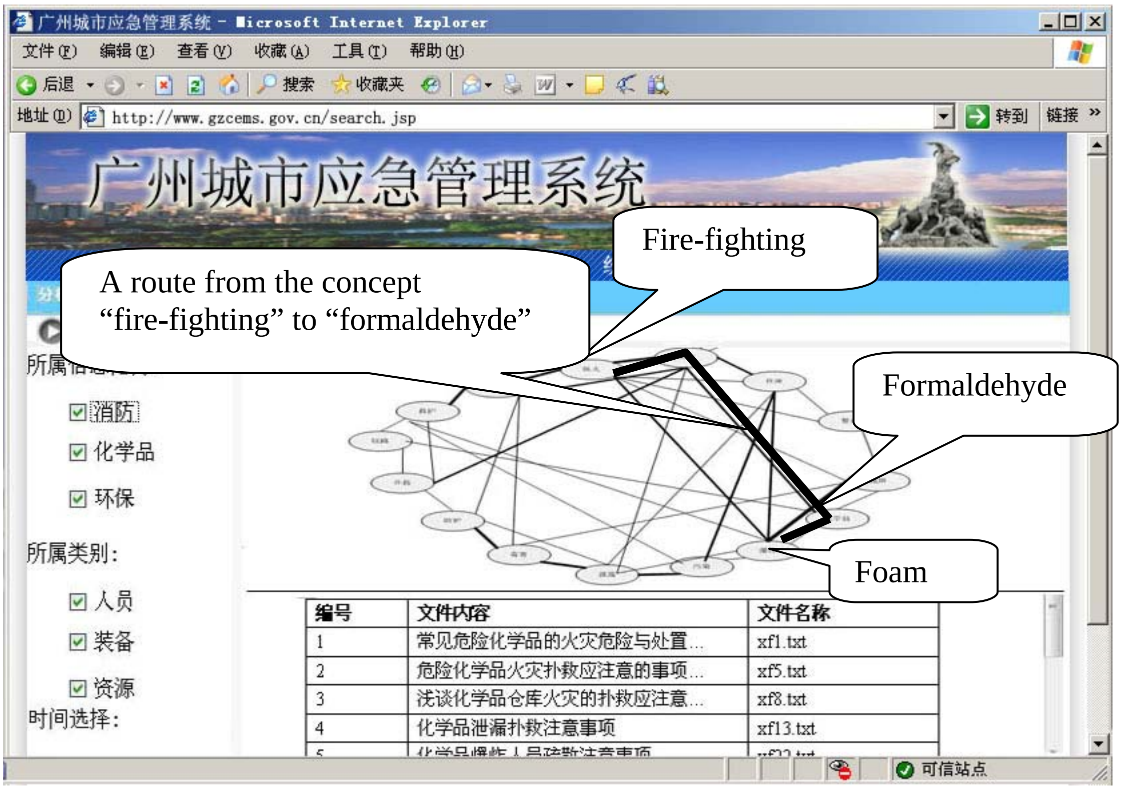
时间选择：

编号	文件内容	文件名称
1	常见危险化学品的火灾危险与处置…	xf1.txt
2	危险化学品火灾扑救应注意的事项…	xf5.txt
3	浅谈化学品仓库火灾的扑救应注意…	xf8.txt
4	化学品泄漏扑救注意事项	xf13.txt

Figure 6.15 A concept relationship model for emergency management.

As an example of verification testing the system, is to find out how to handle the explosion of chemical “formaldehyde”. In the concept relationship model in Figure 6.16, there is a route between the two concepts, chemical product “formaldehyde” and “fire-fighting”. There is also a route which links the concept “foam” with “formaldehyde” and “fire-fighting”. It prompts the possibility to fight the fire caused by the explosion of “formaldehyde” with the material “foam”.

Through clicking the line between two concepts associated with each other, the relevant articles contain the two concepts are listed. This kind of associating concepts may help us to explore various solutions.


广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp
转到 链接
广州城市应急管理系统
所属信息
☑ 消防
☑ 化学品
☑ 环保
所属类别:
☑ 人员
☑ 装备
☑ 资源
时间选择:
可信站点

Fire-fighting

A route from the concept “fire-fighting” to “formaldehyde”

Formaldehyde

Foam

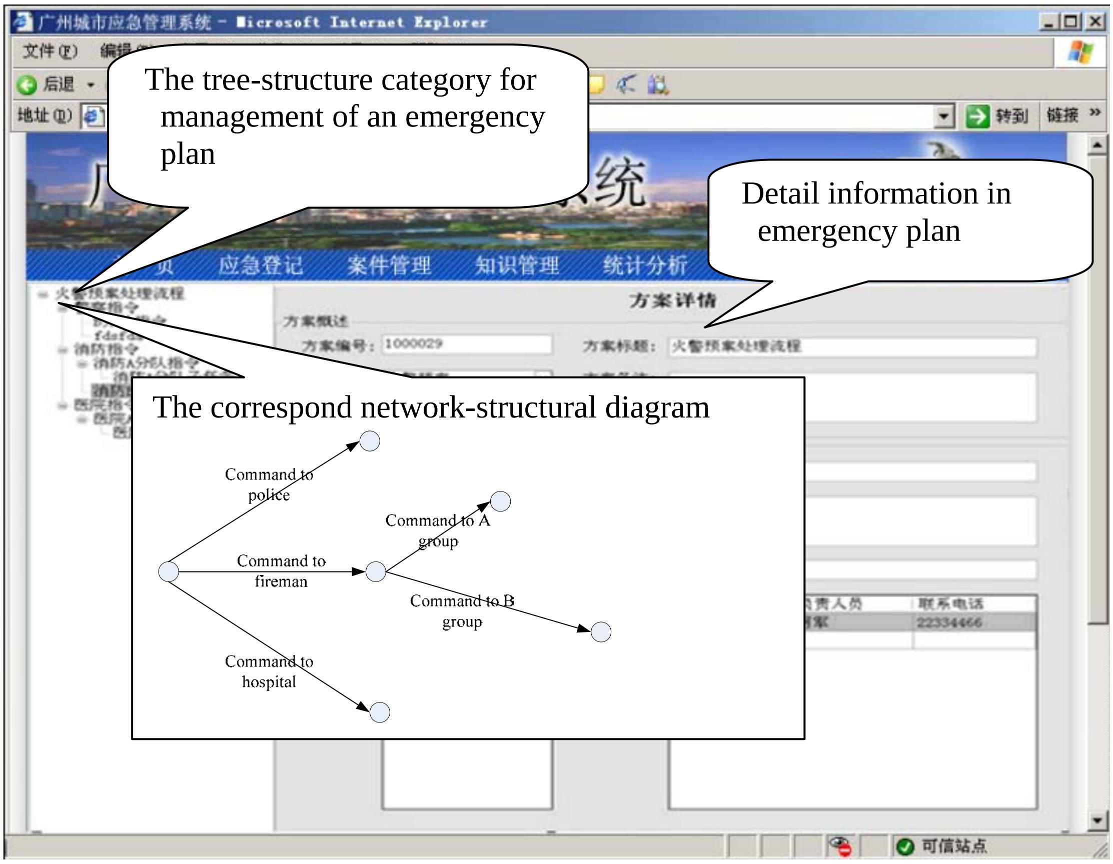
编号	文件内容	文件名称
1	常见危险化学品的火灾危险与处置…	xf1.txt
2	危险化学品火灾扑救应注意的事项…	xf5.txt
3	浅谈化学品仓库火灾的扑救应注意…	xf8.txt
4	化学品泄漏扑救注意事项	xf13.txt
5	化学品爆炸人员疏散注意事项	xf22.txt

Figure 6.16 A concreted concept relationship model for emergency management.

Step 3: Dynamic knowledge flow modeling.

Dynamic knowledge flow model is a statistics-based model through capturing the information in knowledge workflow. In this case, GZ-CEMS builds a mechanism to trace the performance of every emergency handling process. Based on the information obtained from the workflow about emergency management, MAKES in GZ-CEMS constructs the dynamic knowledge flow model. Knowledge entities and emergency information are captured and managed.

Because a tree-structure is simple and is not easy to make error in operation, GZ-CEMS adopts the tree-structure to represent emergency processing workflow and manage information of emergency incidents. Dynamic knowledge flow model can be represented by the tree-structure instead of one network-structure. Figure 6.17 describes the interface of information management in a case of fire-fighting. The correspondent network-structural diagram for the commanding information flow is mapped in Figure 6.17.



方案详情

方案概述

方案编号：1000029

方案标题：火警预案处理流程

负责人员	联系电话
[illegible]	22334466

Figure 6.17 An interface of information management for an emergency incident.

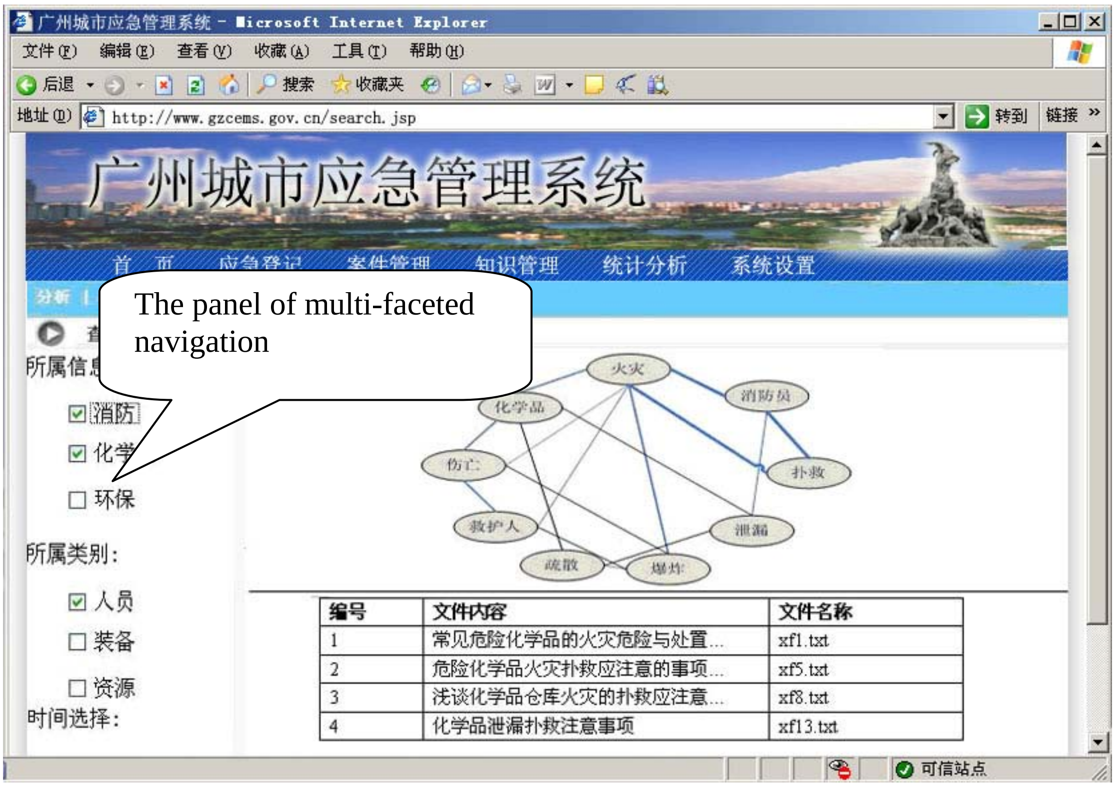
6.4.4 Multi-faceted navigation for decision making of emergency incident

In the decision-making of emergency incident, the first step is collecting and analyzing a variety of information and knowledge related to the emergency incident as soon as possible. The key is deducing the action plans and its influences to the emergency incident. Then, a plan for handling emergency incident can be made. An Input-Process-Output model of information system can be used to describe the process of emergency management. The main input information includes the emergency cases, the dynamic taxonomy, the concept relationship model and social networks, as well as the characteristics of the emergency plans and events described in the output of information for emergency work.

On this basis, the GZ-CEMS can be described based on the characteristics of the emergency incidents. It can locate the concept collection and retrieve the relevant knowledge. Through the multi-faceted knowledge navigation built into the system (Figure 6.18), a multi-dimensional view of the available knowledge assets is provided Emergency management focuses on monitoring and handling complex situations. In this case, the chemical plant explosion needs an integrated deployment of resources, fire-fighting, environmental protection, health care, law and order, and other departments to conduct co-disposal. In such a complex coordination of work to deal with emergencies, GZ-CEMS is able to quickly locate and start the relevant contingency plans and emergency response knowledge to provide effective support through multi-faceted navigation and retrieval of useful knowledge,

6.5 Evaluation architecture

In the verification test, The Committee of City Management of Guang Zhou City Municipal Government organized an expert group to evaluate the effectiveness and performance of the MAKES in Guang Zhou City Emergency Management System. Five experts from several universities and the Committee of City Management of Guang Zhou City Municipal Government were invited to be the assessors. Experts in related fields were invited to assess and verify the effectiveness of MAKES. The list of experts is referred in Appendix 1.


广州城市应急管理系统 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.gzcems.gov.cn/search.jsp
转到 链接
广州城市应急管理系统
首 页 应急登记 案件管理 知识管理 统计分析 系统设置
分析
The panel of multi-faceted navigation
火灾
消防员
化学品
伤亡
扑救
救护人
泄漏
疏散
爆炸
可信站点

所属信息

☑ 消防

☑ 化学

□ 环保

所属类别:

☑ 人员

□ 装备

□ 资源

时间选择:

编号	文件内容	文件名称
1	常见危险化学品的火灾危险与处置...	xf1.txt
2	危险化学品火灾扑救应注意的事项...	xf5.txt
3	浅谈化学品仓库火灾的扑救应注意...	xf8.txt
4	化学品泄漏扑救注意事项	xf13.txt

Figure 6.18 A snapshot for multi-faceted knowledge navigation for emergency incident.

Since decision-making under emergency conditions is a very dynamic and complex process, the availability of related information and knowledge is important for good decision-making. There are many socio-economic factors which are necessary to be considered. To evaluate the effectiveness and efficiency of MAKES in emergency management comprehensively and scientifically, an evaluation architecture of MAKES in emergency management (Figure 6.19) is proposed and implemented. There are five areas to be evaluated. These areas are automatic classification and information retrieval, concept relationship model, dynamic knowledge flow model, social network and multi-faceted navigation. Each area is measured by a quantitative and qualitative measurement indicator according to the characteristics of the area. All of the results of evaluation are synthesized to come up with an overall score.

Two types of measurement indictors are used: namely the quantitative and qualitative indictor. Quantitative measurement is based on some quantifiable aspects of the performance of the area to be evaluated, and qualitative measurement is often a rating based on the expert opinion of the assessor. The effectiveness of information systems cannot be measured by quantitative means alone, and needs to be supplemented by opinions of the experts and the users. The details of the measurement indicators in each of the five evaluation area are given below.

(i) Automatic classification

The objective of automatic classification is to collect unstructured information and classify it automatically. The objective of evaluation of automatic classification is based on the evaluation of the precision of automatic classification and the usability of information retrieval based on dynamic taxonomy.

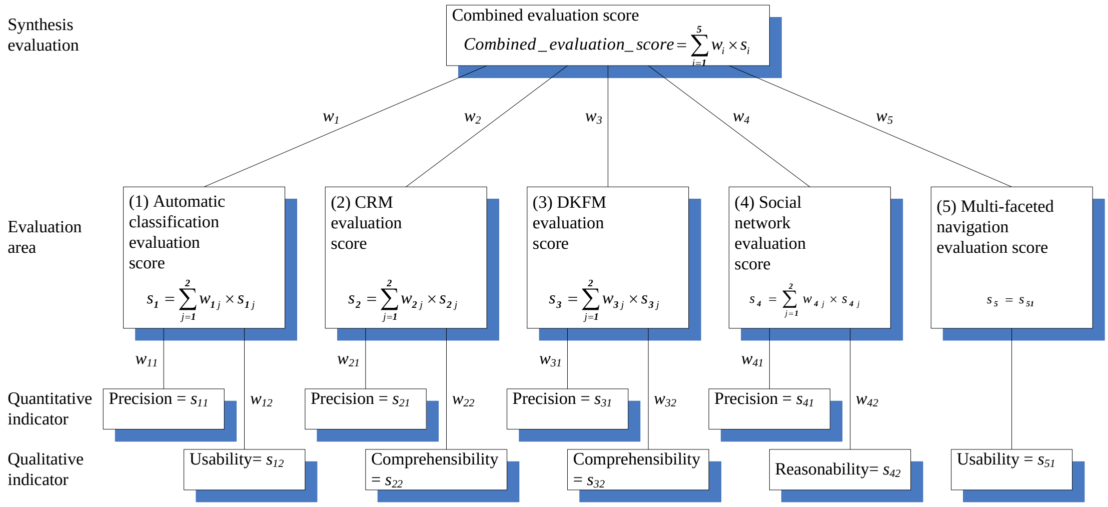

Synthesis evaluation
Combined evaluation score
$Combined\_evaluation\_score = \sum_{i=1}^{5} w_i \times s_i$
w_1
w_2
w_3
w_4
w_5
Evaluation area
(1) Automatic classification evaluation score
$s_1 = \sum_{j=1}^{2} w_{1j} \times s_{1j}$
(2) CRM evaluation score
$s_2 = \sum_{j=1}^{2} w_{2j} \times s_{2j}$
(3) DKFM evaluation score
$s_3 = \sum_{j=1}^{2} w_{3j} \times s_{3j}$
(4) Social network evaluation score
$s_4 = \sum_{j=1}^{2} w_{4j} \times s_{4j}$
(5) Multi-faceted navigation evaluation score
$s_5 = s_{51}$
w_{11}
w_{12}
w_{21}
w_{22}
w_{31}
w_{32}
w_{41}
w_{42}
Quantitative indicator
Precision = s_{11}
Precision = s_{21}
Precision = s_{31}
Precision = s_{41}
Qualitative indicator
Usability= s_{12}
Comprehensibility = s_{22}
Comprehensibility = s_{32}
Reasonability= s_{42}
Usability = s_{51}

Figure 6.19 The evaluation architecture of MAKES in emergency management.

Indices of evaluation criterion:

- Precision of automatic classification is defined as the number of documents which are automatically and correctly classified divided by the total number of documents in a set of testing documents. In the current study, 200 documents are selected by The Committee of City Management of Guang Zhou City Municipal Government.
- Usability is defined as the degree with which the function of automatic classification and information retrieval of MAKES is operated smoothly and with ease according to the experience of the experts.

(ii) Concept relationship model

Concept relationship model is to build the concept relationships about emergency management. The evaluation of the CRM is based on the precision of extracting concept and the comprehensibility of CRM.

Two indices of criterion are defined:

- Precision of extracting concept is defined as the number of concepts extracted by the algorithm of CRET divided by the total number of concepts which are in the testing set of documents.
- Comprehensibility of CRM is defined as the degree about CRM can provide some reasonable clues from the relationships of concepts to handle the emergency incident in this verification testing case according to the experience of the experts.

(iii)Dynamic knowledge flow mode

The Dynamic knowledge flow model (DKFM) is used to construct the pattern of knowledge flow in emergency management. The objective of the evaluation of the dynamic knowledge flow model is based on evaluating the precision of identifying knowledge entities and the comprehensibility of DKFM.

Two indices of evaluation criterion are defined as follows:

- Precision of identifying knowledge entity is defined as the number of automatic identified knowledge entities divided by the total number of knowledge entities which are in the testing set of documents.
- Comprehensibility of DKFM is defined as the degree to which the DKFM can provide some reasonable clues to trace and acquire emergency knowledge according to the characteristics of emergency incident in this

verification case according to the experience of experts.

(iv) Social network.

The social network shows the communication channels of people and their connectivity. The objective of evaluating the social network is to evaluate the precision of the identifying appropriate persons in the social network and the reasonability of the social network.

Two indices of evaluation criterion are given as follows:

- Precision of identifying person, is defined as the number appropriate persons automatically identified divided by the total number of persons which are in the testing set of documents.
- Reasonability of social network, is defined as the degree to which the social network can provide some reasonable clues appropriate persons to handle the emergency incident in this verification case according to the experience of experts.

(v) Multi-faceted navigation

Multi-faceted navigation provides a platform to navigate the knowledge models about emergency management, and the objective of the evaluation of multi-faceted navigation is to evaluate its usability.

Index of evaluation criterion:

- Usability is defined as the degree in which the function of multi-faceted navigation in knowledge models is operated with ease or not according to the judgment of the experts using the platform of multi-faceted navigation.

In this evaluation, a 5-score grading is adopted in which, 0 means the performance is worst and 5 means the performance is excellent. The criteria of 5-score grading are described in Table 6.3.

6.6 Measurement and discussion

It is worthy to note that the Guang Zhou City Emergency Management System is a government project, which has won the Science and Technology Progress Award of Guang Zhou Municipal Government in 2008. For confidentiality reasons, many cases could not be openly discussed and analyzed. As a result, only a small number of non-sensitive cases are selected for analysis analyzed in this study.

Table 6.3 The criteria of 5-score grading for evaluation of MAKES in emergency management.

Score	**Criterion**	**Comment**
1	Knowledge and information retrieved from MAKES is trivial and the function of MAKES does not serve the purpose it is intended	Bad
2	Knowledge and information retrieved from MAKES is of marginal use and the function of MAKES does not serve well its original purpose	Poor
3	Knowledge and information retrieved from MAKES is helpful and the function of MAKES is realized	Satisfactory
4	Knowledge and information retrieved from MAKES is found to be useful and the function of MAKES serves the purpose well	Good

5	Knowledge and information retrieved from MAKES is very useful and the function of MAKES achieves all its designed purposes .	Excellent

The measurement process is executed by the evaluation expert group and the staff from The Committee of City Management of Guang Zhou City Municipal Government. The results of this evaluation of MAKES in emergency management are described as below.

(i) Automatic classification

In the evaluation of MAKES, 200 documents were selected as a testing set. After classifying these documents from the MAKES, 162 documents were identified manually to be classified correctly by the staff from The Committee of City Management of Guang Zhou City Municipal Government. The experts then verified the results. As a result, the precision of automatic classification = (162 x 100)/200 = 81%.

According to the rule of 5 score grading, the score for precision of automatic classification is 4.05. So, the rating is considered as “good”.

Moreover, the experts give their own scores on the usability of automatic classification and information retrieval according to their experience in using MAKES as shown in Table 6.4.

In this evaluation, the names of experts are anonymous. Combining both the precision of automatic classification and the usability, the weights of two indices in the evaluation of automatic classification are equal, the combined score = (4.05+4)/2 = 4.025. The combined score on the evaluation of automatic classification is 4.025.

Table 6.4 Results of evaluation of automatic classification and information retrieval.

No	**Criterion**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**Average score**
1	Usability	4.0	3.5	4.2	4.4	3.9	4

(ii) Concept relationship model

In the testing set of 200 documents, 7536 concepts were extracted automatically according to the thesaurus model, while 8438 concepts were identified manually by the staff in The Committee of City Management of Guang Zhou City Municipal Government. The results were verified by the experts. As a result, the precision of extracting concept = 7536 /8438 = 89.3%, and the score for precision of extracting concept is 4.47.

In addition, the experts gave their scores for comprehensibility according to their experience from the CRM in MAKES as shown in Table 6.5.

Table 6.5 Results of evaluation of concept relationship model.

No	**Criterion**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**Average score**
1	Comprehensibility	3.8	3.9	4.0	3.9	4.0	3.92

Combing both the precision of extracting concept and the evaluation of comprehensibility, the weights of two indices are equal, therefore

the combined score = (4.47+3.92)/2 = 4.195.

The combined score of evaluation of concept relationship model is 4.195.

(iii)Dynamic knowledge flow mode

In the testing set of 200 documents, 286 knowledge entities were identified automatically, including various knowledge sources, while 345 knowledge entities were identified by the staff in The Committee of City Management of Guang Zhou City Municipal Government. The results were verified by the experts. Thus, the precision of identifying knowledge entity = 286/345 = 82.9%.The score of precision of identifying knowledge entity is 4.14.

Furthermore, the experts gave their scores for comprehensibility according to their experience from DKFM in MAKES as shown in Table 6.6.

Table 6.6 Results of evaluation of dynamic knowledge flow model.

No	**Criterion**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**Average score**
1	Comprehensibility	3.4	3.8	4.2	3.7	3.9	3.8

Combining both the precision of identifying knowledge entity and comprehensibility, when the weights of two indices are equal, the combined evaluation score = (4.14+3.8) / 2 = 3.97.The combined score of evaluation of dynamic knowledge flow model is 3.97.

(iv) Social network.

In the testing set of 200 documents, 132 persons were identified automatically, while 148 persons were identified manually by the staff in The Committee of City Management of Guang Zhou City Municipal Government. The results were verified. Thus,

the precision of identifying person = 132/148 = 89.2%.

The score for precision of identifying person is 4.46.

In addition, the experts gave their scores for reasonability according to their experience from the social network in MAKES as shown in Table 6.7.

Table 6.7 Results of evaluation of social network.

No	**Criterion**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**Average score**
1	Reasonability	4.5	4.6	4.3	4.3	4.4	4.42

Combining both the precision of identifying person and the reasonability, with the weights of two indices being equal, and the combined evaluation score = (4.46+4.42)/2 = 4.44.The combined score of evaluation of social network is 4.44.

(v) Multi-faceted navigation

The experts rate the usability according to their experience from multi-faceted navigation in MAKES as shown in Table 6.8.

As a result, the score of evaluation of multi-faceted navigation is 4.0.

Table 6.8 Results of evaluation of multi-faceted navigation.

No	Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Average score
1	Usability	3.8	4.1	4.1	3.9	4.1	4.0

(vi) Synthesis evaluation

Based on the above, a series of evaluations for the key functions of MAKES, the result of the overall evaluation of MAKES in the Guang Zhou City Emergency Management System is shown in Table 6.9.

Table 6.9 Results of overall evaluation of MAKES in emergency management.

No.	Criteria	Weight	Score
1	Automatic classification and information retrieval	0.15	4.025
2	Concept relationship model (CRM)	0.25	4.195
3	Dynamic knowledge flow model (DKFM)	0.2	3.97
4	Social network	0.2	4.44
5	Multi-faceted navigation	0.2	4.0
Total score		**4.1345**	

All measurement scores are organized as a score tree which is shown in Figure

6.20. The overall combined evaluation score is 4.1345.

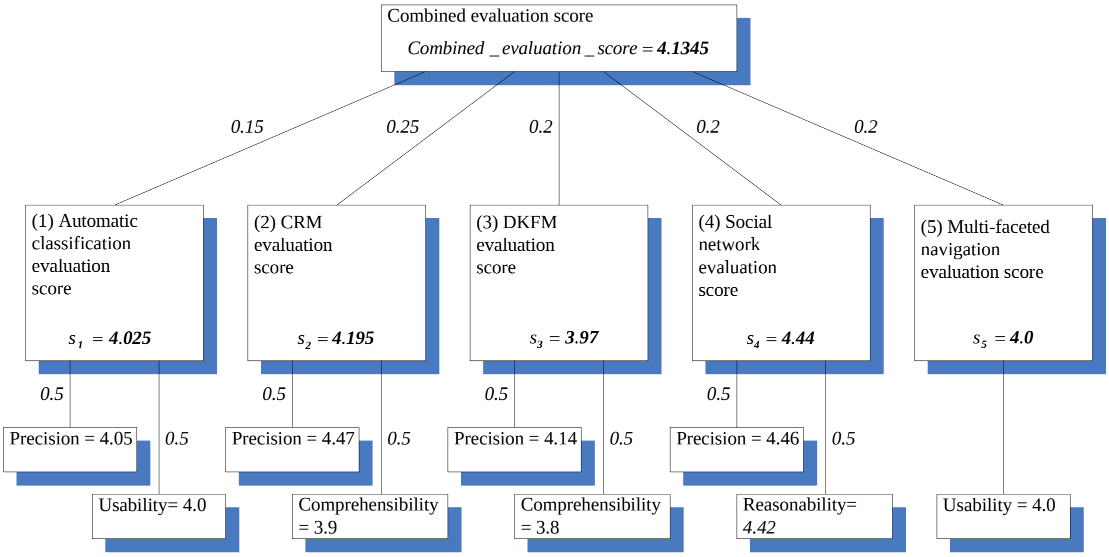


Figure 6.20 Score tree of synthesis evaluation.

In the current concept relationship model of emergency knowledge, technologies for semantic analysis are not used. As a result, the ability to describe concepts and their relationships is relatively simple which limits the intelligent reasoning functions. In the next stage of development, semantic analysis will be deployed to come up with a semantic knowledge network model for intelligent reasoning to enable emergency decision-making to be more effective and more accurate.

The establishment of emergency knowledge based on semantic analysis is feasible. However, in practice, semantic analysis is very complex because this involves the morphology of language, syntax, and grammar analysis. This requires a large team, including the participation of linguists. There is a future need to enhance the system’s ability of using semantic analysis in order to improve the intelligence and effectiveness of emergency knowledge management.

Chapter 7. Verification Testing in Knowledge Auditing of Masses of Unstructured Information (Case Two)

Knowledge auditing is an important aspect of knowledge management. The results of a knowledge audit are derived from the analysis of a huge amount of unstructured information, e.g. email. Effective management of the unstructured information is essential to enable an organization to uncover valuable knowledge hidden underneath the sea of information and knowledge. In this chapter, the results of a trial implementation of MAKES in supporting automatic auditing of valuable knowledge from a mass of unstructured information in a trading company are presented. This provides an important means to verify the capability of MAKES when applied to the knowledge auditing of unstructured information.

7.1 Background of the case

A trading company named Angus Electronics Co. Ltd., was planning to implement knowledge management. The company was selected as a reference site for the verification testing of MAKES. Angus started business in 1990 and has more than 60 staff distributed between the Hong Kong headquarters and the Shenzhen branch office. The main products are electronic components for audio, video and home appliances. There are over 3,500 customers and nearly 100 vendors. Angus aims to provide value-added services, new product ideas and total solutions for its customers to meet the changing market environment.

Trading industries are special industries because most of them do not have their own core technology. They depend on their knowledge of purchasing the right products from suppliers and selling them to the right customers. As a result, updated market information, customer analysis and knowledge of new technologies are critical to the survival of trading companies. As a trading company, an excellent customer and supplier network as well as the ability to form good relationships with the stakeholders are the core competences of Angus. It is also vital to manage rapid changes in product concepts and knowledge so as to develop effective knowledge sharing and acquisition strategies.

Angus started its Knowledge Management (KM) initiatives in 2003. MAKES is one of the KM strategies for supporting the knowledge work in the company. The aim of MAKES is to perform a knowledge audit initially and then manage the valuable knowledge embedded in large numbers of emails among different parties. On average, over 500 emails about quotations, sales orders, purchase orders and technical problems are received every day. As shown in Figure 7.1, all staff members are habitually using Outlook Express, which is a client application, for their email management. There is no centralized email management system in the company.

Nowadays, Angus has shifted from being a traditional trading business to providing integrated solutions and services to their customers. In Angus, a large amount of information and knowledge comes from the email among staff, vendors, and customers. As can be seen from Figure 7.1, the staffs use Outlook application to send and receive emails. With this working pattern, it is very difficult for staff to share information among themselves because email messages are scattered over different personal computers. To solve the problems of information sharing, staff of Angus usually forward email to relevant locations. However, a large amount of forwarded emails may lead to confusion and inconsistency of information. In this situation, the company staff has to spend more time to search and locate the email knowledge as there is no effective method to manage the ever increasing amount of unstructured information that is flowing in the company.

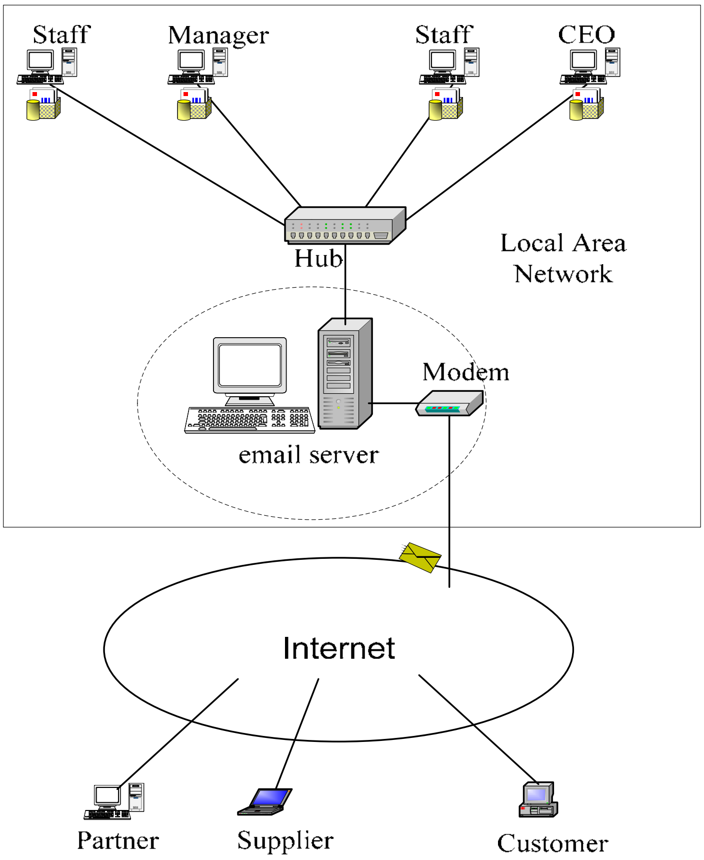


Figure 7.1 The topology of traditional email work system.

The traditional email system has several limitations. Firstly, the emails are distributed in the desktop computers of different staff members and it is very difficult to share them among different users' desktops. Intranet and email forwarding are used for email sharing in Angus. Since different users have different criteria and behavior for classifying emails, it is hard to search for the right emails without having a good taxonomy system. It is not only time consuming but also ineffective when the knowledge workers want to find the right knowledge from a sea of emails. Table 7.1 summarizes the challenges encountered by Angus in managing its unstructured information, i.e. emails. To manage a massive number of emails every day and support knowledge workers in searching and browsing emails effectively, a multi-facet taxonomy system called MAKES was developed and trial implemented in Angus.

Table 7.1 Challenges in current email management system.

Detailed problem descriptions	Consequence of problems
Outlook Express is a client application. The emails are generally deleted from the email server and stored in personal computers of staff. There is no central database for the storage of the emails	i) Difficult to share the emails among different staff ii) Loss of emails after staff resign iii) Duplication of emails in staff's computers

Intranet is the platform for email sharing. Since there are too many categories in the hierarchical structure in the company intranet, it is ineffective for sharing emails.	iv) Hard to identify the right email in a huge and complex taxonomy in the intranet during email searching. v) Difficult to classify the emails in the intranet. vi) The efficiency of email sharing in the intranet is low.
Email forwarding is an alternative method for sharing the knowledge in the email. Usually, the staff needs to forward all the emails about business issues to their supervisors or managers. If the manager supervises 3 staff, then he/she may receive 3 copies of the same email which are forwarded by his/her subordinates.	vii) Increases non-value adding workload on the staff (time consuming) viii) Increases duplication of emails. ix) It is difficult for Managers to find the right emails
Only a basic searching function is provided by Windows is available on the intranet (This can only provide searching of file names but cannot search the emails according to contents). However, it does not	x) The accuracy and relevance of the searching is low xi) A large number of irrelevant files are searched.

focus on email searching. The performance of this searching function is low for such a huge number of files.	
Different staffs have different email management behaviors. Emails are classified in different times and using different criteria.	xii) The emails in the intranet are not up to date. xiii) Staff may not know where the emails are when they are uploaded by other staff.
Emails are classified and filed manually.	xiv) Increase in workload xv) Increase in human errors of email classification
Emails can only be checked in offices.	xvi) It is inaccessible for the users outside office hour.

7.2 Design of verification testing

MAKES was built in Angus for the pilot verification testing. In this case, MAKES was developed as a web-based system to be plugged into the original email system as shown in Figure 7.2. It receives emails from the email server and stores them in the centralized dynamic taxonomy. The staffs browse the entire dynamic taxonomy by using the browsers from their personal computers. It is easy to share information and quickly search and acquire useful information. Meanwhile, the email handling function enables staff to use the MAKES to process their email-based knowledge work, instead of Outlook.

The first step of the implementation of MAKES is the development of a thesaurus model for the company. A thesaurus study is carried out via knowledge worker interviews and survey questionnaires so as to collect the initial input for the thesaurus model. The thesaurus model includes a controlled vocabulary for the company containing such items as organization names, the names of the contact persons for the customers, vendors, partners and internal users, product types, product series and product items, and the terminology of the operation; the synonyms of all those names and terminology; and the stop words or words that are not of any interest. New concepts are discovered automatically when words are found that do not exist in the thesaurus model. How significant the new concepts are depends on how they are rated by the rule-based popularity and intensity analyses. As a result, the thesaurus model is continuously updated. This operation dramatically reduces the effort involved in maintaining the thesaurus model.

As shown in Figure 7.3, an automatic engine is built to extract the Really Simple Syndication (RSS) of thesauruses from the websites of the public thesauruses (e.g. http://www.electronicproducts.com/rss/). The updated data is continuously checked and compared with the company’s thesaurus. A “To Be Confirmed Concept List” is generated and stored in the company’s knowledge repository for approval. At the same time, new emails of the company and WebPages of the competitors are continuously monitored according to the methodology described in the previous section. When the intensity and popularity of a new concept are higher than their pre-defined thresholds, the concept is added to the “To Be Confirmed Concept List”. The authorized personnel needs to revise the concepts in the list and retains them in the thesaurus for future reuse. Figure 7.4 shows a list of elicited concepts from the Angus’s MAKES.

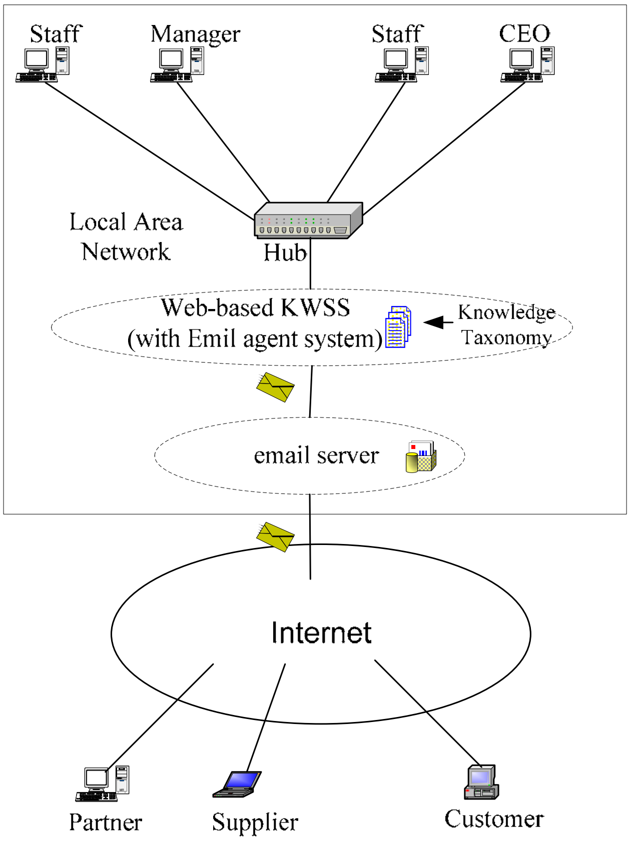
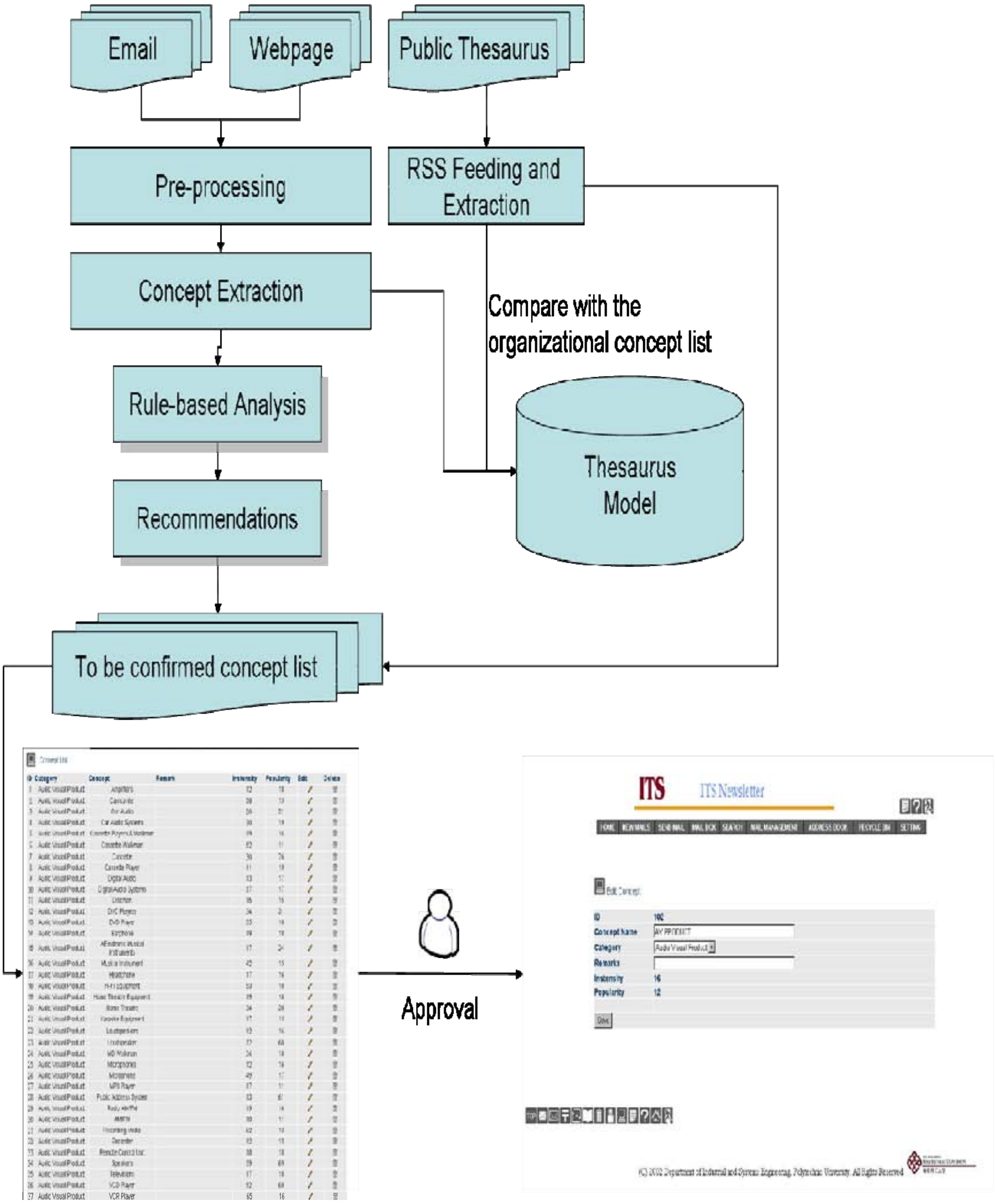
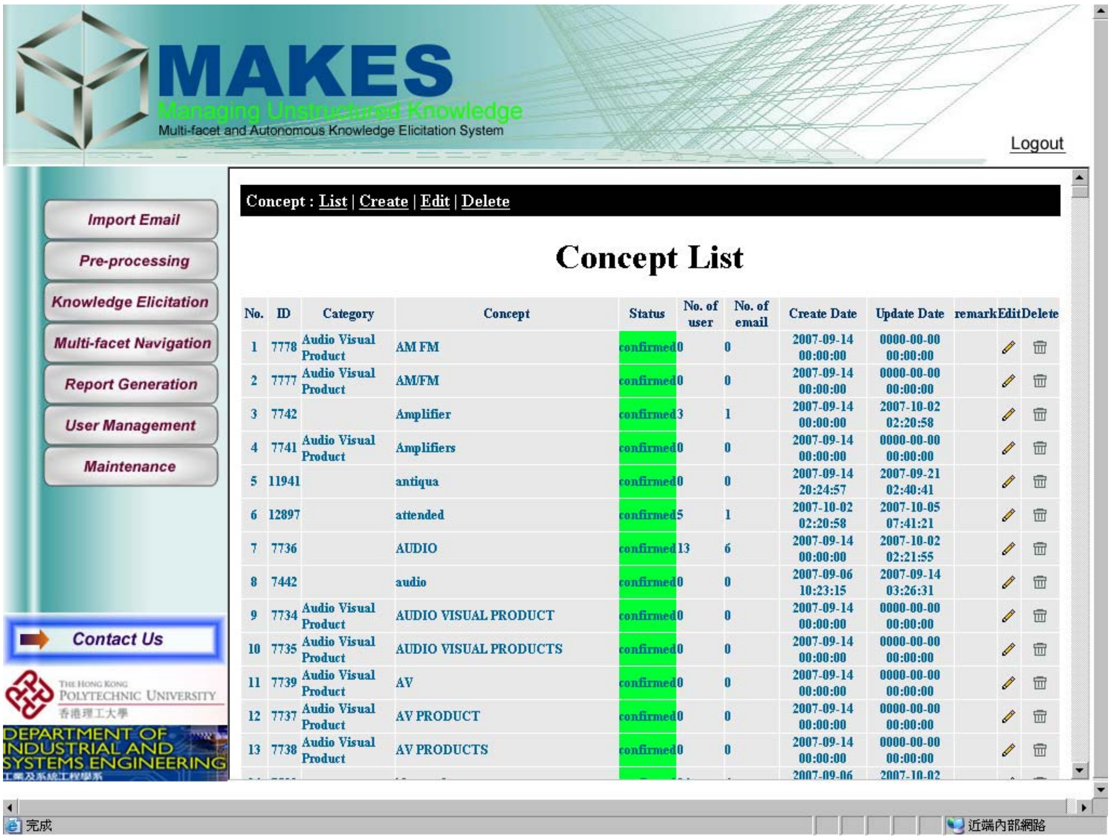


Figure 7.2 The topology of Angus’s MAKES.



ID 102
Concept Name AV PRODUCT
Category Audio Visual Product
Remarks
Intensity 16
Popularity 12

Figure 7.3 The schema of concept elicitation and maintenance process.


MAKES
Managing Unstructured Knowledge
Multi-facet and Autonomous Knowledge Elicitation System
Logout
Import Email
Pre-processing
Knowledge Elicitation
Multi-facet Navigation
Report Generation
User Management
Maintenance
Contact Us
The Hong Kong Polytechnic University
香港理工大學
DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING
完成
近端內部網路

Concept : List | Create | Edit | Delete

Concept List

No.	ID	Category	Concept	Status	No. of user	No. of email	Create Date	Update Date	remark	Edit	Delete
1	7778	Audio Visual Product	AM FM	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
2	7777	Audio Visual Product	AM/FM	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
3	7742		Amplifier	confirmed	3	1	2007-09-14 00:00:00	2007-10-02 02:20:58		✎	🗑
4	7741	Audio Visual Product	Amplifiers	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
5	11941		antiqua	confirmed	0	0	2007-09-14 20:24:57	2007-09-21 02:40:41		✎	🗑
6	12897		attended	confirmed	5	1	2007-10-02 02:20:58	2007-10-05 07:41:21		✎	🗑
7	7736		AUDIO	confirmed	13	6	2007-09-14 00:00:00	2007-10-02 02:21:55		✎	🗑
8	7442		audio	confirmed	0	0	2007-09-06 10:23:15	2007-09-14 03:26:31		✎	🗑
9	7734	Audio Visual Product	AUDIO VISUAL PRODUCT	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
10	7735	Audio Visual Product	AUDIO VISUAL PRODUCTS	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
11	7739	Audio Visual Product	AV	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
12	7737	Audio Visual Product	AV PRODUCT	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑
13	7738	Audio Visual Product	AV PRODUCTS	confirmed	0	0	2007-09-14 00:00:00	0000-00-00 00:00:00		✎	🗑

Figure 7.4 A list of elicited concepts from Angus's MAKES.

Figure 7.5 shows the process of Multi-faceted Navigation based on the searching input. Users can input the search criteria, which include department, personnel, concepts, date, and they can select the analysis type (i.e. concept relationship or social network). Then the knowledge assets are filtered based on the selected criteria, and the corresponding network is generated. By clicking on the nodes or edges of the network, corresponding knowledge assets are displayed.

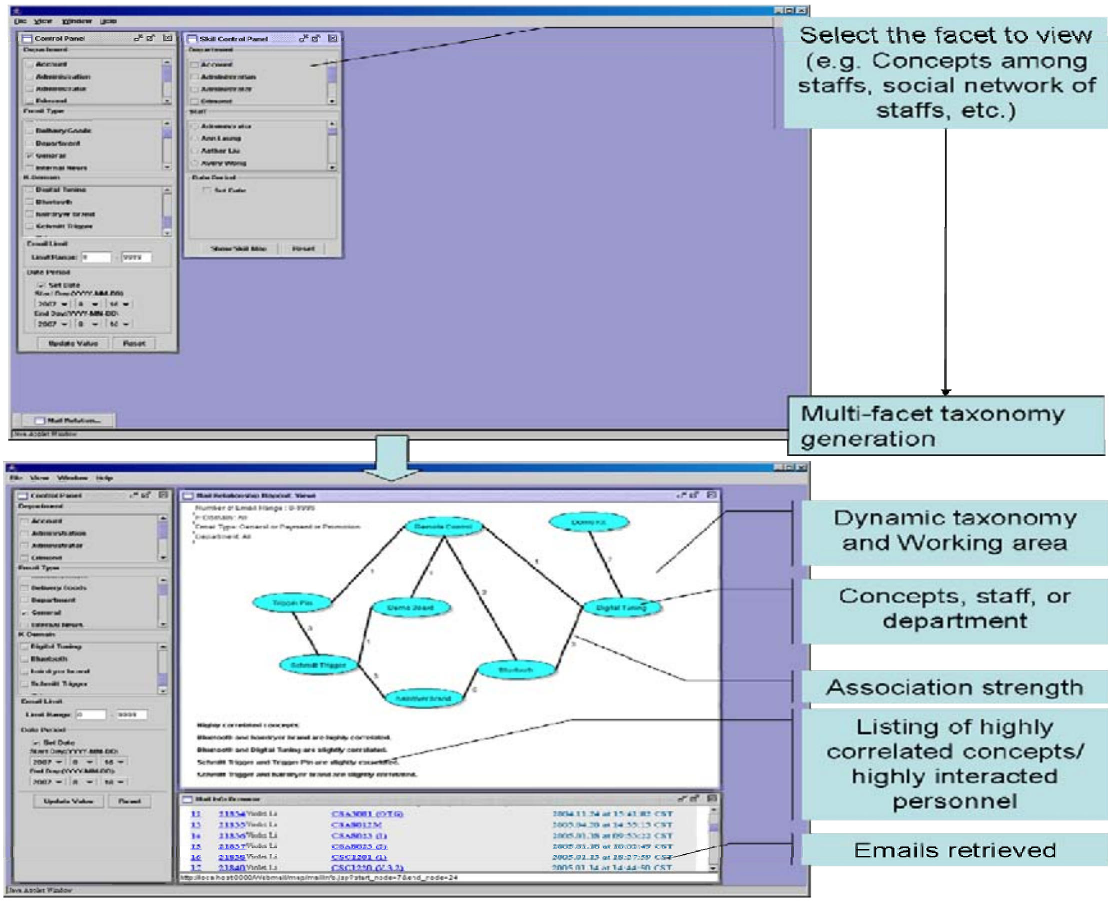

Select the facet to view (e.g. Concepts among staffs, social network of staffs, etc.)
Multi-facet taxonomy generation
Dynamic taxonomy and Working area
Concepts, staff, or department
Association strength
Listing of highly correlated concepts/ highly interacted personnel
Emails retrieved

Figure 7.5 The schema of the multi-faceted navigation process of unstructured information.

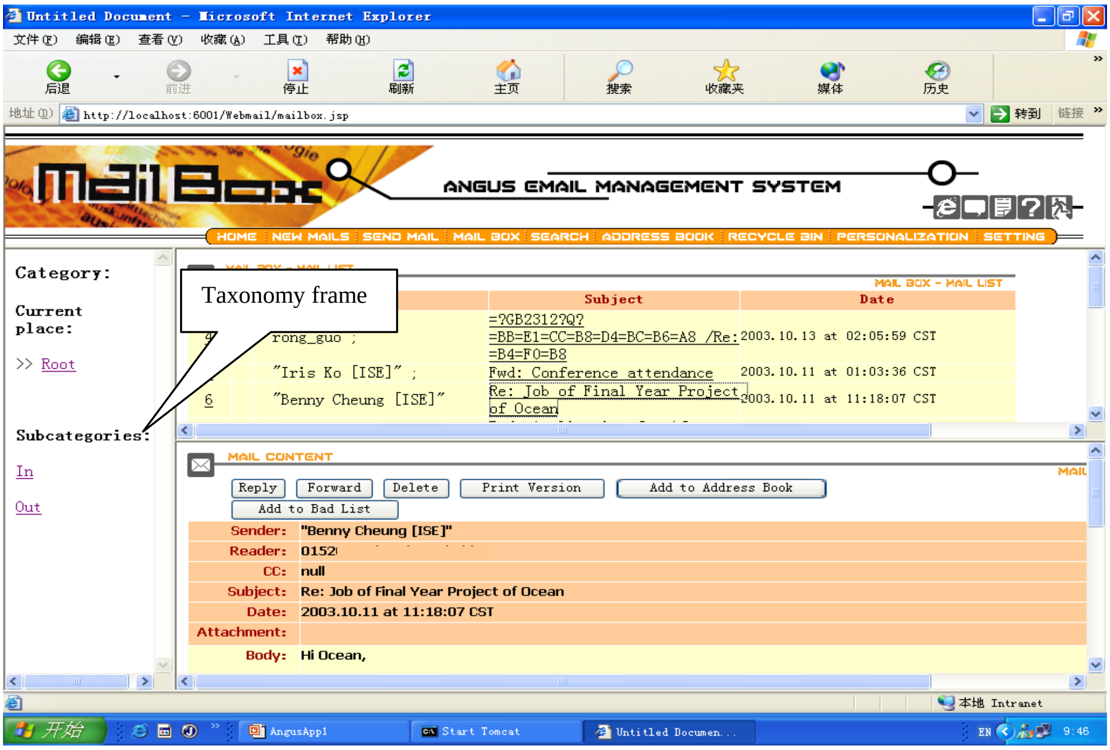
7.3 Operation of application

To test its efficiency and verify the system MAKES was implemented in the Angus Co. Ltd. The main operational steps and results are illustrated below:

7.3.1 Text mining and taxonomy

In the web-based MAKES the interface, with a personal taxonomy as shown in Figure 7.6 is provided according to the user's role. Figure 7.7(a) shows the interface for inputting search keywords while the search results are shown in Figure 7.7(b). A category is depicted according to the field associated with the search keywords. Figure 7.7(b) expresses one kind of pattern of a dynamic taxonomy. Figure 7.8 displays a sub-window with a new taxonomy entity produced by the agent-based dynamic taxonomy system. The concepts shown in the new concept frame, as illustrated in Figure 7.8, are derived from the statistics derived from a large number of email messages.

The employees of Angus send and receive emails in MAKES, and all email messages are stored in a centralized web server which is constructed using Linux and MySQL. The desired information can be searched and retrieved in the Intranet and Internet easily and quickly by using the relevant functions.


Untitled Document - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 前进 停止 刷新 主页 搜索 收藏夹 媒体 历史
地址(D) http://localhost:6001/Webmail/mailbox.jsp
转到 链接
MailBox
ANGUS EMAIL MANAGEMENT SYSTEM
HOME NEW MAILS SEND MAIL MAIL BOX SEARCH ADDRESS BOOK RECYCLE BIN PERSONALIZATION SETTING
Taxonomy frame
MAIL BOX - MAIL LIST
MAIL CONTENT
Reply
Forward
Delete
Print Version
Add to Address Book
Add to Bad List
本地 Intranet
开始
AngusApp1
Start Tomcat
Untitled Documen...
EN
9:46

Category:

Current place:

>> Root

Subcategories:

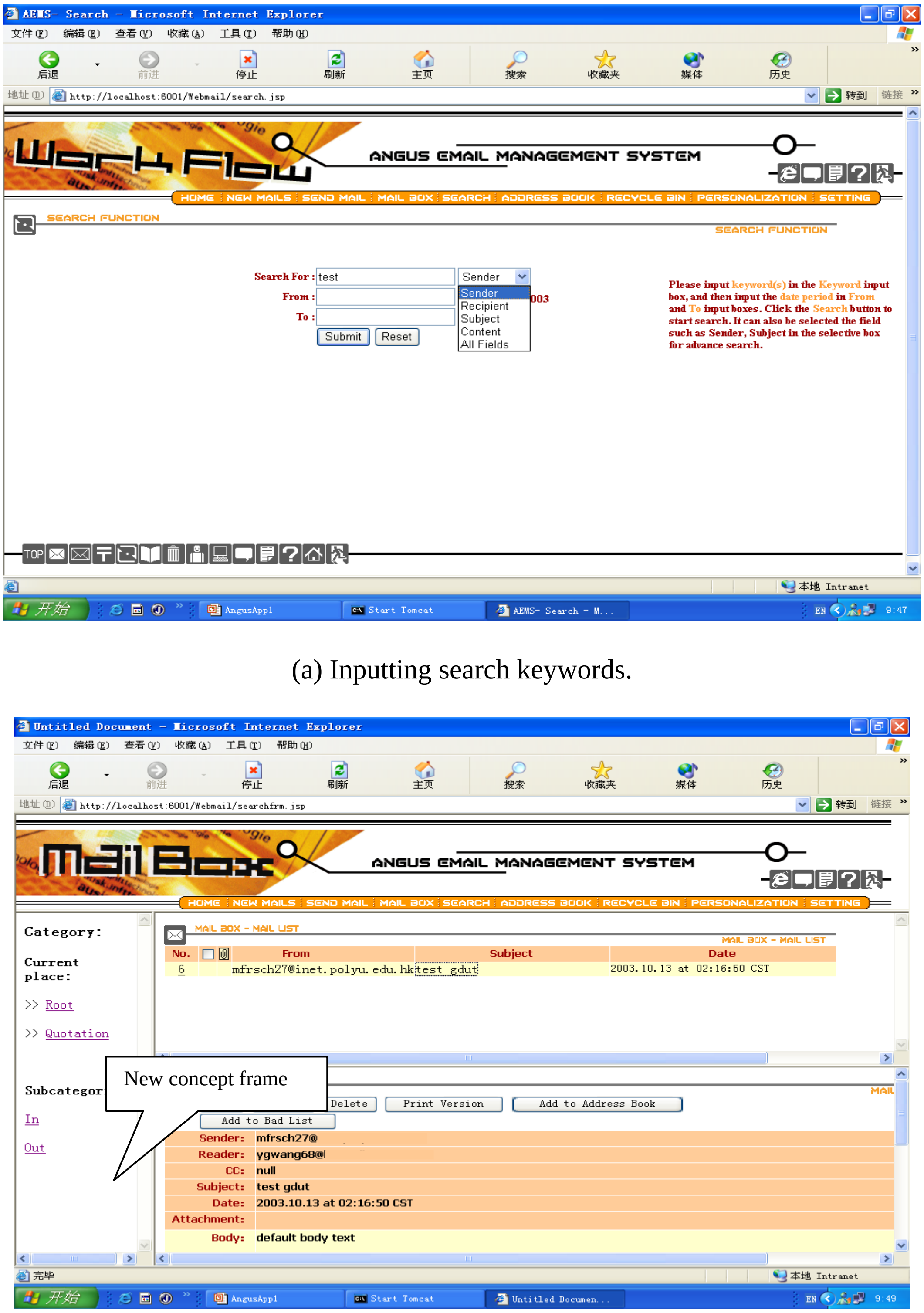
In

Out

		Subject	Date
4	rong_guo ;	=?GB2312?Q? =BB=E1=CC=B8=D4=BC=B6=A8 /Re: =B4=F0=B8	2003.10.13 at 02:05:59 CST
	"Iris Ko [ISE]" ;	Fwd: Conference attendance	2003.10.11 at 01:03:36 CST
6	"Benny Cheung [ISE]"	Re: Job of Final Year Project of Ocean	2003.10.11 at 11:18:07 CST

Sender:	"Benny Cheung [ISE]"
Reader:	0152
CC:	null
Subject:	Re: Job of Final Year Project of Ocean
Date:	2003.10.11 at 11:18:07 CST
Attachment:	
Body:	Hi Ocean,

Figure 7.6 Browsing email message.


AEMS- Search - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 前进 停止 刷新 主页 搜索 收藏夹 媒体 历史
地址(D) http://localhost:6001/Webmail/search.jsp
转到 链接
Work Flow
ANGUS EMAIL MANAGEMENT SYSTEM
HOME NEW MAILS SEND MAIL MAIL BOX SEARCH ADDRESS BOOK RECYCLE BIN PERSONALIZATION SETTING
SEARCH FUNCTION
Submit
Reset
TOP
本地 Intranet
开始 AngusApp1 Start Tomcat AEMS- Search - M... EN 9:47

Search For : test [Sender]

Sender
Recipient
Subject
Content
All Fields

From :

To :

Please input keyword(s) in the Keyword input box, and then input the date period in From and To input boxes. Click the Search button to start search. It can also be selected the field such as Sender, Subject in the selective box for advance search.

(a) Inputting search keywords.

Untitled Document - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 前进 停止 刷新 主页 搜索 收藏夹 媒体 历史
地址(D) http://localhost:6001/Webmail/searchfrm.jsp
转到 链接
Mail Box
ANGUS EMAIL MANAGEMENT SYSTEM
HOME NEW MAILS SEND MAIL MAIL BOX SEARCH ADDRESS BOOK RECYCLE BIN PERSONALIZATION SETTING
MAIL BOX - MAIL LIST
Delete
Print Version
Add to Address Book
Add to Bad List
MAIL
完毕
本地 Intranet
开始 AngusApp1 Start Tomcat Untitled Documen... EN 9:49

Category:

Current place:

>> Root

>> Quotation

Subcategor

In

Out

New concept frame

No.	□	From	Subject	Date
6		mfrsch27@inet.polyu.edu.hk	test gdut	2003.10.13 at 02:16:50 CST

Sender: mfrsch27@

Reader: ygwang68@

CC: null

Subject: test gdut

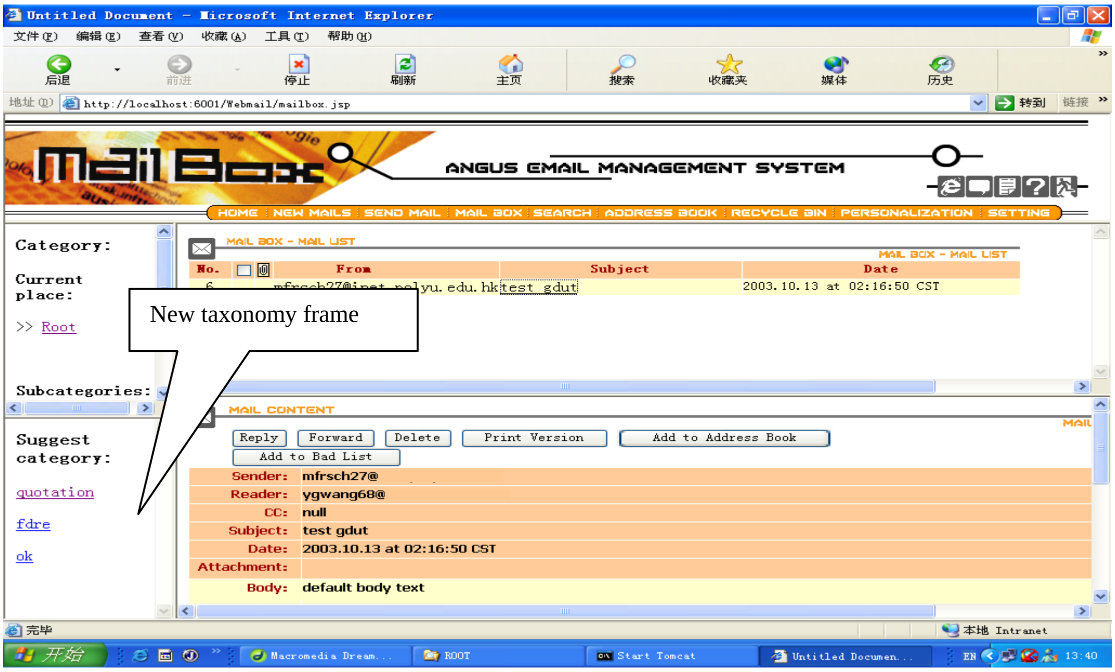
Date: 2003.10.13 at 02:16:50 CST

Attachment:

Body: default body text

(b) Browsing search result.

Figure 7.7 Search function with dynamic taxonomy.


Untitled Document - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 前进 停止 刷新 主页 搜索 收藏夹 媒体 历史
地址(D) http://localhost:6001/Webmail/mailbox.jsp
转到 链接
MailBox
ANGUS EMAIL MANAGEMENT SYSTEM
HOME NEW MAILS SEND MAIL MAIL BOX SEARCH ADDRESS BOOK RECYCLE BIN PERSONALIZATION SETTING
MAIL BOX - MAIL LIST
MAIL CONTENT
Reply
Forward
Delete
Print Version
Add to Address Book
Add to Bad List
New taxonomy frame
完毕
本地 Intranet
开始
Macromedia Dream...
ROOT
Start Tomcat
Untitled Documen...
EN
13:40

Category:

Current place:

>> Root

Subcategories:

Suggest category:

quotation

fdre

ok

No.	□	From	Subject	Date
6		mfrsch27@inet.polyu.edu.hk	test gdut	2003.10.13 at 02:16:50 CST

Sender: mfrsch27@

Reader: ygwang68@

CC: null

Subject: test gdut

Date: 2003.10.13 at 02:16:50 CST

Attachment:

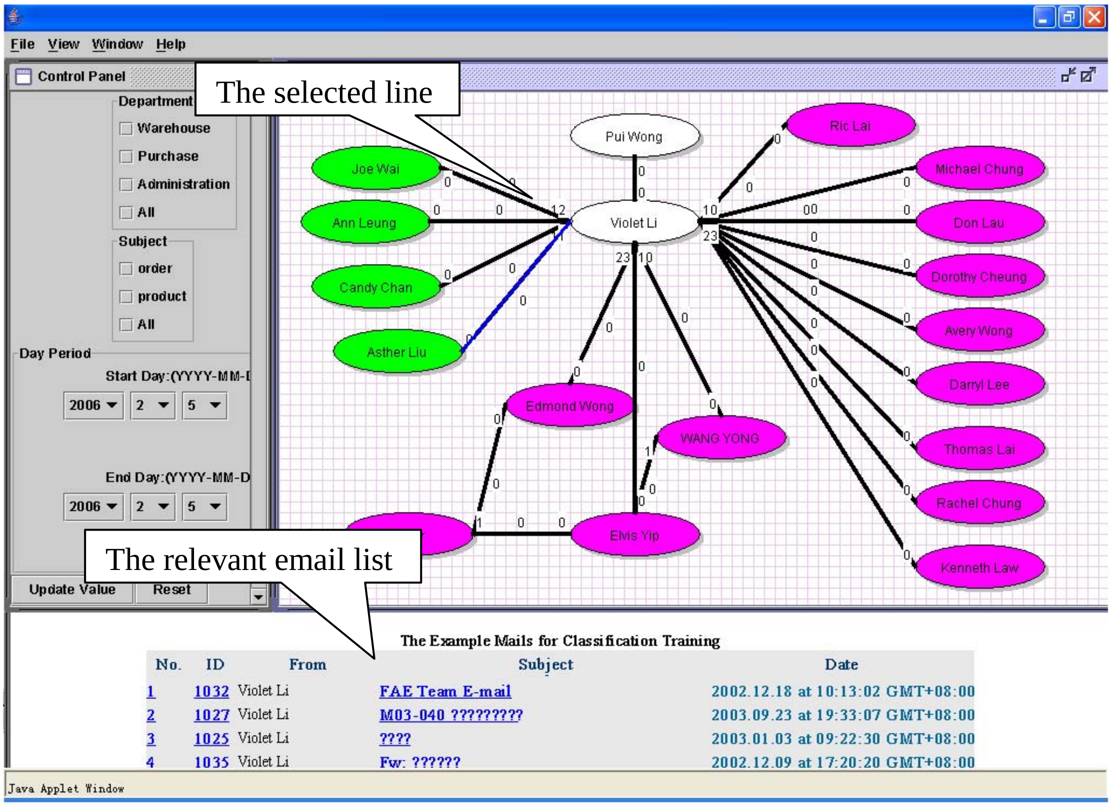
Body: default body text

Figure 7.8 A dynamic taxonomy with suggestion about new categories.

7.3.2 Modeling the knowledge flow

As a result of text mining and classification, all email messages are featured. DKFM is built and is shown in Figure 7.9. The nodes denote the names of knowledge entities, such as employees and vendors. The backgrounds of the nodes in different colours, shown in Figure 7.9, denote different departments or organizations. The lines denote the relationships of correspondence among the nodes. The numbers at the two ends of a line denotes the number of email messages. At the bottom of the interface of Figure 7.9, there is a sub-window which tabulates the list of relevant email messages attached to the line selected in the DKFM.

The initial diagram of DKFM is automatically produced by the system, and then the nodes can be dragged to any place in the diagram window by the user so as to re-align the diagram according to the user's preferences. The new diagram of the pattern is recorded and redisplayed.



The selected line

The relevant email list

Department
- □ Warehouse
- □ Purchase
- □ Administration
- □ All

Subject
- □ order
- □ product
- □ All

Day Period

Start Day:(YYYY-MM-D

2006 2 5

End Day:(YYYY-MM-D

2006 2 5

The Example Mails for Classification Training

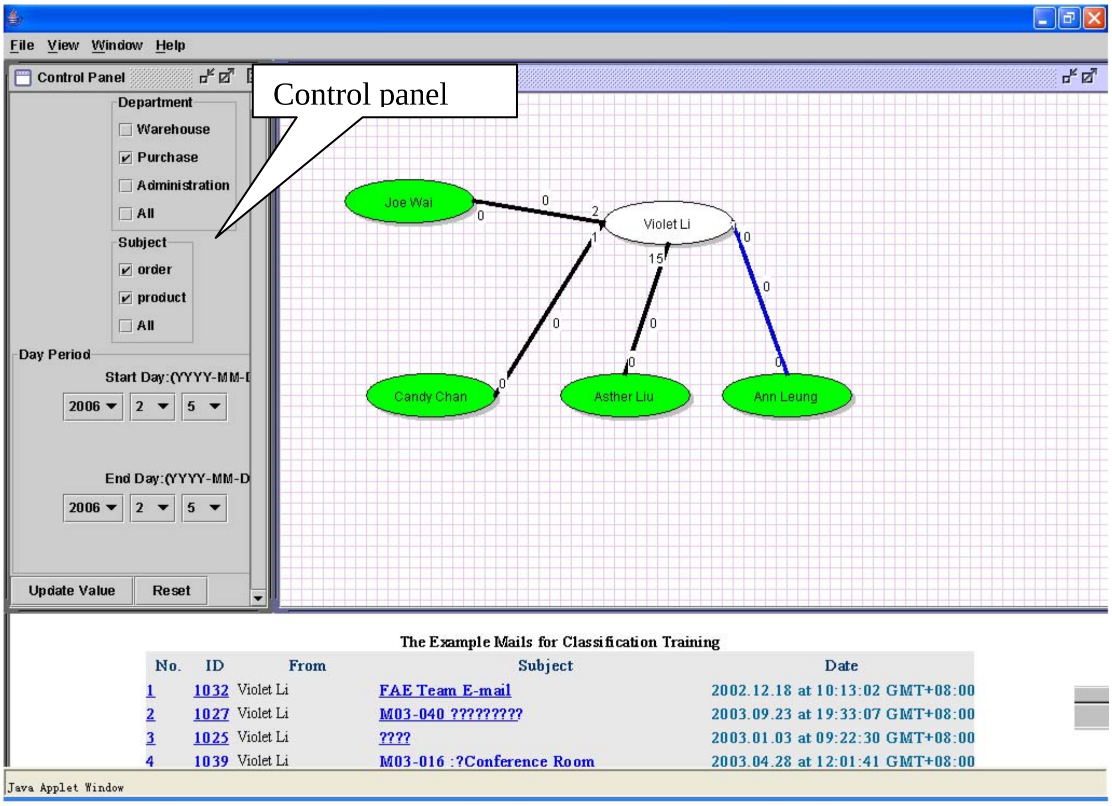
No.	ID	From	Subject	Date
1	1032	Violet Li	FAE Team E-mail	2002.12.18 at 10:13:02 GMT+08:00
2	1027	Violet Li	M03-040 ?????????	2003.09.23 at 19:33:07 GMT+08:00
3	1025	Violet Li	????	2003.01.03 at 09:22:30 GMT+08:00
4	1035	Violet Li	Fw: ??????	2002.12.09 at 17:20:20 GMT+08:00

Figure 7.9 An interface of dynamic knowledge flow model.

If people want to analyze the knowledge work, they can use the control panel to set the conditions to view the relevant model of the knowledge flow. There are three options which denote three dimensions which are: department, subject, and time. When the user sets the options, the relevant pattern of DKFM is displayed in Figure 7.10.

The conditions in Figure 7.10 are set so as the Department is 'Purchase' and the subject includes 'order and product', but the duration is not set. This means that people want to view the pattern of knowledge flow concerning the email messages that are sent and received by the employees of the Purchasing Department where the content of the emails is concerned with matters pertaining to order, and product. There is no time limitation about the time when the email messages are sent or received.

As shown in Figure 7.10, the new pattern of DKFM shows only five nodes, four of which have a green background and belong to the Purchasing Department. The numbers attached to the lines denote the number of email messages which are related to the subjects of order and product. The email messages which satisfy the conditions are listed in the bottom sub-window of Figure 7.10.



Department
☐ Warehouse
☑ Purchase
☐ Administration
☐ All

Subject
☑ order
☑ product
☐ All

Day Period
Start Day:(YYYY-MM-D
2006 2 5
End Day:(YYYY-MM-D
2006 2 5

The Example Mails for Classification Training

No.	ID	From	Subject	Date
1	1032	Violet Li	FAE Team E-mail	2002.12.18 at 10:13:02 GMT+08:00
2	1027	Violet Li	M03-040 ?????????	2003.09.23 at 19:33:07 GMT+08:00
3	1025	Violet Li	????	2003.01.03 at 09:22:30 GMT+08:00
4	1039	Violet Li	M03-016 :?Conference Room	2003.04.28 at 12:01:41 GMT+08:00

Figure 7.10 An new pattern of DKFM according to the settings of control panel.

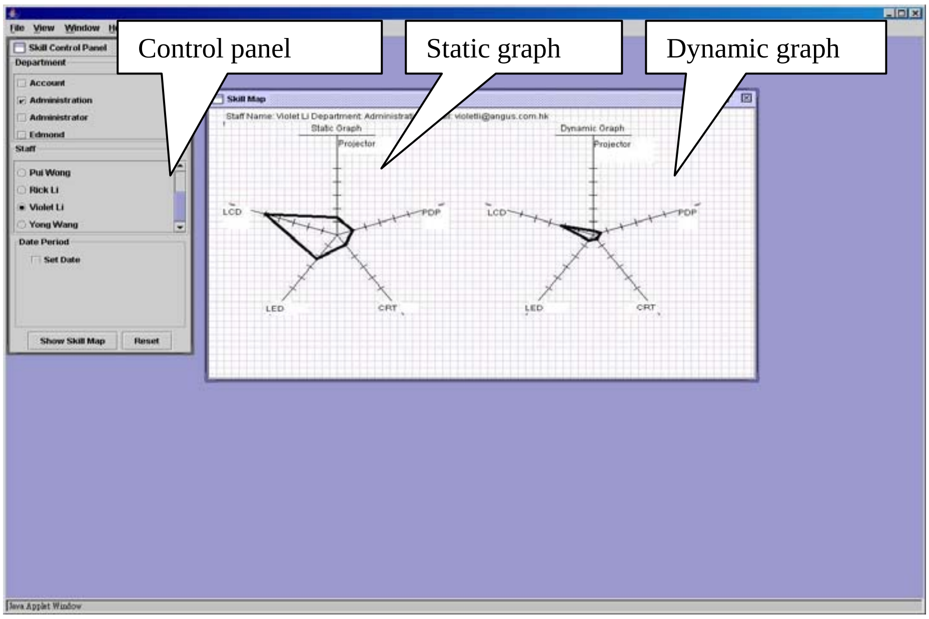
From the DKFM, the managers of Angus can see and analyze the patterns of knowledge work from various views and aspects, and try to find out the corresponding relationships between their pattern of knowledge work and the best practice. For example, they may discover and deduce some characteristics through analyzing the work pattern of the best employees from the pattern of DKFM in order to share and diffuse the best work pattern.

7.3.3 Analyzing knowledge capability of employees

In the model of knowledge flow, managers can click any nodes to demonstrate its distribution in knowledge domains in the form of a radar graph. This means they can view the knowledge capability of employees by using a radar graph which includes some knowledge domains. In other words, the radar graph of knowledge domains reflects the knowledge entity’s competence to some extent.

Figure 7.11 shows the model of knowledge capability of employee “Violet Li” from “Administration”. There are five knowledge domains on LED, LCD, CRT, PDP, and Projector, in the radar graph. This means the messages which are processed by Violet Li are mainly about these five knowledge domains. LCD is the most important field in Violet Li’s work in these five knowledge domains.

Similar to the window of DKFM, the left part of the screen in Figure 7.11 is the control panel. The user can use the panel to choose the department or member of staff, and even a data set. The related radar graph is constructed according to the conditions set by the control panel. There are two graphs in KCM. The left graph is a static graph which displays the measures of five pre-defined knowledge domains. The right one is called a dynamic graph which expresses the measures of the top five knowledge domains the employee works on. The pre-defined knowledge domains denote the area the user of MAKES is interested in, and the top knowledge domains are the areas the employee works mainly on.



Department
☐ Account
☑ Administration
☐ Administrator
☐ Edmond

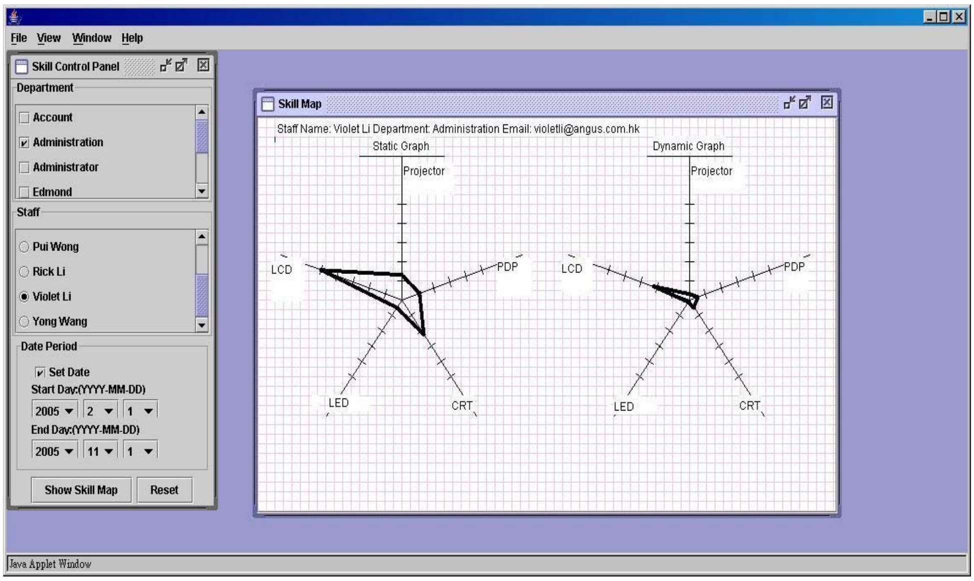
Staff
○ Pui Wong
○ Rick Li
● Violet Li
○ Yong Wang

Date Period
☐ Set Date

Staff Name: Violet Li Department: Administrat violetli@angus.com.hk

Figure 7.11 A radar graph of knowledge capability model of a knowledge entity.

MAKES informs the user about the trends in technology development by comparing the changes of knowledge domains in different radar graphs of KCM of the same knowledge entity in different time periods. Figure 7.12 shows the measure of knowledge domains of KCM of Angus in 2005, where Angus is regarded as a knowledge entity. Meanwhile, Figure 7.13 displays the new measure of knowledge domains of KCM of Angus in 2006. Comparing Figure 7.12 and Figure 7.13, the frequency of use of the knowledge domain “LED” is greater in 2006 than in 2005. It may mean LED is an increasingly valuable technology to which more attention should be paid in the future. In fact, LCD is always the hottest technology field in these two years.



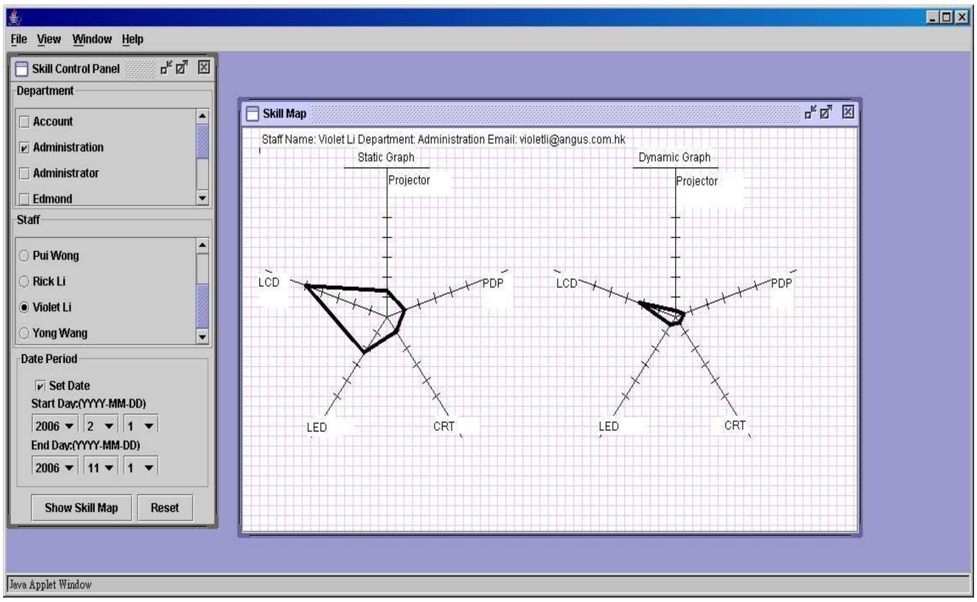
Department
□ Account
☑ Administration
□ Administrator
□ Edmond

Staff
○ Pui Wong
○ Rick Li
● Violet Li
○ Yong Wang

Date Period
☑ Set Date
Start Day:(YYYY-MM-DD) 2005 2 1
End Day:(YYYY-MM-DD) 2005 11 1

Staff Name: Violet Li Department: Administration Email: violetli@angus.com.hk

Figure 7.12 A radar graph of KCM of Angus in 2005.



Department
□ Account
☑ Administration
□ Administrator
□ Edmond

Staff
○ Pui Wong
○ Rick Li
● Violet Li
○ Yong Wang

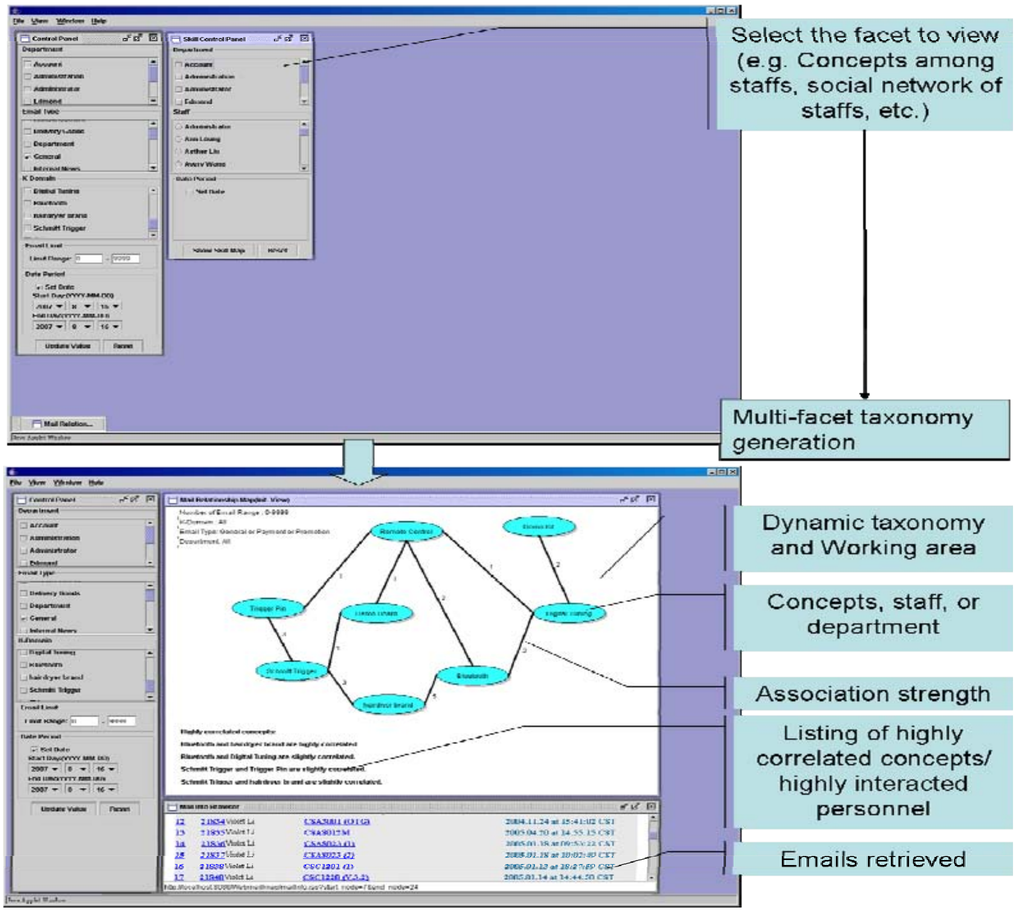
Date Period
☑ Set Date
Start Day:(YYYY-MM-DD) 2006 2 1
End Day:(YYYY-MM-DD) 2006 11 1

Staff Name: Violet Li Department: Administration Email: violetli@angus.com.hk

Figure 7.13 A radar graph of KCM of Angus in 2006.

7.3.4 Multi-faceted navigation for unstructured information

The multi-faceted navigation interface of MAKES for unstructured information is shown in Figure 7.14. In this interface, a user can setup the searching criteria to include department, personnel, concepts, date, and select the analysis type (i.e. Concept relationship or social network). Then the knowledge assets are filtered. The filtering is based on the criteria that are entered, and then, the corresponding knowledge model is generated.


Select the facet to view (e.g. Concepts among staffs, social network of staffs, etc.)
Multi-facet taxonomy generation
Dynamic taxonomy and Working area
Concepts, staff, or department
Association strength
Listing of highly correlated concepts/ highly interacted personnel
Emails retrieved
Control Panel
Skill Control Panel
Show Skill Map
Update Value
Reset
Remote Control
Trigger Pin
Schmitt Trigger
Bluetooth
Digital Tuning
Mail Relationship Map

Number of Email Range : 0-9999
K-Domain : All
Email Type: General or Payment or Promotion
Department: All

Highly correlated concepts:
Bluetooth and Digital Tuning are slightly correlated.
Schmitt Trigger and Trigger Pin are slightly correlated.

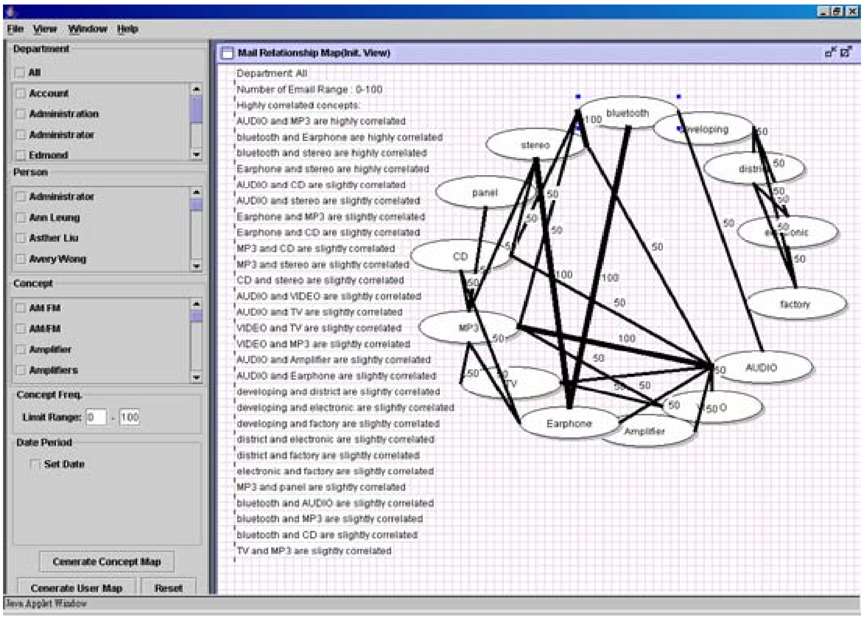
12	21854 Violet Li	CSA5801 (O1G)	2004.11.24 at 15:41:02 CST
13	21855 Violet Li	CSA8017M	2005.04.20 at 14:55:15 CST
14	21836 Violet Li	CSA5023 (1)	2005.01.18 at 09:53:22 CST
15	21837 Violet Li	CSA5023 (2)	2005.01.18 at 10:02:40 CST
16	21838 Violet Li	CSC1201 (1)	2005.01.12 at 18:27:59 CST
17	21840 Violet Li	CSC1210 (V.3.2)	2005.01.14 at 14:44:50 CST

Figure 7.14 A schema of multi-faceted navigation for unstructured information.

The schemas of a concept relationship map and social network are shown in Figures 7.15 and Figure 7.16, respectively. The concept relationship map in Figure 7.15 depicts the relationships among various concepts and the number of emails containing those concepts, while the social network depicts the relationships among various knowledge workers and the unstructured information, as shown in Figure 7.16. The text narrative about the relationships among the concepts and knowledge works is listed at the left of the graphs in Figure 7.15 and Figure 7.16, respectively.

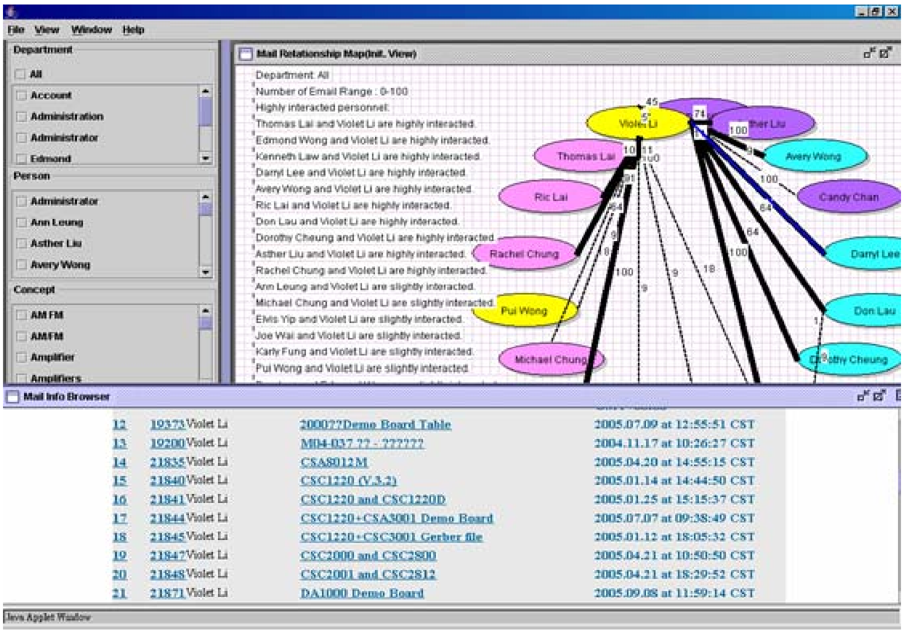
The knowledge inventory is reported as shown in Figure 7.17 and Figure 7.18 where the elicited concepts and critical users report generated from the Angus's MAKES can be found. The list of concepts in Figure 7.17 shows the name, the total number of occurrences of the concept in all the messages, and the total number of messages in which the concept occurs. The importance of every concept is measured and classified in the field of concept type, in the list of concepts in Figure 7.17. Figure 7.18 lists the information about the users. The information includes the group or department which the user belongs to, and the total number of concepts which occur in the user's email messages. The importance of the user is rated according to his or her performance in knowledge work.

Graphical representations of the clusters of the elicited concepts and knowledge workers are shown in Figures 7.19 and Figure 7.20, respectively. A snapshot of the knowledge inventory report in Figure 7.19 represents the distribution of concepts in two-dimensional space where the horizontal axis is the number of users and the vertical axis is the number of emails. Figure 7.20 displays the distribution of concepts of a user in the same two-dimensional space. From these graphs, people can identify some key concepts and how often they are referred to.



Department: All
Number of Email Range : 0-100
Highly correlated concepts:
AUDIO and MP3 are highly correlated
bluetooth and Earphone are highly correlated
bluetooth and stereo are highly correlated
Earphone and stereo are highly correlated
AUDIO and CD are slightly correlated
AUDIO and stereo are slightly correlated
Earphone and MP3 are slightly correlated
Earphone and CD are slightly correlated
MP3 and CD are slightly correlated
MP3 and stereo are slightly correlated
CD and stereo are slightly correlated
AUDIO and VIDEO are slightly correlated
AUDIO and TV are slightly correlated
VIDEO and TV are slightly correlated
VIDEO and MP3 are slightly correlated
AUDIO and Amplifier are slightly correlated
AUDIO and Earphone are slightly correlated
developing and district are slightly correlated
developing and electronic are slightly correlated
developing and factory are slightly correlated
district and electronic are slightly correlated
district and factory are slightly correlated
electronic and factory are slightly correlated
MP3 and panel are slightly correlated
bluetooth and AUDIO are slightly correlated
bluetooth and MP3 are slightly correlated
bluetooth and CD are slightly correlated
TV and MP3 are slightly correlated

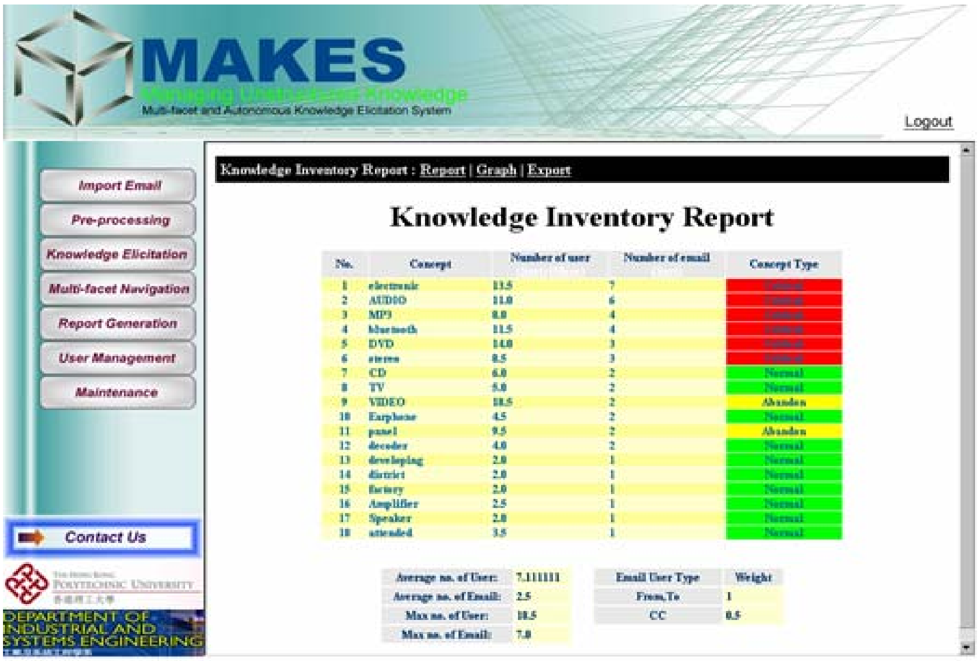
Figure 7.15 A snapshot of a concept relationship model.



Department: All
Number of Email Range : 0-100
Highly interacted personnel:
Thomas Lai and Violet Li are highly interacted.
Edmond Wong and Violet Li are highly interacted.
Kenneth Law and Violet Li are highly interacted.
Darryl Lee and Violet Li are highly interacted.
Avery Wong and Violet Li are highly interacted.
Ric Lai and Violet Li are highly interacted.
Don Lau and Violet Li are highly interacted.
Dorothy Cheung and Violet Li are highly interacted.
Asther Liu and Violet Li are highly interacted.
Rachel Chung and Violet Li are highly interacted.
Ann Leung and Violet Li are slightly interacted.
Michael Chung and Violet Li are slightly interacted.
Elvis Yip and Violet Li are slightly interacted.
Joe Wai and Violet Li are slightly interacted.
Karly Fung and Violet Li are slightly interacted.
Pui Wong and Violet Li are slightly interacted.

12	19373	Violet Li	200077Demo Board Table	2005.07.09 at 12:55:51 CST
13	19200	Violet Li	M04-037 ?? - ??????	2004.11.17 at 10:26:27 CST
14	21835	Violet Li	CSA8012M	2005.04.20 at 14:55:15 CST
15	21840	Violet Li	CSC1220 (V.3.2)	2005.01.14 at 14:44:50 CST
16	21841	Violet Li	CSC1220 and CSC1220D	2005.01.25 at 15:15:37 CST
17	21844	Violet Li	CSC1220+CSA3001 Demo Board	2005.07.07 at 09:38:49 CST
18	21845	Violet Li	CSC1220+CSC3001 Gerber file	2005.01.12 at 18:05:32 CST
19	21847	Violet Li	CSC2000 and CSC2800	2005.04.21 at 10:50:50 CST
20	21848	Violet Li	CSC2001 and CSC2812	2005.04.21 at 18:29:52 CST
21	21871	Violet Li	DA1000 Demo Board	2005.09.08 at 11:59:14 CST

Figure 7.16 A snapshot of a social network.


MAKES
Managing Unstructured Knowledge
Multi-facet and Autonomous Knowledge Elicitation System
Logout
Import Email
Pre-processing
Knowledge Elicitation
Multi-facet Navigation
Report Generation
User Management
Maintenance
Contact Us
The Hong Kong Polytechnic University
香港理工大學
DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

Knowledge Inventory Report : Report | Graph | Export

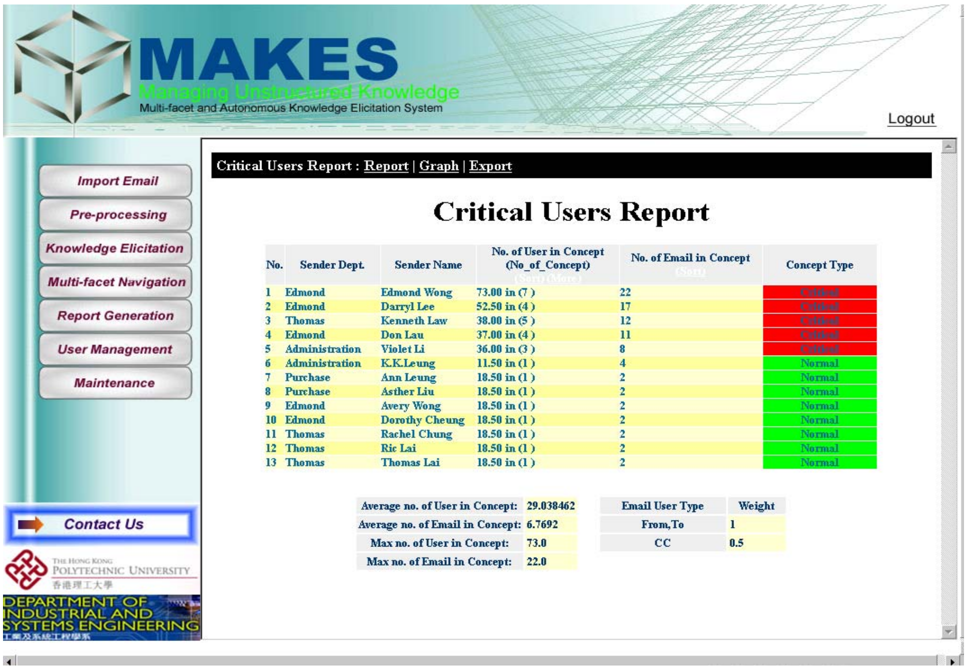
Knowledge Inventory Report

No.	Concept	Number of user	Number of email	Concept Type
1	electronic	13.5	7	Critical
2	AUDIO	11.0	6	Critical
3	MP3	8.0	4	Critical
4	bluetooth	11.5	4	Critical
5	DVD	14.0	3	Critical
6	stereo	8.5	3	Critical
7	CD	6.0	2	Normal
8	TV	5.0	2	Normal
9	VIDEO	18.5	2	Abandon
10	Earphone	4.5	2	Normal
11	panel	9.5	2	Abandon
12	decoder	4.0	2	Normal
13	developing	2.0	1	Normal
14	district	2.0	1	Normal
15	factory	2.0	1	Normal
16	Amplifier	2.5	1	Normal
17	Speaker	2.0	1	Normal
18	attended	3.5	1	Normal

Average no. of User:	7.111111
Average no. of Email:	2.5
Max no. of User:	18.5
Max no. of Email:	7.0

Email User Type	Weight
From,To	1
CC	0.5

Figure 7.17 The snapshot of knowledge inventory report.


MAKES
Managing Unstructured Knowledge
Multi-facet and Autonomous Knowledge Elicitation System
Logout
Import Email
Pre-processing
Knowledge Elicitation
Multi-facet Navigation
Report Generation
User Management
Maintenance
Contact Us
The Hong Kong Polytechnic University
香港理工大學
DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

Critical Users Report : Report | Graph | Export

Critical Users Report

No.	Sender Dept.	Sender Name	No. of User in Concept (No_of_Concept) (Sort) (More)	No. of Email in Concept (Sort)	Concept Type
1	Edmond	Edmond Wong	73.00 in (7)	22	Critical
2	Edmond	Darryl Lee	52.50 in (4)	17	Critical
3	Thomas	Kenneth Law	38.00 in (5)	12	Critical
4	Edmond	Don Lau	37.00 in (4)	11	Critical
5	Administration	Violet Li	36.00 in (3)	8	Critical
6	Administration	K.K.Leung	11.50 in (1)	4	Normal
7	Purchase	Ann Leung	18.50 in (1)	2	Normal
8	Purchase	Asther Liu	18.50 in (1)	2	Normal
9	Edmond	Avery Wong	18.50 in (1)	2	Normal
10	Edmond	Dorothy Cheung	18.50 in (1)	2	Normal
11	Thomas	Rachel Chung	18.50 in (1)	2	Normal
12	Thomas	Ric Lai	18.50 in (1)	2	Normal
13	Thomas	Thomas Lai	18.50 in (1)	2	Normal

Average no. of User in Concept:	29.038462
Average no. of Email in Concept:	6.7692
Max no. of User in Concept:	73.0
Max no. of Email in Concept:	22.0

Email User Type	Weight
From,To	1
CC	0.5

Figure 7.18 The snapshot of critical users report.

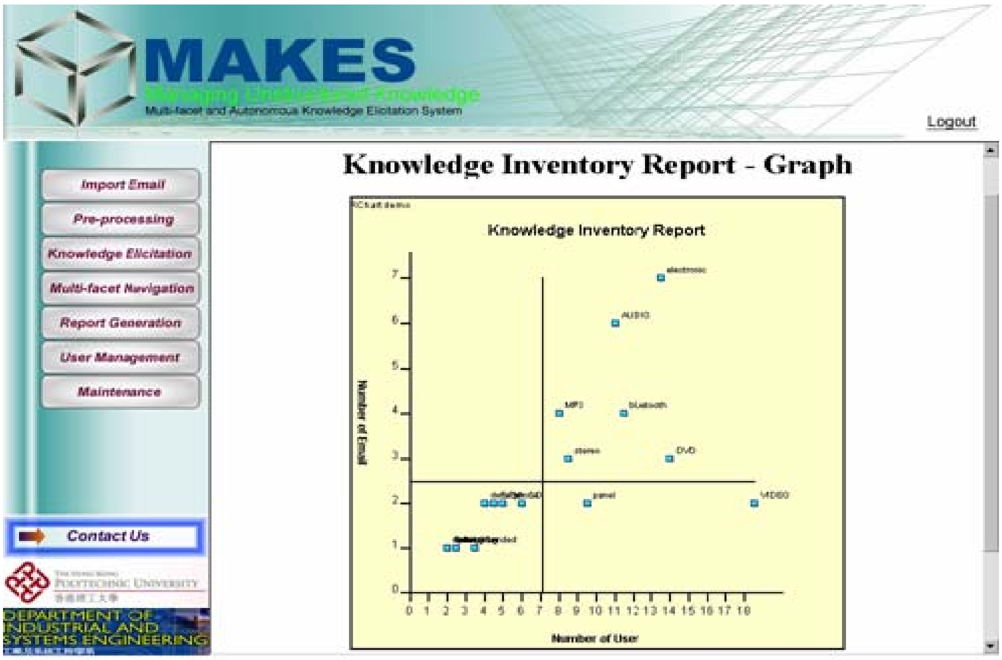


Figure 7.19 Graphical representation of the clustered concepts.

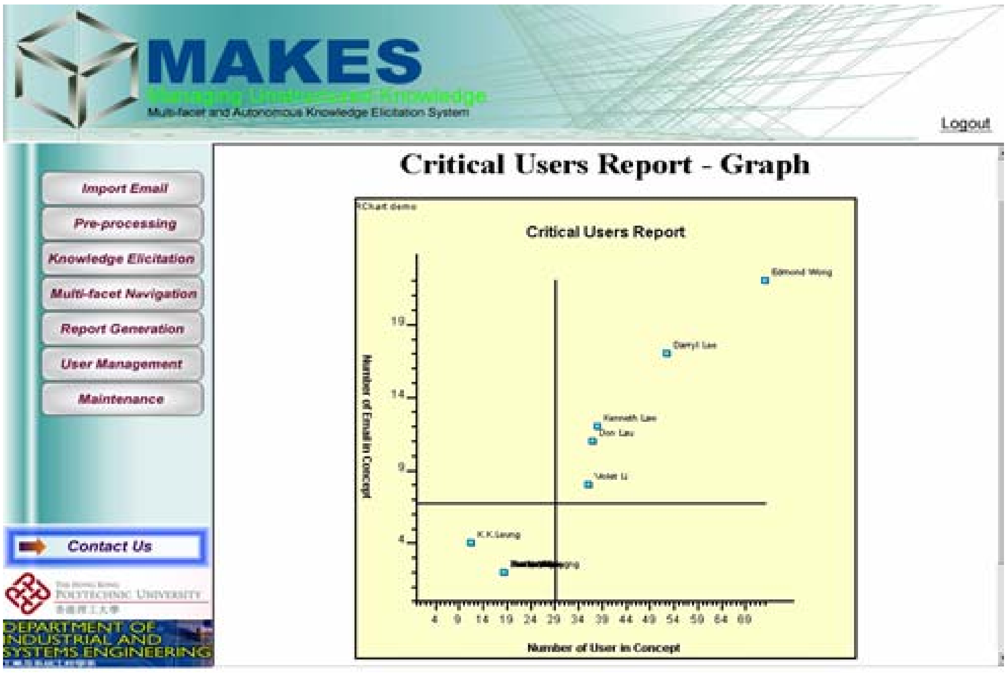


Figure 7.20 Graphical representation of the clustered knowledge workers.

7.4 Quantitative analysis of the system performance

A quantitative experiment has been carried out for measuring the performance of the Concept Relationship Model (CRM). The experiment flow is shown in Figure 7.21. A total of 29 emails were selected from the company. They were analyzed by an expert group which was composed of members of staff of the company. The key concepts were first extracted manually. Then, they were converted into concept relationship models both by the expert group and by CRM described as below:

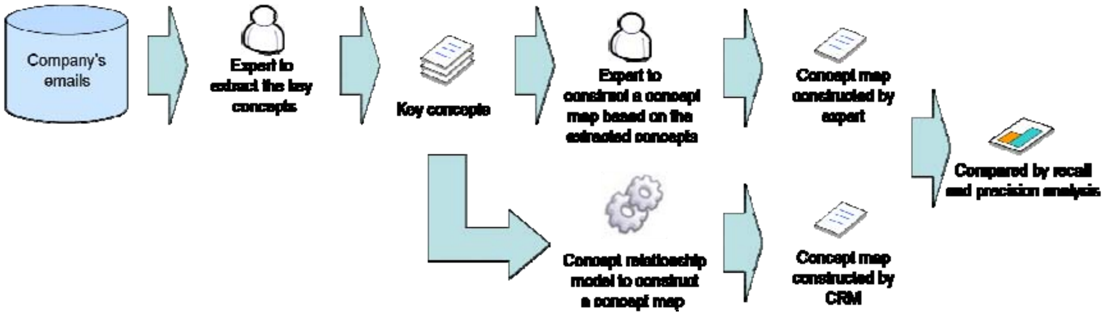

Company's emails
Expert to extract the key concepts
Key concepts
Expert to construct a concept map based on the extracted concepts
Concept map constructed by expert
Concept relationship model to construct a concept map
Concept map constructed by CRM
Compared by recall and precision analysis

Figure 7.21 Experimental flow of the concept relationship model evaluation.

Part of the concept relationship map constructed by the expert group is shown in Figure 7.22. The boxes indicate the key concepts and the lines represent the relationships among the concepts. The correctness of the concept relationships suggested by the CRM was measured by comparing them with the relationships suggested by the expert group, based on the recall and precision analysis. The recall and precision are determined based on Equations (7.1) and (7.2) respectively.

$$recall = \frac{M}{E} \tag{7.1}$$

$$precision = \frac{M}{C} \tag{7.2}$$

Where

M is the number of matched associations between the expert group and the CRM,

E is the number of associations suggested by expert group and

C is the number of associations suggested by CRM.

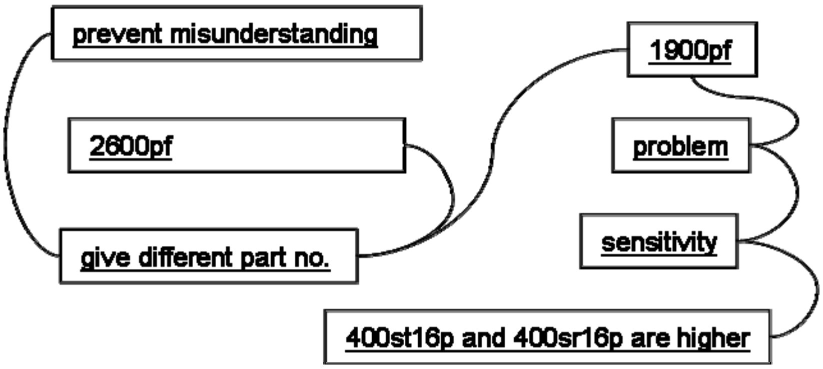


Figure 7.22 Part of the concept relationship constructed by the expert group.

7.5 Results and discussion

With the initial trial implementation of MAKES in Angus, encouraging results were achieved and a number of qualitatively and quantitative potential advantages were realized as shown below:

i) Valuable knowledge inherent in the massive email communication between the company staff, the customers and the suppliers is uncovered. Concept relationships, social network, and the patterns of knowledge flow, are derived based on mining the unstructured information in emails. Some key actors and concepts are identified by Angus' managers, evidently from these graphs. It is very useful to human resource management and for the development of technology.

ii）The knowledge needed for providing the most advanced products to meet customer needs is leveraged. Analyzing the evolution of concept map along with the time reveals the trends in the use of technology and in the appearance of new technology. In particular, many technical enterprises became more interested learning about LED through analyzing the concept of map and knowledge inventory report. Hence, Angus will allocate more resources to collect information about LED and increase the company's knowledge about LED.

iii）Rapid change of taxonomy for the company's unstructured information (especially regarding new products and new market development) is achieved, and new knowledge and concepts can be automatically discovered. The dynamic taxonomy increases the efficiency of managing unstructured information. It enables new, additional categories relative to a concept to be identified as soon as the new concept appears frequently. Hence, LED has been added into the taxonomy of Angus.

iv）More relevant and concise search results are provided so that the users can identify the right knowledge more easily and efficiently. The Web-based search interface facilitates more efficient searching for information and knowledge. Dynamic taxonomy provides an efficient way and structure for users to browse the relevant information, and avoids the limitations of the traditional approach to searching.

v）Personal interest oriented and concise taxonomies are provided. The dynamic taxonomy in Angus’ MAKES is personal. It improves the efficiency of the user when browsing for information and knowledge.

vi）The time, the cost and the workload on taxonomy maintenance are reduced. The time spent on email classification and filing is dramatically reduced. As a rough estimation, the average time spent on email filing by each member of staff every day is 30 minutes x 66% + 90 minutes x 10% + 150 minutes x 19% + 210 minutes x 5% = 73.8 minutes = 1.2 hours. That mean all the staff may save about an hour each after the implementation of the MAKES. From the point of view of a company with 60 staff, about 18,936 working hours (263 days x 60 staff x 1.13 hours) can be saved in email classification annually.

vii）The results of the quantitative analysis of the performance of the MAKES show that 593 and 62 associations were suggested by the CRM and the expert group, respectively, and 62 associations are matched. Hence, the recall and precision rate are 10.5% and 100%, respectively.

The verification testing of MAKES for knowledge audit in Angus has been successfully performed and this trial application produced some valuable results. The results are encouraging with excellent precision although the recall rate is not high.

Chapter 8. Conclusion and Further Work

8.1 Contribution to new knowledge

A huge amount of unstructured information is scattered amongst various electronic resources in an organization. A lot of useful knowledge is embedded in this information. This information is a valuable intellectual asset of an organization. Very often, it is not attended to or managed. Traditional information management deals with formal, orderly and structured information in various databases and repositories, whereas from the information system's perspectives, knowledge management deals with informal and unstructured information. Such information is carried by the people associated with it and who work with it. Useful knowledge that can be extracted from these people-information pairs includes information about the knowledge owners and users, the knowledge items themselves, their demand and supply, their utilization and information about who are the critical knowledge workers, and details about the knowledge network. None of this information can be found in an organizational chart of an organization or from the work process. This information is often dynamic in nature and changes with the people, the working environment and in response to external relationships. This is especially true for knowledge work which deals with the manipulation of all sorts of information from various internal and external sources.

To manage such unstructured information effectively, a Multi-Faceted and

Automatic Knowledge Elicitation System (MAKES) is proposed, which allows for retrieving, automatically classifying, capturing and sharing valuable knowledge from a mass of unstructured information including multiple concepts at many levels of abstraction. In this thesis, several diagrammatic knowledge models are adopted to elicit, represent and organize this unstructured information and turn it into useful knowledge as desired by the users. These knowledge models depict knowledge entities and their relationships in the form of graphical patterns. The knowledge models in this thesis are: Concept Relationship Model (CRM), Dynamic Knowledge Flow Model (DKFM) and Knowledge Capability Model (KCM). These models uncover the inherent relationships among people and knowledge entities in various forms and from multiple views.

A dynamic taxonomy has also been constructed to classify a huge amount of complex, heterogeneous information, specifically email. The technology of term spotting is used to decompose unstructured information and build eigenvectors. A supervised learning mechanism is constructed and a statistics-based similarity algorithm is designed to classify messages automatically. At the same time, a knowledge mining algorithm is developed to automatically generate and self-maintain the evolution of the multi-faceted taxonomy maps based on the searching criteria, searching keywords, and the user behaviors of the knowledge workers. It saves the time and reduces the workload of the staff.

Concept Relationship Model (CRM) illustrates the relationships of knowledge concepts based on a thesaurus model. An algorithm called the Concept Relationship Exploring Technique (CRET) algorithm is designed to mine the co-occurrence relationships of concepts and build the probability model of the dependent relationships of concepts extracted from a very large amount of unstructured information. An assessment report of the concepts thus associated is generated automatically. For example, if the concepts are related to the technology items, then, the evolving trends of the technologies are shown.

Understanding how knowledge flows through an organization is critical to the increase of productivity of knowledge work. DKFM describes the patterns of flow of information among knowledge entities. Through analyzing information captured from knowledge work, DKFM can identify the sources and destinations of information and track the flow of information. Knowledge entities are identified and knowledge inventory is organized through analyzing the structures of DKFM. Critical knowledge entities and users are evaluated according to the amount of information which flows to and from them. Upper level analysis of DKFM for certain knowledge domains is provided and helps to uncover the pattern of information flow in those knowledge domains, e.g. financial, information technology. By making the communication flow transparent, processes of knowledge work can be made more explicit, thus allowing organizations to make better use of people by freeing them from being buried in conventional multilayer hierarchies and inefficient business processes. While the knowledge entities are only defined as people, groups, and organizations, a model of knowledge flow can be regarded as a social network. And then, the technologies of Social Network Analysis (SNA) can be adopted to analyze the pattern knowledge flow.

KCM is used to evaluate the knowledge capability of every knowledge worker.

The knowledge domains which a knowledge worker works on reveal which domains are the master areas the knowledge worker works on. KCM reflects which knowledge work a knowledge worker is engaged in and what knowledge a knowledge worker is frequently associated with. A report about knowledge capability of a knowledge worker or an organization can be generated based on KCM.

The MAKES has been applied in two cases for its verification, one in emergency management and one in a knowledge audit. Emergency management aims to reduce vulnerability to hazards and to increase the ability to cope with disasters. A key issue in emergency management is to have access to all relevant information from various experts, authorities, documents, cases, procedures, people affected which are often scattered, uncoordinated and unstructured; and to make well-informed decisions. Generally, the response to an emergency incident needs a lot of time to search for the necessary information. In the verification testing of this thesis, MAKES is integrated into an actual system in use for emergency management in Guang Zhou city, a metropolis of China, to provide relevant information and knowledge in the handling of emergency incidents. The relevant information about an emergency incident can be modeled automatically and dynamically. CRM gives hints for searching and capturing the relevant information and knowledge according to the key characters of the emergency. DKFM shows the pattern for making a rescue plan. Some key knowledge entities needed in response to the emergency are identified. KCM illustrates the knowledge capability of a knowledge entity. Finally, the main information and knowledge on how to manage the emergency can be accessed from MAKES through a multi-faceted navigation platform. Reports about the knowledge assets needed to handle an emergency are generated in order to improve the capability of decision making.

In an enterprise, a knowledge audit is usually recommended in organizations as an important first step prior to the launching of any knowledge management program. The purpose of a knowledge audit is to determine what knowledge is needed, from where knowledge is available and what is missing, who needs this knowledge, and how this knowledge can be applied. Traditional knowledge audits depend mainly on the labor of the auditor and take a long time to complete. Very often, the subjective factors from informants and interviewees and auditors can not be avoided. As a means of conducting a systematic examination and evaluation of organizational knowledge assets, MAKES has been used in the knowledge auditing of an electronic trading company, Angus Co. Through reconstructing the email system of Angus Co., the all the emails are captured to build knowledge models. A dynamic taxonomy is used for storing and searching emails. CRM has found new concepts and has enabled the trends in new technology to be forecasted. The key knowledge workers in Angus Co. have been identified and evaluated based on analyzing DKFM. Several reports about its knowledge assets are generated. All of the work is done by MAKES automatically and more efficiently than by using a traditional knowledge audit.

The major contributions of the research are described as follows:

i) The work has demonstrated that the approach to handling informal and unstructured data from electronic sources is different from handling orderly structured information which falls into the domain of information management. From the information systems perspective, knowledge management refers to the managing of informal and unstructured information that is associated with the people or knowledge workers possessing it. It is the people and their information that we are managing, not the information per se in knowledge management. One of the aims of managing the knowledge is to reveal the knowledge flow, the relationship between the knowledge entity (its people, knowledge items, networks etc.), the critical knowledge workers and the knowledge teams involved in a business process, through the various knowledge models constructed.

ii) The process to elicit the knowledge is designed and implemented in a system named as MAKES (multi-faceted and automatic knowledge elicitation system) to facilitate searching, navigation, analyzing, discovering and visualizing the right knowledge assets from among a very large amount of unstructured information. A knowledge mining algorithm is developed to automatically generate and self-maintain the evolution of the multi-faceted taxonomy maps based on the searching criteria, searching keywords, and the user behaviors of the knowledge workers. A purpose-built knowledge elicitation algorithm called the "Concept Relationship Exploring Technique (CRET)" to put new categories or abstractions in the multi-faceted taxonomy maps and various knowledge models for generating different reports is assembled. The modular design enables different applications and reports to be generated.

iii) A multi-linguistic processing ability for handling unstructured information is built into the system. The advantage of using technology for term spotting and statistics-based classification is that multi-lingual messages, such as those in English, Simplified Chinese, and Traditional Chinese, can be processed easily and efficiently. In Hong Kong and the mainland of China, multi-lingual messages are common. The ability to process multi-lingual messages is a very important feature of the system.

iv) The above system has been implemented for verification in two application cases, namely in emergency management deployed in Guang Zhou a municipal city and in knowledge auditing of a trading firm in Hong Kong. It is found that the system promises to be able to handle a vast amount of unstructured electronic information, to enable quickly access to relevant expertise in the handling of emergencies and thus to improve the quality of the decision making process in emergency management. MAKES can also identify important knowledge assets such as knowledge flow, critical knowledge workers and the knowledge network, all of which are embedded in the large amount of organizational emails.

8.2 Future work and suggestions

Although MAKES has been successfully prototyped and applied in two real cases, a lot of improvements need to be made to improve its performance, robustness and reliability. The future work is recommended as below.

i) Although MAKES can handle multi-lingual messages, the algorithm for text mining only adopts statistics-based technologies. The degree of precision with which the analyzing of the unstructured information is carried out is still not high. Adopting and synthesizing multiple technologies of natural language processing, such as conceptual dependency and lexical cohesion, will be a good approach to improve the efficiency and accuracy of the text mining for unstructured information.

ii) In this thesis, network graphs are the main forms of knowledge models. Some special structures of knowledge models imply something special about the nature of knowledge work. For example, a structure that contains a long chain may mean there is a long process in the work. A star structure denotes that the node in the center of the star structure may be a key actor or concept, etc. Analyzing the network structure of DKFM can be a valuable approach. Technologies of network analysis which are embedded with more artificial intelligence should be considered to enhance the capability of finding special patterns of knowledge work from DKFM. Based on complex and precise algorithms of text mining, the capability of identifying roles in knowledge work, e.g. suppliers and customers of knowledge in knowledge work, is very important. Analyzing and tracing the flow of knowledge will enhance knowledge searching and creation.

iii) Along with the development of knowledge work in an organization, patterns of knowledge work change from time to time. Users can compare the patterns which are formed in different periods of time to find out the gaps in patterns of knowledge flow and improve the process. Comparative analysis based on the time dimension will enhance the capability of forecasting trends in the use of technology and in knowledge.

iv) Further development of intelligent agents and the collaborative mechanism among multiple agents to develop the intelligent processing functions for MAKES, e.g. natural language processing, is needed.

v) Apart from the two specific cases in emergency management and knowledge audit, MAKES can be further generalized into other applications which need to manage a huge amount of information

With suitable enhancement, the approach proposed in this thesis can help the development of affordable and effective tools for enterprises and organizations to manage their unstructured information and turn it into a useful organizational knowledge asset.

Appendix 1. List of Experts for Evaluation of MAKES

Head of the expert group:

Tang, Yong, Vice Head of Faculty of Information, Sun Yat-Sen University, expert in information system and knowledge engineering.

Member of the expert group:

Fu, Xiufen, Professor, Vice Head of Faculty of Computer, Guang Dong University of Technology, expert in artificial intelligence.

He, Jin, Senior Engineer, Department of Technology, Bureau of Environmental Protection Bureau of Guang Dong Province Municipal Government.

Lei, Gao, Senior Engineer, Department of Fire-prevention, Fire Services Bureau of Guang Zhou City Municipal Government, expert in fire-prevention and fire-fighting.

Wu, Linhai, Senior Engineer, Vice Head of Department, Command Department, The Committee of City Management of Guang Zhou Municipal Government, expert in emergency management.

Appendix 2. List of Main Program Files of MAKES

Main program files are,

home.jsp	//home page
addressbook.jsp	//Manage address book
aailattach.jsp	//handle attachment of mail
mailextract.jsp	//extract concept from mail
mailfunction.jsp	//provide functions of email process
mailsendbean.jsp	//bean of sending email
mailupload.jsp	//upload files to MAKES
mailboxcategory.jsp	//show category of mails
mailcategorysuggest.jsp	//give suggestion for category
mailboxmain.jsp	//provide framework for browsing emails
mailboxbycontent.jsp	//browse content of mail
mailboxbyemailapplet.jsp	//display emails in applet
mailboxbyextract.jsp	//display emails according to concepts
smailboxapplet.jsp	//display searching result of emails in mailbox
mailcategoryfv.jsp	//construct eigenvector of category
mailconstructfv.jsp	//construct eigenvector of email
mailautoclassfy.jsp	//classify emails automatically
mailexamplelist.jsp	//list example emails
mailextract.jsp	//extract concepts from email
mailfrequency.jsp	//statistic frequency of concept in emails
AbstractEdge.java	//show edges of DKFM

AbstractNode.java	//show nodes of DKFM
ArrowHead.java	//show arrows of DKFM
ClassRelationshipEdge.java	//construct relationship of concpets
DB_initView.servlet.java	//initiate database
Direction.java	//statistic arrow directions of DKFM
Edge.java	//statistic edges of DKFM
FormLayout.java	//layout the form with control panel
Graph.java	//show graph of DKFM
GraphFrame.java	//deploy frame of graph of DKFM
GraphPanel.java	//deploy control panel of DKFM
MailDiagramGraph.java	//show mail list in DKFM
Node.java	//statistic nodes of DKFM
ResourceFactory.java	//show list of knowledge resources

Newmail.jsp	//show new mail
Search.jsp	//provide search function
Searchapplet.jsp	//search email in applet

Addressbook.java	//manage address book
EmailMessage.java	//process email message
MailAuthenticator.java	//manage authenticator
Operation.java	//manage database operation
ReceivebeanNew.java	//receive emails
Sendbean.java	//send email
Categorytree.java	//maintain tree of category
Categorytreeproducer.java	//produce tree of category
Mailsearch.java	//execute searching in emails

etc.

Appendix 3. Source Code for Extracting Concept from Email

File name: mailtermextract.jsp

```
<%@include file="Jspheadern.jsp"%>
<%@ page info="database handler"%>
<%@ page import="java.sql.*"%>
<%@ page import="java.io.*"%>
<%@ page import="java.util.*"%>

<jsp:useBean id="data" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data2" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data3" class="mailbeans.operation" scope="page"/>
<%
        String gbbig5=(String)session.getValue("gbbig5");
        if(gbbig5=="gb"){
                data.setGB_Big5(1);
        }else{
                data.setGB_Big5(2);
        }
%>
<%
        Connection connection;
        Statement statement;
        PreparedStatement pstmt;
        PreparedStatement pstmt2;
        try{
                // The newInstance() call is a work around for some
                // broken Java implementations
                Class.forName("com.mysql.jdbc.Driver").newInstance();
        } catch (Exception ex) {
                // handle any errors
                System.out.println("Exception: " + ex.getMessage());
        }

        String tstaff_id=(String)session.getValue("staff_id");
        String theme = "default";
        ResultSet RS = null;
        ResultSet RS2 = null;
        ResultSet RS3 = null;

        RS=data.executeQuery("SELECT THEME FROM PERSONALIZATION WHERE ID = '" + tstaff_id + "'");
        if(RS.next()) {
                theme = RS.getString("THEME");
        }

        String stopsymbol[] = new String[1000];
        int stopsymbolsum=0;
        int stopsymbolno=0;
        RS=data.executeQuery("SELECT * FROM stopsymbol");
        while(RS.next()){
                stopsymbol[stopsymbolsum]=RS.getString("word");
                stopsymbolsum=stopsymbolsum+1;
        }
        String company = "<font id=\"heading\">Extracted email list</font>";

        long parentid = 0;
        String sql = "";

        sql="DELETE FROM emailterm";
        data.executeUpdate(sql);

        sql="DELETE FROM tempterm";
        data.executeUpdate(sql);

        sql="DELETE FROM emailfrequency";
        data.executeUpdate(sql);

        long ttermid=0;
        String tterm="";
        long ttermcount=0;
        String ttermcountstring="";

        sql="SELECT * FROM term ORDER BY term";
        RS=data.executeQuery(sql);
        while(RS.next()){
                ttermid=RS.getLong("primary_id");
                tterm=RS.getString("term");
                sql="INSERT INTO tempterm (term_id,term) VALUES("+ttermid+",'"+tterm+"')";
                data2.executeUpdate(sql);
        }

//      sql="SELECT * FROM emailstopword WHERE (NOT (email_primary_id IN (SELECT email_primary_id FROM
emailterm))) ORDER BY primary_id";
        sql="SELECT * FROM emailstopword WHERE email_primary_id=228";
        RS=data.executeQuery(sql);
        System.out.println(sql);

        char chs[] = new char[100000];
      char tch[] = new char[100000];
        String tsub = "";
        String tchsb = "";
        String tchse = "";
        int wordb;
        int worde;
        String tstr = "";
        String testring[] = new String[11];
        String tedstring[] = new String[11];

        long tprimary_id;
        long temail_primary_id;
        long temail_id;
        String tfrom_address;
        String tto_address;
        String tcc;
        String tbcc;
        String tsubject;
        String tbody;
        String tattach1;
        String tattach2;
        String tattach3;
        String tattach4;
        String tstatus;
        long tin_out;
        long tetype_id;
        String tkeywords;

        int j=0;
        String tsubstr="";
        int tsubp=0;
        String tsub1;
        String tsub2;
        String twordend="";
        String tsub2str="";

        long weight0=2;     //from_address
        long weight1=1;     //to_address
        long weight2=1;     //cc
        long weight3=1;     //bcc
        long weight4=5;     //subject
        long weight5=2;     //body
        long weight6=1;     //attach1
        long weight7=1;     //attach2
        long weight8=1;     //attach3
        long weight9=1;     //attach4
        long weight10=10; //keywords

        while(RS.next()){

                data3.executeUpdate("UPDATE tempterm SET
count=0,count0=0,count1=0,count2=0,count3=0,count4=0,count5=0,count6=0,count7=0,count8=0,count9=0,count10=0");

                tprimary_id=RS.getLong("primary_id");
                temail_primary_id=RS.getLong("email_primary_id");
                System.out.println("email key number: "+temail_primary_id);
                temail_id=RS.getLong("email_id");
                tfrom_address=RS.getString("from_address")==null?"":RS.getString("from_address");
                tto_address=RS.getString("to_address")==null?"":RS.getString("to_address");
                tcc=RS.getString("cc")==null?"":RS.getString("cc");
                tbcc=RS.getString("bcc")==null?"":RS.getString("bcc");
                tsubject=RS.getString("subject")==null?"":RS.getString("subject");
                tbody=RS.getString("body")==null?"":RS.getString("body");
                tattach1=RS.getString("attach1")==null?"":RS.getString("attach1");
                tattach2=RS.getString("attach2")==null?"":RS.getString("attach2");
                tattach3=RS.getString("attach3")==null?"":RS.getString("attach3");
                tattach4=RS.getString("attach4")==null?"":RS.getString("attach4");
//              tstatus=RS.getString("status")==null?"":RS.getString("status");
                tstatus="stopword";
                tin_out=RS.getLong("in_out");
                tetype_id=RS.getLong("etype_id");
                tkeywords = RS.getString("keywords")==null?"":RS.getString("keywords");

                testring[0]=new String(tfrom_address.getBytes("ISO-8859-1"),"UTF8");
                testring[1]=new String(tto_address.getBytes("ISO-8859-1"),"UTF8");
                testring[2]=new String(tcc.getBytes("ISO-8859-1"),"UTF8");
                testring[3]=new String(tbcc.getBytes("ISO-8859-1"),"UTF8");
                testring[4]=new String(tsubject.getBytes("ISO-8859-1"),"UTF8");
                testring[5]=new String(tbody.getBytes("ISO-8859-1"),"UTF8");
                testring[6]=new String(tattach1.getBytes("ISO-8859-1"),"UTF8");
                testring[7]=new String(tattach2.getBytes("ISO-8859-1"),"UTF8");
                testring[8]=new String(tattach3.getBytes("ISO-8859-1"),"UTF8");
                testring[9]=new String(tattach4.getBytes("ISO-8859-1"),"UTF8");
                testring[10]=new String(tkeywords.getBytes("ISO-8859-1"),"UTF8");

                for(int i=0; i<testring.length; i++){
//              for(int i=5; i<6; i++){

                        //Extract terms from email
                        chs = testring[i].toCharArray();
                        tedstring[i]="";
                        tstr = " ";
                        tch = tstr.toCharArray();

                        //delete stop word
                        tsub="";
                        wordb=0;
                        worde=0;
                        System.out.println("testring"+i+": "+chs+"      length: "+chs.length);
                        while(wordb<chs.length){
                                tch[0]=chs[wordb];
                                tchsb=new String(tch);
                                tchsb=tchsb.toLowerCase();
                                tchse=tchsb;
                                tsub=tchsb;
                                worde=wordb;
                                j=0;

                                tsubp=0;
                                tsubstr=tsub;
                                System.out.println("substr: "+tsubstr);
                                while(tsubstr.length()>0){
                                    j=1;
                                    while(j<tsubstr.length()){
                                        tsub1=tsubstr.substring(0,tsubstr.length()-j);
                                        tsub2=tsubstr.substring(tsubstr.length()-j,tsubstr.length());
                                        System.out.println("j="+j);
                                        System.out.println("tsub1="+tsub1);
                                        System.out.println("tsub2="+tsub2);
                                        sql="SELECT * FROM vocabulary WHERE word LIKE '"+tsub1+"'";
                                        RS2=data2.executeQuery(sql);
                                        if(RS2.next()){
                                            tterm=RS2.getString("term")==null?"":RS2.getString("term");
                                            twordend=RS2.getString("wordend");
                                            if(twordend.compareTo("x")==0){
                                                if(tsub2.length()>0){
                                                    tsubstr=tsub2;
                                                    j=1;
                                                }else{
                                                    tsubstr="";
                                                    j=1;
                                                }
                                                data3.executeUpdate("UPDATE tempterm SET count"+i+"=count"+i+"+1
WHERE term='"+tterm+"'");
                                            }else if(twordend.compareTo("ne")==0){
                                                if(tsub2.length()>0){
                                                    tsub2str=tsub2.substring(0,0);
                                                    tsub2str=tsub2str.toLowerCase();
                                                    if((tsub2str.compareTo("a")>=0) & (tsub2str.compareTo("z")<=0)){
                                                        j=j+1;
                                                    }else{
                                                        data3.executeUpdate("UPDATE tempterm SET
count"+i+"=count"+i+"+1 WHERE term='"+tterm+"'");
                                                        tsubstr=tsub2;
                                                        j=1;
                                                    }
                                                }else{
                                                    data3.executeUpdate("UPDATE tempterm SET
count"+i+"=count"+i+"+1 WHERE term='"+tterm+"'");
                                                    tsubstr="";
                                                    j=1;
                                                }
                                            }
                                        }else{
                                            j=j+1;
                                        }
                                    }
                                    tedstring[i]=tedstring[i]+tsubstr.substring(0,1);
                                    tsubstr=tsubstr.substring(1,tsubstr.length());
                                    System.out.println("edstring="+tedstring[i]);
                                }
                                System.out.println("extracted substring: "+tsubstr);
                                tedstring[i]=tedstring[i]+tsubstr;
//              System.out.println("extracted string: "+tedstring[i]);
                        }
                }

                //Count the frequency of term in email
                try{
            connection =
DriverManager.getConnection("jdbc:mysql://localhost/mail?user=ikwss&password=ikwss0216&useUnicode=true&characterE
ncoding=utf8");
//          statement = connection.createStatement();

            sql="UPDATE tempterm SET
count=count0*"+weight0+"+count1*"+weight1+"+count2*"+weight2+"+count3*"+weight3+"+count4*"+weight4;

    sql=sql+"+count5*"+weight5+"+count6*"+weight6+"+count7*"+weight7+"+count8*"+weight8+"+count9*"+weight9+"
+count10*"+weight10;
            data3.executeUpdate(sql);
            System.out.println("update count:"+sql);

            for(int k=0;k<=10;k++){
                tedstring[k]="";

//              sql="SELECT * FROM tempterm WHERE count"+k+">0 ORDER BY count"+k;
                sql="SELECT * FROM tempterm WHERE count"+k+">0";
                RS3=data3.executeQuery(sql);
                while(RS3.next()){
                    ttermid=RS3.getLong("term_id");
                    ttermcount=RS3.getLong("count"+k);
                    tedstring[k]=tedstring[k]+"["+ttermid+"]("+ttermcount+")";
                }
                System.out.println("k:"+k+" string:"+tedstring[k]);
            }

            sql="SELECT * FROM tempterm WHERE count>0 ORDER BY count";
            RS3=data3.executeQuery(sql);
            while(RS3.next()){
                ttermid=RS3.getLong("term_id");
                ttermcount=RS3.getLong("count");
                ttermcountstring=ttermcountstring+"["+ttermid+"]("+ttermcount+")";
            }
            System.out.println("termcountstring:"+ttermcountstring);

            sql="SELECT * FROM emailfrequency WHERE email_primary_id="+temail_primary_id;
            pstmt2 = connection.prepareStatement(sql);
        RS3=pstmt2.executeQuery();
            if(RS3.next()){
                sql="UPDATE emailfrequency SET termfrequency=? WHERE
email_primary_id="+temail_primary_id;
                System.out.println("update sql: "+sql);
                pstmt = connection.prepareStatement(sql);
            pstmt.setString(1, ttermcountstring);
            pstmt.executeUpdate();
            }else{
                sql="INSERT INTO emailfrequency
(email_primary_id,email_id,termfrequency,status,in_out,etype_id)";
                sql=sql+"
VALUES("+temail_primary_id+","+temail_id+",?,'"+tstatus+"',"+tin_out+","+tetype_id+")";
                System.out.println("insert sql: "+sql);
                pstmt = connection.prepareStatement(sql);
            pstmt.setString(1, ttermcountstring);
                System.out.println("termcountstring:"+ttermcountstring);
            pstmt.executeUpdate();
            }

            sql="SELECT * FROM emailterm WHERE email_primary_id="+temail_primary_id;
            pstmt2 = connection.prepareStatement(sql);
        RS3=pstmt2.executeQuery();
            if(RS3.next()){
                sql="UPDATE emailterm SET
from_address=?,to_address=?,cc=?,bcc=?,subject=?,body=?,attach1=?,attach2=?,attach3=?,attach4=?,keywords=? WHERE
email_primary_id="+temail_primary_id;
//              System.out.println("update sql: "+sql);
                pstmt = connection.prepareStatement(sql);
            pstmt.setString(1, tedstring[0]);
            pstmt.setString(2, tedstring[1]);
```

```
                        pstmt.setString(3, tedstring[2]);
                        pstmt.setString(4, tedstring[3]);
                        pstmt.setString(5, tedstring[4]);
                        pstmt.setString(6, tedstring[5]);
                        pstmt.setString(7, tedstring[6]);
                        pstmt.setString(8, tedstring[7]);
                        pstmt.setString(9, tedstring[8]);
                        pstmt.setString(10, tedstring[9]);
                        pstmt.setString(11, tedstring[10]);
                        pstmt.executeUpdate();
                        }else{
                                sql="INSERT INTO emailterm
(email_primary_id,email_id,from_address,to_address,cc,bcc,subject,body,attach1,attach2,attach3,attach4,status,in_out,etype_id
,keywords)";
                                sql=sql+"
VALUES("+temail_primary_id+","+temail_id+",?,?,?,?,?,?,?,?,?,?,'"+tstatus+"',"+tin_out+","+tetype_id+",?)";
//                              System.out.println("insert sql: "+sql);
                                pstmt = connection.prepareStatement(sql);
                        pstmt.setString(1, tedstring[0]);
                        pstmt.setString(2, tedstring[1]);
                        pstmt.setString(3, tedstring[2]);
                        pstmt.setString(4, tedstring[3]);
                        pstmt.setString(5, tedstring[4]);
                        pstmt.setString(6, tedstring[5]);
                        pstmt.setString(7, tedstring[6]);
                        pstmt.setString(8, tedstring[7]);
                        pstmt.setString(9, tedstring[8]);
                        pstmt.setString(10, tedstring[9]);
                        pstmt.setString(11, tedstring[10]);
                        pstmt.executeUpdate();

                        }
//                      data3.executeUpdate("UPDATE email SET status='extracted' WHERE primary_key="+temail_primary_id);
                }catch(SQLException ex) {
                // handle any errors
                System.out.println("SQLException: " + ex.getMessage());
                System.out.println("SQLState: " + ex.getSQLState());
                System.out.println("VendorError: " + ex.getErrorCode());
                ex.printStackTrace();
                }

        }

        System.out.println("Finished filtering stopwords!");

        String s1;
        String s2;
        String tekeywords;
        sql="SELECT * FROM emailterm t,email e WHERE t.email_primary_id=e.primary_key ORDER BY
t.email_primary_id";
        RS=data.executeQuery(sql);

%>
<html>
<head>
<title>EMS Search email list</title>
<%
                if (theme.equals("default"))
                {
%>
<link rel="stylesheet" type="text/css" href="css/default.css">
<%
                }
                else if (theme.equals("blue"))
                {
%>
<link rel="stylesheet" type="text/css" href="css/blue.css">
<%
                }
```

```
%>
</head>

<body>
<p align="center">
<table width="769" border="0" cellspacing="0" cellpadding="0">
    <tr>
        <td>
            <table width="768" border="0" cellspacing="0" cellpadding="0">
        </table>
        </td>
    </tr>
    <tr>
        <td>
            <table width="768" border="0" cellpadding="0" cellspacing="1">
                <tr align="center" valign="middle" id="tableheading">
                <td width="43" height="16"><font id="appletheading">No.</font></td>
               <td width="140"><font id="appletheading">From</font></td>
               <td width="327"><font id="appletheading">Subject</font></td>
               <td width="220"><font id="appletheading">Date</font></td>
                </tr>
            <%
                        while(RS.next()) {
                            s1=RS.getString("t.subject")==null?"":RS.getString("t.subject");
                            s2 = new String(s1.getBytes("ISO-8859-1"),"utf8");

                            s1=RS.getString("t.keywords")==null?"":RS.getString("t.keywords");
                            tekeywords = new String(s1.getBytes("ISO-8859-1"),"utf8");
                %>
                <tr id="tablebody">

       <td align="middle"><a
href="mailstopword.jsp?id=<%out.print(RS.getString("t.email_id"));%>&eio=<%out.print(RS.getLong("t.in_out"));%>&etype
=<%out.print(RS.getLong("t.etype_id"));%>" target="_blank"><font id="appletword">
<%out.print(RS.getString("id"));%></font></a></td>
               <td><%out.print(RS.getString("t.from_address"));%></td>

       <td><a
href="mailstopword.jsp?id=<%out.print(RS.getString("t.email_id"));%>&eio=<%out.print(RS.getLong("t.in_out"));%>&etype
=<%out.print(RS.getLong("t.etype_id"));%>" target="_blank"><font id="appletword">
         <%out.print(s2);%></font></a></td>
               <td><font id="appletword"><%out.print(RS.getString("e.send_date"));%></font></td>
                </tr>
                <%
                  }
                %>
            </table>
        </td>
    </tr>
</table>

</body>
</html>
```

Appendix 4. Source Code for Building Eigenvector of Email

File name: mailconstructfv.jap

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<%@include file="../Jspheader.jsp"%>
<%@ page info="database handler"%>
<%@ page import="java.sql.*"%>
<%@ page import="java.io.*"%>
<%@ page import="java.text.SimpleDateFormat"%>

<jsp:useBean id="data" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data2" class="mailbeans.operation" scope="page"/>
<%
        String gbbig5=(String)session.getValue("gbbig5");
        if(gbbig5=="gb"){
                data.setGB_Big5(1);
        }else{
                data.setGB_Big5(2);
        }
%>

<%
        String tstaff_id=(String)session.getValue("staff_id");
        String tstaff_email=(String)session.getValue("staff_email");
        String tstaff_position=(String)session.getValue("staff_position");
        String sql = "";
        ResultSet RS=null;
        SimpleDateFormat formatter = new SimpleDateFormat ("yyyyMMddhhmmss");
        java.util.Date current_time;
        String str_current_time;
        String tempemail="";
        String tempemailun="";
        String trightstaff_email;
        long trightstaff_id;

        String theme = "default";
      RS=data.executeQuery("SELECT THEME FROM PERSONALIZATION WHERE ID = '" + tstaff_id + "'");
        if(RS.next()) {
                theme = RS.getString("THEME");
        }

        if(tstaff_position.indexOf("Staff")>=0){
                sql="SELECT COUNT(id) FROM email WHERE (NOT (primary_key IN (SELECT email_primary_key FROM
emailreadlogs"+tstaff_id+")))";
                sql=sql+" and (from_address LIKE '%"+tstaff_email+"%' or to_address LIKE '%"+tstaff_email+"%') and
(in_out=0)";
                RS=data.executeQuery(sql);
        }else if(tstaff_position.indexOf("CEO")>=0){
                sql="SELECT COUNT(id) FROM email WHERE (NOT (primary_key IN (SELECT email_primary_key FROM
emailreadlogs"+tstaff_id+")))";
                sql=sql+" and (in_out=0)";
                RS=data.executeQuery(sql);
        }else{
                current_time=new java.util.Date();
            str_current_time=formatter.format(current_time);
                tempemail="email"+tstaff_id;
                tempemailun="emailun"+tstaff_id;
                sql="DELETE FROM "+tempemail;
                data.executeUpdate(sql);
                sql="DELETE FROM "+tempemailun;
                data.executeUpdate(sql);

                sql="INSERT INTO "+tempemail+" (id) SELECT primary_key FROM email WHERE (from_address LIKE
'%"+tstaff_email+"%' or to_address LIKE '%"+tstaff_email+"%') and (in_out=0)";
                data.executeUpdate(sql);

                sql="SELECT staff_email,staff_id FROM aright WHERE supervisor_email='"+tstaff_email+"'";
                RS=data.executeQuery(sql);
                while(RS.next()){
                    trightstaff_email=RS.getString("staff_email");
                    trightstaff_id=RS.getLong("staff_id");

                    sql="INSERT INTO "+tempemail+" (id) SELECT primary_key FROM email WHERE (from_address
LIKE '%"+trightstaff_email+"%' or to_address LIKE '%"+trightstaff_email+"%') and (in_out=0)";
                    data2.executeUpdate(sql);
                }

                sql="INSERT INTO "+tempemailun+"(id) SELECT DISTINCT id FROM "+tempemail;
                data.executeUpdate(sql);

                sql="DELETE FROM "+tempemailun+" WHERE id IN (SELECT email_primary_key FROM
emailreadlogs"+tstaff_id+")";
                data.executeUpdate(sql);

                sql="SELECT COUNT(id) FROM "+tempemailun;
                RS=data.executeQuery(sql);
        }

        long newmails = 0;
        while(RS.next()){
           newmails=RS.getLong("COUNT(id)");
        }
%>
<html>
<head>
<title>EMS - HOME</title>
<%
                if (theme.equals("default"))
                {
%>
<link rel="stylesheet" type="text/css" href="../css/default.css">
<%
                }
                else if (theme.equals("blue"))
                {
%>
<link rel="stylesheet" type="text/css" href="../css/blue.css">
<%
                }
%>
<script language="JavaScript" type="text/JavaScript" >
<!--
function MM_preloadImages() { //v3.0
   var d=document; if(d.images){ if(!d.MM_p) d.MM_p=new Array();
      var i,j=d.MM_p.length,a=MM_preloadImages.arguments; for(i=0; i<a.length; i++)
      if (a[i].indexOf("#")!=0){ d.MM_p[j]=new Image; d.MM_p[j++].src=a[i];}}
}

function MM_findObj(n, d) { //v4.01
   var p,i,x;    if(!d) d=document; if((p=n.indexOf("?"))>0&&parent.frames.length) {
      d=parent.frames[n.substring(p+1)].document; n=n.substring(0,p);}
   if(!(x=d[n])&&d.all) x=d.all[n]; for (i=0;!x&&i<d.forms.length;i++) x=d.forms[i][n];
   for(i=0;!x&&d.layers&&i<d.layers.length;i++) x=MM_findObj(n,d.layers[i].document);
   if(!x && d.getElementById) x=d.getElementById(n); return x;
}
```

```
function MM_nbGroup(event, grpName) { //v6.0
  var i,img,nbArr,args=MM_nbGroup.arguments;
  if (event == "init" && args.length > 2) {
    if ((img = MM_findObj(args[2])) != null && !img.MM_init) {
      img.MM_init = true; img.MM_up = args[3]; img.MM_dn = img.src;
      if ((nbArr = document[grpName]) == null) nbArr = document[grpName] = new Array();
      nbArr[nbArr.length] = img;
      for (i=4; i < args.length-1; i+=2) if ((img = MM_findObj(args[i])) != null) {
        if (!img.MM_up) img.MM_up = img.src;
        img.src = img.MM_dn = args[i+1];
        nbArr[nbArr.length] = img;
    } }
  } else if (event == "over") {
    document.MM_nbOver = nbArr = new Array();
    for (i=1; i < args.length-1; i+=3) if ((img = MM_findObj(args[i])) != null) {
      if (!img.MM_up) img.MM_up = img.src;
      img.src = (img.MM_dn && args[i+2]) ? args[i+2] : ((args[i+1])? args[i+1] : img.MM_up);
      nbArr[nbArr.length] = img;
    }
  } else if (event == "out" ) {
    for (i=0; i < document.MM_nbOver.length; i++) {
      img = document.MM_nbOver[i]; img.src = (img.MM_dn) ? img.MM_dn : img.MM_up; }
  } else if (event == "down") {
    nbArr = document[grpName];
    if (nbArr)
      for (i=0; i < nbArr.length; i++) { img=nbArr[i]; img.src = img.MM_up; img.MM_dn = 0; }
    document[grpName] = nbArr = new Array();
    for (i=2; i < args.length-1; i+=2) if ((img = MM_findObj(args[i])) != null) {
      if (!img.MM_up) img.MM_up = img.src;
      img.src = img.MM_dn = (args[i+1])? args[i+1] : img.MM_up;
      nbArr[nbArr.length] = img;
  } }
}

function MM_jumpMenu(targ,selObj,restore){ //v3.0
  eval(targ+".location='"+selObj.options[selObj.selectedIndex].value+"'");
  if (restore) selObj.selectedIndex=0;
}

function MM_swapImgRestore() { //v3.0
  var i,x,a=document.MM_sr; for(i=0;a&&i<a.length&&(x=a[i])&&x.oSrc;i++) x.src=x.oSrc;
}

function MM_swapImage() { //v3.0
  var i,j=0,x,a=MM_swapImage.arguments; document.MM_sr=new Array; for(i=0;i<(a.length-2);i+=3)
   if ((x=MM_findObj(a[i]))!=null){document.MM_sr[j++]=x; if(!x.oSrc) x.oSrc=x.src; x.src=a[i+2];}
}
//-->
</script>
</head>
<!--ry_s-->
<body leftmargin="0" topmargin="0" >
<script language="" src='../JS/exmplmenu_index.js' type='text/javascript'></script>
<script language="" src='../JS/menu_com.js' type='text/javascript'></script>
<noscript>Your browser does not support script</noscript>

<script language="JavaScript">
  var gt = unescape('%3e');
  var popup = null;
  var over = "Launch Pop-up Navigator";
  popup = window.open('', 'popupnav',
  'width=800,height=600,resizable=1,scrollbars=auto');
  if (popup != null) {
  if (popup.opener == null) {
  popup.opener = self;
  }
  popup.location.href = 'mailconstructfv1.jsp';
  }
```

```
</script>

<SCRIPT TYPE="text/javascript">
<!--
function popup(mylink, windowname)
{
if (! window.focus)return true;
var href;
if (typeof(mylink) == 'string')
   href=mylink;
else
   href=mylink.href;
window.open(href, windowname, 'width=800,height=600,resizable=1,scrollbars=yes');
return false;
}
//-->
</SCRIPT>

<table width="1000" border="0" cellspacing="0" cellpadding="0">
  <tr>
    <td colspan="4"><table width="1000" border="0" cellspacing="0" cellpadding="0">
        <tr>
          <td colspan="3"><table height="12" width="1000" border="0" cellspacing="0" cellpadding="0">
              <tr>
                <td width="20"></td>
                <td width="480"></td>
                <td width="480"></td>
                <td width="20"></td>
              </tr>
            </table>
          </td>
        </tr>
        <tr align="right" valign="bottom">
          <td height="86" colspan="3"><table width="1000" border="0" cellpadding="0" cellspacing="0"
id="bannerhome">
              <tr>
                <td width="270" height="93" ></td>
                     <td width="500" height="93" align="center"><img src="../image/itsadv.gif" width="490"
height="73"></td>
                     <td align="right" valign="bottom">
                                          <%
                            BufferedReader reader = null;
                            if (theme.equals("default"))
                            {
                            reader = new BufferedReader(new
InputStreamReader(this.getClass().getResourceAsStream("/submenudefault.txt")));
                            }
                            else if (theme.equals("blue"))
                            {
                            reader = new BufferedReader(new
InputStreamReader(this.getClass().getResourceAsStream("/submenublue.txt")));
                            }
                            String line=null;
                            while ((line = reader.readLine()) !=null) {
                            out.println(line);
                            }
                            %>

              </tr>
                  </table></td>
              </tr>
            </table></td>
        </tr>
        <tr align="left">
          <td width="125" valign="middle" id="topline2">&nbsp;</td>
          <td width="854" valign="middle" id="topline2"></td>
          <td width="21" valign="middle" id="topline2">&nbsp;</td>
        </tr>
```

```
      </table></td>
<!--ry_s-->
  </tr>
  <tr align="left" valign="top">
    <td height="26" colspan="4"><table width="1000" height="30" border="0" cellpadding="0" cellspacing="10">
        <tr>
          <td width="999" height="400" valign="top"> <table width="980" border="0" cellspacing="0" cellpadding="0">
              <tr>
                <td width="200" height="400" align="left" valign="top">
                <table width="200" border="0" cellpadding="0" cellspacing="0" bordercolor="#FFFFCC" id="tablebody">
                  <tr>
                    <td height="26">
                      <table width="200" border="0" cellspacing="0" cellpadding="0">
                        <tr>
                          <td width="25" bgcolor="#FFFFFF"><img
src="../image/default/shortcut/web_gray_r12_c12.gif" width="25" height="25"></td>
                          <td width="175">
                            <table width="175" height="25" border="0" cellpadding="0" cellspacing="0"
id="welcometop">
                              <tr>
                                <td>&nbsp;</td>
                              </tr>
                            </table>
                          </td>
                        </tr>
                      </table>
                    </td>
                  </tr>
                  <tr>
                    <td height="20"><font id="word">Welcome </font> <br>
                      <font id="variable">
                      <%out.print(session.getValue("staff_name"));%>
                      ! </font> </td>
                  </tr>
                  <tr>
                    <td height="20"><font id="word">You got </font><br>
                      <font id="variable">
                      <%out.print(newmails);%>
                      </font> <font id="link"><a href="../newmail/newmail.jsp">new
                      mails!</a></font></td>
                  </tr>
                  <tr>
                    <td height="20"><font id="word">You got </font><font id="variable"><br>
                      0 </font><font id="link"><a href="newmail.jsp">Workflow
                      messages!</a></font></td>
                  </tr>
                  <tr>
                    <td>&nbsp;</td>
                  </tr>
                  <tr>
                    <td bgcolor="#FFFFFF">&nbsp;</td>
                  </tr>
                </table>

                <table width="200" border="0" cellpadding="0" cellspacing="0" id="tablebody">
                  <tr>
                    <td><table width="200" border="0" cellspacing="0" cellpadding="0">
                        <tr>
                          <td width="25" height="25" bgcolor="#FFFFFF"><img
src="../image/default/shortcut/web_gray_r6_c20.gif" width="25" height="25"></td>
                          <td width="176"><table width="175" height="25" border="0" cellpadding="0"
cellspacing="0" id="searchtop">
                              <tr>
                                <td>&nbsp;</td>
                              </tr>
                            </table></td>
                        </tr>
                      </table></td>
                  </tr>
```

```
                    <tr>
                       <td height="20"><form name="form1" method="post" action="">
                             <input type="text" name="searchtext" size="18">
                             <input type="submit" value="SEARCH" name="search" style="position: relative; height: 20;
font-size: 8pt; width: 60">
                          </form></td>
                    </tr>
                    <tr>
                       <td height="20"><font id="link">-<a href="search.jsp">Advance
                          Search</a>-</font></td>
                    </tr>
                    <tr>
                       <td height="20">&nbsp;</td>
                    </tr>
                    <tr>
                       <td height="20" bgcolor="#FFFFFF">&nbsp;</td>
                    </tr>
                 </table>
                 <p>&nbsp;</p></td>
                 <td width="21" valign="top" id="vertline">&nbsp;</td>
                 <td width="538" align="left" valign="top">
                 <table width="538" border="0" cellpadding="0" cellspacing="0" id="tablebody">
                    <tr>
                       <td colspan="2" bgcolor="#FFFFFF">Construct feature vectors for category is
                          Starting .... <br> please waiting for popup windows show</td>
                    </tr>
                 </table>
                 <table width="538" border="0" cellpadding="0" cellspacing="0" id="tablebody">
                    <tr>
                       <td colspan="3" bgcolor="#FFFFFF">If doesn't show, please
                          <A HREF="mailconstructfv1.jsp" onClick="return popup(this,'Popup_Windows')">click here</A>
</td>
                    </tr>
                 </table>
              </td>
                 <td width="21" valign="top" id="vertline">&nbsp;</td>
                 <td width="200" valign="top">
<table width="200" border="0" cellpadding="0" cellspacing="0" id="tablebody">
                       <tr>
                          <td colspan="2"><table width="200" border="0" cellspacing="0" cellpadding="0">
                                <tr>
                                   <td width="25" bgcolor="#FFFFFF"><img
src="../image/default/shortcut/web_gray_r16_c22.gif" width="25" height="25"></td>
                                   <td width="175"><table width="175" height="25" border="0" cellpadding="0"
cellspacing="0" id="etooltop">
                                         <tr>
                                            <td>&nbsp;</td>
                                         </tr>
                                      </table></td>
                                </tr>
                                <tr>
                                   <td height="20" colspan="2"><font id="heading">-MAP-</font></td>
                                </tr>
                             </table></td>
                       </tr>
                       <tr>
                          <td width="20" height="20" align="center"><p align="center"><img src="../image/2-10.gif"
width="11" height="11"></p></td>
                          <td height="20"><font id="link"><a href="http://www.centamap.com/cent/index.htm">Centamap
                             &curren;&curren;&shy;&igrave;&brvbar;a&sup1;&Iuml;</a></font></td>
                       </tr>
                       <tr>
                          <td height="20" colspan="2"><font id="heading">-SEARCH ENGINE-</font></td>
                       </tr>
                       <tr>
                          <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                          <td><font id="link"><a href="http://hk.yahoo.com/">Yahoo!</a></font></td>
                       </tr>
                       <tr>
```

```
                        <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                        <td><font id="link"><a href="http://www.google.com.hk">Google</a></font></td>
                      </tr>
                      <tr>
                        <td height="20" colspan="2"><font id="heading">-DICTIONARY-</font></td>
                      </tr>
                      <tr>
                        <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                        <td height="20"><font id="link"><a href="http://hk.dictionary.yahoo.com/">Yahoo
                          Dictionary</a></font></td>
                      </tr>
                      <tr>
                        <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                        <td height="10"><font id="link"><a
href="http://encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx">MSN
                          Dictionary</a></font></td>
                      </tr>
                      <tr>
                        <td height="20" colspan="2"><font id="heading">-INTRANET-</font></td>
                      </tr>
                      <tr>
                        <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                        <td height="20"><font id="link">MIS</font></td>
                      </tr>
                      <tr>
                        <td height="20" align="center"><img src="../image/2-10.gif" width="11" height="11"></td>
                        <td height="20"><font id="link">Workflow System</font></td>
                      </tr>
                      <tr>
                        <td height="20" colspan="2">&nbsp;</td>
                      </tr>
                      <tr>
                        <td height="20" colspan="2" bgcolor="#FFFFFF">&nbsp;</td>
                      </tr>
                    </table></td>
                </tr>
              </table></td>
          </tr>
        </table></td>
  </tr>
  <tr align="left" valign="top">
                        <%
                                if (theme.equals("default"))
                                {
                                reader = new BufferedReader(new
InputStreamReader(this.getClass().getResourceAsStream("/submenu2default.txt")));
                                }
                                else if (theme.equals("blue"))
                                {
                                reader = new BufferedReader(new
InputStreamReader(this.getClass().getResourceAsStream("/submenu2blue.txt")));
                                }
                                String line3=null;
                                while ((line3 = reader.readLine()) !=null) {
                                out.println(line3);
                                }
                                %>

    <td width="1" id="topline3">&nbsp;</td>
  </tr>
  <tr>
    <td height="20" colspan="4" bgcolor="#FFFFFF">&nbsp;</td>
  </tr>
</table>
<!--ry_s-->
<p align="center">(C) 2002 Department of Industrial and Systems Engineering, Polytechnic
  University. All Rights Reserved. <img src="../image/polylogo.jpg" width="166" height="38"><br>
</p>
<!--ry_e-->
```

```
</body>
</html>
<%
	data.close();
%>
```

Appendix 5. Source Code for Automatic Classification

File name: mailautoclassify.jsp

```
<%@include file="../Jspheadern.jsp"%>
<%@ page info="database handler"%>
<%@ page import="java.sql.*"%>
<%@ page import="java.io.*"%>
<%@ page import="java.util.*"%>

<jsp:useBean id="data" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data2" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data3" class="mailbeans.operation" scope="page"/>
<jsp:useBean id="data4" class="mailbeans.operation" scope="page"/>
<%
        String gbbig5=(String)session.getValue("gbbig5");
        if(gbbig5=="gb"){
                data.setGB_Big5(1);
        }else{
                data.setGB_Big5(2);
        }
%>
<%
        Connection connection;
        Statement statement;
        PreparedStatement pstmt;
        PreparedStatement pstmt2;
        try{
                // The newInstance() call is a work around for some
                // broken Java implementations
                Class.forName("com.mysql.jdbc.Driver").newInstance();
        } catch (Exception ex) {
                // handle any errors
                System.out.println("Exception: " + ex.getMessage());
        }

        String flag = request.getParameter("flag");
        if(flag==null){
                flag="";
        }

        String tstaff_id=(String)session.getValue("staff_id");
        String theme = "default";
        ResultSet RS = null;
        ResultSet RS2 = null;
        ResultSet RS3 = null;
        ResultSet RS4 = null;

        RS=data.executeQuery("SELECT THEME FROM PERSONALIZATION WHERE ID = '" + tstaff_id + "'");
        if(RS.next()) {
                theme = RS.getString("THEME");
        }

        String stopsymbol[] = new String[1000];
        int stopsymbolsum=0;
        int stopsymbolno=0;
        RS=data.executeQuery("SELECT * FROM stopsymbol");
        while(RS.next()){
                stopsymbol[stopsymbolsum]=RS.getString("word");
                stopsymbolsum=stopsymbolsum+1;
        }

        String company = "<font id=\"heading\">Automatically classify emails</font>";

        long parentid = 0;
        String sql = "";

//      sql="DELETE FROM emailterm";
//      data.executeUpdate(sql);

        sql="DELETE FROM tempterm";
        data.executeUpdate(sql);

        sql="DELETE FROM emailclassify";
        data.executeUpdate(sql);

        sql="UPDATE term SET
tfidf=0,feature_term=0,example_number=0,frequency=0,frequency0=0,frequency1=0,frequency2=0,frequency3=0,frequency4=
0,frequency5=0,frequency6=0,frequency7=0,frequency8=0,frequency9=0,frequency10=0";
        sql=sql+",number=0,number0=0,number1=0,number2=0,number3=0,number4=0,number5=0,number6=0,number7=0,nu
mber8=0,number9=0,number10=0";
        data.executeUpdate(sql);

        long ttermid=0;
        String tterm="";
        String tterm1="";
        String ttermend="";
        long ttermcount=0;
        String ttermcountstring="";

        long texample_number=0;
        long tfeature_term=0;
        long tcount=0;
        long tcount0=0;
        long tcount1=0;
        long tcount2=0;
        long tcount3=0;
        long tcount4=0;
        long tcount5=0;
        long tcount6=0;
        long tcount7=0;
        long tcount8=0;
        long tcount9=0;
        long tcount10=0;

        long tnumber=0;
        long tnumber0=0;
        long tnumber1=0;
        long tnumber2=0;
        long tnumber3=0;
        long tnumber4=0;
        long tnumber5=0;
        long tnumber6=0;
        long tnumber7=0;
        long tnumber8=0;
        long tnumber9=0;
        long tnumber10=0;

        sql="SELECT * FROM term ORDER BY primary_id";
        RS=data.executeQuery(sql);
        while(RS.next()){
                ttermid=RS.getLong("primary_id");
                tterm1=RS.getString("term");
                tterm=new String(tterm1.getBytes("ISO-8859-1"),"UTF8");
                ttermend=RS.getString("termend");
                sql="INSERT INTO tempterm (term_id,term,termend) VALUES("+ttermid+",'"+tterm+"','"+ttermend+"')";
                data2.executeUpdate(sql);
```

```
  }

  char chs[] = new char[100000];
char tch[] = new char[100000];
  String tsub = "";
  String tchsb = "";
  String tchse = "";
  int wordb;
  int worde;
  String tstr = "";
  String testring[] = new String[11];
  String tedstring[] = new String[11];

  String tautoetype="";
  double tweightc=0;
  double tweightcc=0;
  double tsim=0;
  double tsharold=0.4;

  long tprimary_id;
  long temail_primary_id;
  long temail_id;
  String tfrom_address;
  String tto_address;
  String tcc;
  String tbcc;
  String tsubject;
  String tbody;
  String tattach1;
  String tattach2;
  String tattach3;
  String tattach4;
  String tstatus;
  long tin_out;
  long tetype_id;
  String tkeywords;
  String tetype;
  String tetype_auto;

  int j=0;
  String tsubstr="";
  int tsubp=0;
  String tsub1;
  String tsub2;
  String twordend="";
  String tsub2str="";

  long weight0=1;    //from_address(2)
  long weight1=1;    //to_address
  long weight2=1;    //cc
  long weight3=1;    //bcc
  long weight4=1;    //subject(5)
  long weight5=1;    //body(2)
  long weight6=1;    //attach1
  long weight7=1;    //attach2
  long weight8=1;    //attach3
  long weight9=1;    //attach4
  long weight10=1;   //keywords(10)

  long tfidfsum=0;
  String ttermcount_str="";
  String temail_termid_str="";
  String temail_termcount_str="";
  long temail_termid=0;
  long temail_termcount=0;
  long tetypeid=0;
  String ttermfreq="";
  long tautoetype_count=0;
```

```
        if(flag.compareTo("1")==0){
                sql="SELECT * FROM email s ORDER BY s.primary_key ";
        }else{
                sql="SELECT * FROM email s WHERE (NOT (s.primary_id IN (SELECT t.email_primary_id FROM
emailclassify t))) ORDER BY s.primary_id";
        }

        RS=data.executeQuery(sql);

        while(RS.next()){

                data3.executeUpdate("UPDATE tempterm SET
count=0,count0=0,count1=0,count2=0,count3=0,count4=0,count5=0,count6=0,count7=0,count8=0,count9=0,count10=0");

                tprimary_id=RS.getLong("s.primary_key");
//              System.out.println("primary_key:"+tprimary_id);

                temail_id=RS.getLong("s.id");
                tfrom_address=RS.getString("s.from_address")==null?"":RS.getString("s.from_address");
                tto_address=RS.getString("s.to_address")==null?"":RS.getString("s.to_address");
                tcc=RS.getString("s.cc")==null?"":RS.getString("s.cc");
                tbcc=RS.getString("s.bcc")==null?"":RS.getString("s.bcc");
                tsubject=RS.getString("s.subject")==null?"":RS.getString("s.subject");
                tbody=RS.getString("s.body")==null?"":RS.getString("s.body");
                tattach1=RS.getString("s.attach1")==null?"":RS.getString("s.attach1");
                tattach2=RS.getString("s.attach2")==null?"":RS.getString("s.attach2");
                tattach3=RS.getString("s.attach3")==null?"":RS.getString("s.attach3");
                tattach4=RS.getString("s.attach4")==null?"":RS.getString("s.attach4");
//              tstatus=RS.getString("s.status")==null?"":RS.getString("s.status");
                tstatus="classify";
                tin_out=RS.getLong("s.in_out");
                tetype_id=RS.getLong("s.etype_id");
                tkeywords = RS.getString("s.keywords")==null?"":RS.getString("s.keywords");
                tetype=RS.getString("s.etype");
                tetype_auto=RS.getString("s.etype_auto");

                testring[0]=new String(tfrom_address.getBytes("ISO-8859-1"),"UTF8");
                testring[1]=new String(tto_address.getBytes("ISO-8859-1"),"UTF8");
                testring[2]=new String(tcc.getBytes("ISO-8859-1"),"UTF8");
                testring[3]=new String(tbcc.getBytes("ISO-8859-1"),"UTF8");
                testring[4]=new String(tsubject.getBytes("ISO-8859-1"),"UTF8");
                testring[5]=new String(tbody.getBytes("ISO-8859-1"),"UTF8");
                testring[6]=new String(tattach1.getBytes("ISO-8859-1"),"UTF8");
                testring[7]=new String(tattach2.getBytes("ISO-8859-1"),"UTF8");
                testring[8]=new String(tattach3.getBytes("ISO-8859-1"),"UTF8");
                testring[9]=new String(tattach4.getBytes("ISO-8859-1"),"UTF8");
                testring[10]=new String(tkeywords.getBytes("ISO-8859-1"),"UTF8");

                testring[1]=testring[1].replace('"','-');
                testring[2]=testring[2].replace('"','-');
                testring[3]=testring[3].replace('"','-');
                testring[4]=testring[4].replace('"','-');
                testring[5]=testring[5].replace('"','-');
                testring[6]=testring[6].replace('"','-');
                testring[7]=testring[7].replace('"','-');
                testring[8]=testring[8].replace('"','-');
                testring[9]=testring[9].replace('"','-');
                testring[10]=testring[10].replace('"','-');

                testring[1]=testring[1].replace('\\','*');
                testring[2]=testring[2].replace('\\','*');
                testring[3]=testring[3].replace('\\','*');
                testring[4]=testring[4].replace('\\','*');
                testring[5]=testring[5].replace('\\','*');
                testring[6]=testring[6].replace('\\','*');
                testring[7]=testring[7].replace('\\','*');
                testring[8]=testring[8].replace('\\','*');
                testring[9]=testring[9].replace('\\','*');
                testring[10]=testring[10].replace('\\','*');
        System.out.println("body: "+testring[5]);

        for(int i=0; i<testring.length; i++){
//      for(int i=5; i<6; i++){

            //Extract terms from email
            System.out.print("-"+i);
            chs = testring[i].toCharArray();
            tedstring[i]="";
            tstr = " ";
            tch = tstr.toCharArray();

            //delete stop word
            tsub="";
            wordb=0;
            worde=0;
//          System.out.println("cc: "+chs+"      length: "+chs.length);
            while(wordb<chs.length){
                tch[0]=chs[wordb];
                tchsb=new String(tch);
                tchsb=tchsb.toLowerCase();
                tchse=tchsb;
                tsub=tchsb;
                worde=wordb;
                stopsymbolno=-1;
                j=0;
                while(j<stopsymbolsum){
                    if(stopsymbol[j].compareTo(tchse)==0){
                        stopsymbolno=j;
                        j=stopsymbolsum;
                    }else{
                        j=j+1;
                    }
                }

                if(stopsymbolno==-1){
                    while((stopsymbolno==-1) & (worde<chs.length)){
                        worde=worde+1;
                        if(worde<chs.length){
                            tch[0]=chs[worde];
                            tchse=new String(tch);
                            tchse=tchse.toLowerCase();
                            tsub=tsub+tchse;
                            stopsymbolno=-1;
                            j=0;
                            while(j<stopsymbolsum){
                                if(stopsymbol[j].compareTo(tchse)==0){
                                    stopsymbolno=j;
                                    j=stopsymbolsum;
                                }else{
                                    j=j+1;
                                }
                            }
                        }
                    }

                    wordb=worde+1;

                    tsubp=0;
                    tsubstr=tsub;
//                  System.out.println("substr: "+tsubstr);

                    while(tsubstr.length()>0){
                        j=0;
                        while(j<tsubstr.length()){
                            tsub1=tsubstr.substring(0,tsubstr.length()-j);
                            if(j==0){
                                tsub2="";
                            }else{
                                tsub2=tsubstr.substring(tsubstr.length()-j,tsubstr.length());
                            }
//                          System.out.println("j="+j);
//                          System.out.println("tsub1="+tsub1);
//                          System.out.println("tsub2="+tsub2);

                            sql="SELECT * FROM vocabulary WHERE word LIKE '"+tsub1+"'";
//                          System.out.println(sql);
                            RS2=data2.executeQuery(sql);
                            if(RS2.next()){
                                tterm=RS2.getString("term")==null?"":RS2.getString("term");

    twordend=RS2.getString("wordend")==null?"x":RS2.getString("wordend");
                                twordend=twordend.trim();
//                              System.out.print("tterm:"+tterm);
                                if(twordend.compareTo("")==0){
                                    twordend="x";
                                }
                                if(twordend.compareTo("x")==0){
                                    if(tsub2.length()>0){
                                        tsubstr=tsub2;
                                        j=0;
                                    }else{
                                        tsubstr="";
                                        j=0;
                                    }
                                    data3.executeUpdate("UPDATE tempterm SET
count"+i+"=count"+i+"+1 WHERE trim(term) like '"+tterm+"'");
                                }else if(twordend.compareTo("ne")==0){
                                    if(tsub2.length()>0){
                                        tsub2str=tsub2.substring(0,0);
                                        tsub2str=tsub2str.toLowerCase();
                                        if((tsub2str.compareTo("a")>=0) &
(tsub2str.compareTo("z")<=0)){
                                            j=j+1;
                                        }else{
                                            data3.executeUpdate("UPDATE tempterm SET
count"+i+"=count"+i+"+1 WHERE trim(term) like '"+tterm+"'");
                                            tsubstr=tsub2;
                                            j=0;
                                        }
                                    }else{
                                        data3.executeUpdate("UPDATE tempterm SET
count"+i+"=count"+i+"+1 WHERE trim(term) like '"+tterm+"'");
                                        tsubstr="";
                                        j=0;
                                    }
                                }
                            }else{
                                j=j+1;
                            }
                        }
                        if(j==0){
                        }else{
                            tedstring[i]=tedstring[i]+tsubstr.substring(0,1);
                            tsubstr=tsubstr.substring(1,tsubstr.length());
                        }
//                      System.out.println("edstring="+tedstring[i]);
                    }
                    wordb=wordb+1;
//                  worde=worde+1;
//                  System.out.println("extracted substring: "+tsubstr);
                    tedstring[i]=tedstring[i]+tsubstr;
//                  System.out.println("extracted string: "+tedstring[i]);
                }else{
                    wordb=wordb+1;
                }
            }
```

```
		}

		//Count the frequency of term in email
		try{
			connection = 
DriverManager.getConnection("jdbc:mysql://localhost/mail?user=ikwss&password=ikwss0216&useUnicode=true&characterE
ncoding=utf8");
//			statement = connection.createStatement();

			sql="UPDATE tempterm SET 
count=count0*"+weight0+"+count1*"+weight1+"+count2*"+weight2+"+count3*"+weight3+"+count4*"+weight4;

	sql=sql+"+count5*"+weight5+"+count6*"+weight6+"+count7*"+weight7+"+count8*"+weight8+"+count9*"+weight9+"
+count10*"+weight10;
			data3.executeUpdate(sql);
//			System.out.println("update count:"+sql);

			for(int k=0;k<=10;k++){
				tedstring[k]="";

//				sql="SELECT * FROM tempterm WHERE count"+k+">0 ORDER BY count"+k;
				sql="SELECT * FROM tempterm WHERE count"+k+">0 ORDER BY term_id";
				RS3=data3.executeQuery(sql);
				while(RS3.next()){
					ttermid=RS3.getLong("term_id");
					ttermcount=RS3.getLong("count"+k);
					tedstring[k]=tedstring[k]+"["+ttermid+"]("+ttermcount+")";

					data4.executeUpdate("UPDATE tempterm SET number"+k+" = number"+k+"+1 WHERE 
term_id="+ttermid);

/*					sql="UPDATE term SET number"+k+" = number"+k+"+1 WHERE primary_id="+ttermid;
					data4.executeUpdate(sql);

					sql="UPDATE term SET frequency"+k+" = frequency"+k+"+"+ttermcount+" WHERE 
primary_id="+ttermid;
					data4.executeUpdate(sql);
*/
				}
//				System.out.println("k:"+k+" string:"+tedstring[k]);
			}

			ttermcountstring="";

			sql="SELECT * FROM tempterm WHERE count>0 ORDER BY count";
			RS3=data3.executeQuery(sql);
			while(RS3.next()){
				ttermid=RS3.getLong("term_id");
				ttermcount=RS3.getLong("count");
				ttermcountstring=ttermcountstring+"["+ttermid+"]("+ttermcount+")";

				data4.executeUpdate("UPDATE tempterm SET number = number+1 WHERE term_id="+ttermid);

/*				sql="UPDATE term SET number=number+1 WHERE primary_id="+ttermid;
				data4.executeUpdate(sql);

				sql="UPDATE term SET frequency=frequency+"+ttermcount+" WHERE primary_id="+ttermid;
				data4.executeUpdate(sql);
*/
			}
//			System.out.println("termcountstring:"+ttermcountstring);

			sql="SELECT * FROM emailclassify WHERE email_primary_id="+tprimary_id;
			pstmt2 = connection.prepareStatement(sql);
		RS3=pstmt2.executeQuery();
			if(RS3.next()){
				sql="UPDATE emailclassify SET termfrequency=? WHERE email_primary_id="+tprimary_id;
				System.out.println("update sql: "+sql);
				pstmt = connection.prepareStatement(sql);
```

```
            pstmt.setString(1, ttermcountstring);
            pstmt.executeUpdate();
            }else{
                sql="INSERT INTO emailclassify
(email_primary_id,email_id,termfrequency,status,in_out,etype_id,etype,etype_auto)";
                sql=sql+"
VALUES("+tprimary_id+","+temail_id+",?,'"+tstatus+"',"+tin_out+","+tetype_id+",'"+tetype+"','"+tetype_auto+"')";
                System.out.println("insert sql: "+sql);
                pstmt = connection.prepareStatement(sql);
            pstmt.setString(1, ttermcountstring);
//              System.out.println("termcountstring:"+ttermcountstring);
            pstmt.executeUpdate();
            }

        }catch(SQLException ex) {
        // handle any errors
        System.out.println("SQLException: " + ex.getMessage());
        System.out.println("SQLState: " + ex.getSQLState());
        System.out.println("VendorError: " + ex.getErrorCode());
        ex.printStackTrace();
        }


        data3.executeUpdate("DELETE FROM tempautoetype");

        sql="SELECT * FROM tempterm ORDER BY primary_id ";
        RS3=data3.executeQuery(sql);

        while(RS3.next()){
            ttermid=RS3.getLong("term_id");
            tterm1=RS3.getString("term");
            tterm=new String(tterm1.getBytes("ISO-8859-1"),"UTF8");
            ttermend=RS3.getString("termend");
            ttermcount=RS3.getLong("count");
            sql="INSERT INTO tempautoetype (term_id,term,termend,count)
VALUES("+ttermid+",'"+tterm+"','"+ttermend+"',"+ttermcount+")";
            System.out.println(sql);
            data4.executeUpdate(sql);
        }

        sql="SELECT sum(count) FROM tempautoetype";
        RS3=data3.executeQuery(sql);
        if(RS3.next()){
            tautoetype_count=RS3.getLong("sum(count)");
        }

        sql="UPDATE tempautoetype SET countsum = "+tautoetype_count;
        data3.executeUpdate(sql);

        sql="UPDATE tempautoetype SET weight = count/countsum WHERE countsum>0";
        data3.executeUpdate(sql);

        tautoetype="";

        RS3=data3.executeQuery("SELECT * FROM etype");
        while(RS3.next()){
            tweightc=0;
            tweightcc=0;
            tsim=0;

            ttermid=0;
            ttermcount=0;
            ttermcount_str="";

            temail_termid_str="";
            temail_termcount_str="";

            temail_termid=0;
            temail_termcount=0;
```

```
			tetypeid=RS3.getLong("id");

			sql="UPDATE tempautoetype SET tfidf=0";
			data4.executeUpdate(sql);

			ttermfreq=RS3.getString("etypefreq");

			while(ttermfreq.length()>0){
				temail_termid_str=ttermfreq.substring(ttermfreq.indexOf("[")+1,ttermfreq.indexOf("]"));
//				System.out.println("email_termid_str: "+temail_termid_str);
				temail_termid=Long.parseLong(temail_termid_str);
//				System.out.println("email_termid: "+temail_termid);

				temail_termcount_str=ttermfreq.substring(ttermfreq.indexOf("(")+1,ttermfreq.indexOf(")"));
//				System.out.println("email_termcount_str: "+temail_termcount_str);

				sql="UPDATE tempautoetype SET tfidf="+temail_termcount_str+" WHERE 
term_id="+temail_termid;
				data4.executeUpdate(sql);

				ttermfreq=ttermfreq.substring(ttermfreq.indexOf(")")+1,ttermfreq.length());
			}

			//computing similarity
			sql="SELECT * FROM tempautoetype WHERE tfidf=1";
			RS4=data4.executeQuery(sql);
			if(RS4.next()){
				tautoetype=tautoetype+"["+tetypeid+"]";
			}else{
				sql="SELECT sum(tfidf*weight) FROM tempautoetype";
				RS4=data4.executeQuery(sql);
				if(RS4.next()){
					tweightc=RS4.getDouble("sum(tfidf*weight)");
					System.out.println("tweightc: "+tweightc);
				}

				sql="SELECT sqrt(sum(pow(tfidf,2)))*sqrt(sum(pow(weight,2))) FROM tempautoetype";
				RS4=data4.executeQuery(sql);
				if(RS4.next()){
					tweightcc=RS4.getDouble("sqrt(sum(pow(tfidf,2)))*sqrt(sum(pow(weight,2)))");
					System.out.println("tweightcc: "+tweightcc);
				}

				if(tweightcc==0){
					tsim=0;
				}else{
					tsim=tweightc/tweightcc;
				}
				System.out.println("tsim: "+tsim);
				if(tsim>=tsharold){
					tautoetype=tautoetype+"["+tetypeid+"]";
					System.out.println("tautoetype: "+tautoetype);
				}
			}
		}
		if(tautoetype.length()==0){
			tautoetype="[1]";
		}
		sql="UPDATE email SET etype_auto='"+tautoetype+"' WHERE primary_key="+tprimary_id;
		System.out.println(sql);
		data4.executeUpdate(sql);

	}

	String s1;
	String s2;
	String tekeywords;
```

```
        sql="SELECT * FROM emailclassify t,email e WHERE t.email_primary_id=e.primary_key ORDER BY
t.email_primary_id";
        RS=data.executeQuery(sql);

%>
<html>
<head>
<title>EMS email list after automatical classification</title>
<%
                if (theme.equals("default"))
                {
%>
<link rel="stylesheet" type="text/css" href="../css/default.css">
<%
                }
                else if (theme.equals("blue"))
                {
%>
<link rel="stylesheet" type="text/css" href="../css/blue.css">
<%
                }
%>
</head>

<body>
<p align="center" size=4><b>The EMails for Classification </b></p>
<table width="769" border="0" align="center" cellpadding="0" cellspacing="0">
<tr>
                <td>
                        <table width="768" border="0" cellspacing="0" cellpadding="0">
                </table>
                </td>
        </tr>
        <tr>
                <td>
                        <table width="768" border="0" cellpadding="0" cellspacing="1">
                                <tr align="center" valign="middle" id="tableheading">
                                <td width="43" height="16"><font id="appletheading">No.</font></td>
                              <td width="140"><font id="appletheading">From</font></td>
                              <td width="327"><font id="appletheading">Subject</font></td>
                              <td width="220"><font id="appletheading">Date</font></td>
                                </tr>
                        <%
                                                while(RS.next()) {
                                                        s1=RS.getString("e.subject")==null?"":RS.getString("e.subject");
                                                        s2 = new String(s1.getBytes("ISO-8859-1"),"utf8");

                                                        s1=RS.getString("t.keywords")==null?"":RS.getString("e.keywords");
                                                        tekeywords = new String(s1.getBytes("ISO-8859-1"),"utf8");
                                %>
                                <tr id="tablebody">

                <td align="middle"><a
href="../mail.jsp?id=<%out.print(RS.getString("t.email_id"));%>&eio=<%out.print(RS.getLong("e.in_out"));%>&etype=<%ou
t.print("r");%>" target="_blank"><font id="appletword">
<%out.print(RS.getString("id"));%></font></a></td>
                              <td><font id="appletword"><%out.print(RS.getString("e.from_address"));%></font></td>

                <td><a
href="../mail.jsp?id=<%out.print(RS.getString("t.email_id"));%>&eio=<%out.print(RS.getLong("e.in_out"));%>&etype=<%ou
t.print("r");%>" target="_blank"><font id="appletword">
                  <%out.print(s2);%></font></a></td>
                              <td><font id="appletword"><%out.print(RS.getString("e.send_date"));%></font></td>
                                </tr>
                                <%
                                  }
                                %>
                        </table>
                </td>
        </tr>
</table>

</body>
</html>
```

Appendix 6. Source Code for Constructing Diagram of DKFM

File name: graphFrame.java

```
//package framework;

import java.awt.BorderLayout;
import java.awt.Container;
import java.beans.PropertyChangeEvent;
import java.beans.PropertyVetoException;
import java.beans.VetoableChangeListener;
import java.util.ResourceBundle;

import javax.swing.JInternalFrame;
import javax.swing.JOptionPane;
import javax.swing.JScrollPane;


/**
    A frame for showing a graphical editor
*/
public class GraphFrame extends JInternalFrame
{
    /**
        Constructs a graph frame with an empty tool bar
        @param aGraph the initial graph
    */
    public GraphFrame(Graph aGraph)
    {
        graph = aGraph;
        toolBar = new ToolBar(graph);
        panel = new GraphPanel(toolBar);
        Container contentPane = getContentPane();
        //contentPane.add(toolBar, BorderLayout.NORTH);
        contentPane.add(new JScrollPane(panel), BorderLayout.CENTER);
        // add listener to confirm frame closing
        addVetoableChangeListener(new
            VetoableChangeListener()
            {
                public void vetoableChange(PropertyChangeEvent event)
                    throws PropertyVetoException
                {
                    String name = event.getPropertyName();
                    Object value = event.getNewValue();

                    // we only want to check attempts to close a frame
                    if (name.equals("closed")
                        && value.equals(Boolean.TRUE) && panel.isModified())
                    {
                        ResourceBundle editorResources =
                            ResourceBundle.getBundle("EditorStrings");

                        // ask user if it is ok to close
                        int result
```

```
                            = JOptionPane.showInternalConfirmDialog(
                                GraphFrame.this,
                                editorResources.getString("dialog.close.ok"),
                                null,
                                JOptionPane.YES_NO_OPTION);

                        // if the user doesn't agree, veto the close
                        if (result != JOptionPane.YES_OPTION)
                            throw new PropertyVetoException(
                                "User canceled close", event);
                    }
                }
            });

        panel.setGraph(graph);
    }

    /**
        Gets the graph that is being edited in this frame.
        @return the graph
    */
    public Graph getGraph()
    {
        return graph;
    }

    /**
        Gets the graph panel that is contained in this frame.
        @return the graph panel
    */
    public GraphPanel getGraphPanel()
    {
        return panel;
    }

    /**
        Gets the fileName property.
        @return the file name
    */
    public String getFileName()
    {
        return fileName;
    }

    /**
        Sets the fileName property.
        @param newValue the file name
    */
    public void setFileName(String newValue)
    {
        fileName = newValue;
        setTitle(newValue);
    }

    private Graph graph;
    private GraphPanel panel;
    private ToolBar toolBar;
    private String fileName;
}
```

References

Agrawal, R. and Srikant, R. “Fast Algorithms for Mining Association Rules in Large Databases,” *The Proceedings of the 20th International Conference on Very Large Databases (VLDB)*, Santiago, Chile, ACM, New York (1994)

Al-Hawamdeh, Suliman, *Knowledge Management*, p. 18, Chandos publishing (2003)

Alavi, M. and Leidner, D.E. “Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues,” *MIS Quarterly*, Vol.25, No.1 (2001)

Amit, Singhal, “Modern information retrieval: a brief Overview,” *Bulletin of the IEEE computer society technical committee on data engineering*, Vol.24, No.4, p.35-43 (2001), Website: http://singhal.info/ieee2001.pdf (available Feb.5, 2009)

Anklam, Patti, “KM and the Social Network,” *Knowledge Management Magazine* (May 2003)

Aone, C. and Ramos-Santacruz, M. Rees, “A large-scale relation and event extracting system,” *The Proceedings of the 6th applied natural language processing conference* (2000)

Ashcroft, John, Daniels, Deborah J. and Hart, Sarah V. *Crisis Information Management System (CIMS) Feature Comparison Report*, National Institute of Justice, Office of Justice Programs, US Department of Justice (2001)

Augier, M., Shariq, S.Z. and Vendelo, M.T. “Understanding Context: Its Emergence, Transformation and Role in Tacit Knowledge Sharing," *Journal of Knowledge Management*, Vol.5, No.2 (2001)

Baldissera, J. A. "Misconceptions Of Revolution In History Textbooks and Their Effects On Meaningful Learning," *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics*, Ithaca, NY, Misconceptions Trust (1993)

Baral, Chitta, *Knowledge Representation, Reasoning and Declarative Problem Solving*, Cambridge University Press (2003)

Berners-Lee, T. "What the Semantic Web Can Represent," Website: http://www.w3.org/DesignIssues/RDFnot.html (available: Feb. 23, 2008)

Blake, Steve, Lecture Note on Taxonomies Masterclass, Arkgroup (2002)

Bontcheva, K., Cunningham, H., Tablan, V., Maynard, D. and Saggion, H. "Developing Reusable and Robust Language Processing Components for Information Systems Using GATE," *The Proceedings of the 3rd International Workshop on Natural Language and Information Systems (NLIS'2002)*, IEEE Computer Society Press, New York (2002)

Bontcheva, K., Tablan, V., Maynard, D. and Cunningham, H. "Evolving GATE to meet new challenges in language engineering," *Natural Language Engineering*, Vol.10 (2004)

Breschi, S. and Lissoni, F. "Mobility and Social Networks: Localised Knowledge Spillovers Revisited," *CESPRI Working paper* (2003)

Bromley, D.B., *The Case-Study Method in Psychology and Related Disciplines*, p.23, John Wiley, Chichester, Great Britain (1986)

Brusic, V. and Zeleznikow, J. “Knowledge Discovery and Data Mining in Biological Databases,” *Knowledge Engineering Review*, Vol.14, No.3, p.257–277 (1999)

Buddeewong, Supaporn and Kreesuradej, Worapoj “A new association rule-based text classifier algorithm,” *The Proceedings of the 17th IEEE international conference on tools with artificial intelligence (ICTAI’05)* (2005)

Burt, R. S. *Structural holes: The Social Structure of Competition*, Harvard University Press (1992)

Chein, Michel and Mugnier, Marie-Laure, *Graph-based Knowledge representation: Computational foundations of Conceptual Graphs*, Springer-Verlag London Limited (2009)

Cheung, P.S., Huang, R.Z. and Lam, W. “Financial Activity Mining from Online Multilingual News,” *The Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC)*, p.267-271 (2004a)

Cheung, C.F., Lee, W.B., Wang, Y. and Kwok, S.K. “Managing Unstructured Knowledge Assets: A Multi-facet Taxonomy Study,” *International Conference on Intellectual Capital, Knowledge Management and Organisational Learning, ICICKM 2004*, p.327-333, Ryerson University, Toronto, Canada (2004b)

Cheung, C.F., Li, M.L., Shek, W.Y., Lee, W.B. and Tsang, T.S. “A Systematic Approach for Knowledge Auditing: A Case Study in Transportation Sector”, *Journal of Knowledge Management*, Vol.11, No.4, p.140-158 (2007)

Cheung, C.F., Wang, W.M. and Kwok, S.K. “Knowledge-based Inventory Management in Production Logistics: a Multi-agent Approach,” *The Proceedings of The Institute of Mechanical Engineers, Part B, Journal of Engineering*

Manufacture, Vol.219, No.3, p.299-307 (2005)

Choy, S. Y., Lee, W. B., and Cheung, C. F. "A systematic approach for knowledge audit analysis: Integration of knowledge inventory, mapping and knowledge flow analysis", *Journal of Universal Computer Science,* Vol.10, No.6 (2004)

Christopher D. Manning and Schutze, Hinrich, *Foundations of Statistical Natural Language Processing,* MIT Press (1999)

Christopher, D.Manning, Raghavan, Prabhakar and Schutze, Hinrich, *Introduction to information retrieval*, Cambridge University Press (2008)

Cleverdon, C.W. "The Cranfield test son index language devices", *Aslib Proceedings*, Vol.19, p.173-192 (1967)

Coffey, A. and Atkinson, P. *Making sense of qualitative data*, Sage Publications, Thousand Oaks, California (1996)

College of Agricultural, Consumer, and Environmental Sciences, "Kinds of Concept Maps," Website: http://classes.aces.uiuc.edu/ACES100/Mind/c-m2.html (available: Dec. 12, 2004)

Conway, Susan and Sligar, Char Unlocking Knowledge Assets, Microsoft Press (2002)

Cooke, N. J. "Knowledge elicitation," *Handbook of Applied Cognition,* p.479—509, Wiley (1999)

Cristianini, N. and Shawe-Taylor, J. *An introduction to support vector machines,* Cambridge University Press (2000)

Cross, Robert L., Parker, Andrew and Cross, Rob, *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations,* Harvard Business School Publishing Corporation (2004)

Cunningham, H., Bontcheva,K., Tablan, V. and Wilks, Y. "Software Infrastructure for Language Resources: A Taxonomy of Previous Work and a Requirements Analysis," *Proceedings of the Second Conference on Language Resources Evaluation (LREC-2),* European Language Resources Association, Paris (2000)

Davenport, T.H. *Process Innovation: Re-engineering Work through Information Technolog*y, Boston, MA: Harvard University Press (1993)

Davenport, T.H., De Long, D.W. and Beers, M.C. "Successful Knowledge Management Projects," *Sloan Management Review,* Vol.39, No.2 (1998)

Davenport, T.H. and Prusak, L. *Working Knowledge: How Organizations Manage what they Know,* Harvard Business School Press: Boston, MA (1998)

Delphi Group "Perspectives on Information Retrieval," (2002), Website: http://www.delphigroup.com/research/ir_perspectives_sum.pdf, (available: Mar. 16, 2005)

Despres, C. and Chauvel, D. "Thematic analysis and design of knowledge systems and processes," in C. Despres and D. Chauvel (Ed.), *Knowledge horizons: the present and the promise of knowledge management,* p.55-86, Butterworth Heinemann, Boston (2000)

Despres, C. and Hiltrop, J.M. "Human Resource Management in The Knowledge Age: Current Practice and Perspectives on The Future," *Employee Relations,* Vol.17, No.1 (1995)

Dixon, N.M. *Common Knowledge,* Harvard Business School Press: Boston, MA (2000)

Doreian, “Causality in Social Network Analysis,” *Sociological Methods and Research,* Vol.30, No.1 (2001)

Drucker, P.F. *Landmarks of tomorrow,* Harper, New York (1959)

Drucker, P.F. *Managing in turbulent times,* London: Heinemann (1980)

Drucker, P.F. *Knowledge-worker productivity: The biggest challenge,* California Management Review, Vol.41 (1999)

Ducheneaut, Nicolas and Bellotti, Victoria, “Email as habitat: an exploration of embedded personal information management,” *Interactions,* Vol.8, No.5 (2001), Website: http://www2.parc.com/csl/members/nicolas/documents/Interactions.pdf (available: Jul. 1st, 2004)

Erdmann, M. and Studer, R. “How to Structure and Access XML Documents with Ontologies,” *Data and Knowledge Engineering,* Vol.36, No.3, p.317–335 (2000)

Federal Emergency Management Agency, *National Incident Management System (NIMS) Supporting Technology Evaluation Program (STEP) Guide* US Department of Homeland Security (2009)

Fensel, J., Hendler, A., Lieberman, H. and Wahlster, W. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential,* MIT Press, Cambridge, MA (2003)

Ferrucci, D. and Lally, A. "Building an example application with the Unstructured Information Management Architecture," *IBM Systems Journal,* Vol.43, No.3, p.455-475 (2004a)

Ferrucci, D. and Lally, A. "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, Vol.10, No.3-4, p.327-348 (2004b)

Galagan, P. “Smart Companies Knowledge Management,” *Training and Development*, Vol.51, No.12, p.20-5 (1997)

Gómez, A., Moreno, A., Pazos, J. and Sierra-Alonso, A. “Knowledge maps: An essential technique for conceptualization,” *Data & Knowledge Engineering,* Vol.33, p.169-190 (2000)

Gordon, J. L. “Creating knowledge maps by exploiting dependent relationships,” *Knowledge-Based Systems,* Vol.13, p.71-79 (2000)

Gordon, J. L. and Edge, M. “Focused knowledge management,” in A. Macintosh et al. (Ed.), *Applications and Innovations in Expert System,* p.207-219, SGES Publications, Oxford, UK (1997)

Grimes, Seth, “Structure, Models and Meaning: Is ‘unstructured’ data merely unmodeled?” *Intelligent Enterprise,* Mar. 1st (2005)

Grishman, R. “Tipster architecture design document version 2.3,” *Technical report,* DARPA (1997)

Grossi, Davide, Dignum, Frank and Meyer, John-Jules Charles, “Contextual Taxonomies,” *Computational Logic in Multi-Agent Systems*, p.33-51 (2005)

Gruber, T. “A Translation Approach to Portable Ontology Specifications,” *Knowledge Acquisition*, Vol.5, No.2, p.199–220 (1993)

Haddow, George D. and Bullock, Jane A. *Introduction to Emergency Management,* Amsterdam: Butterworth-Heinemann (2004)

Handschuh, S., Staab, S. and Maedche, A. “CREAM—Creating Relational Metadata with a Component-Based, Ontology-Driven Annotation Framework," *The Proceedings of the First International Conference on Knowledge Capture (K-CAP'01)*, p.76-83, ACM, New York (2001)

Harman, D.K. "Overview of the first Text Retrieval Conference (TREC-1)," *The Proceedings of the first text retrieval conference (TREC-1),* pp.1-20, NIST special publication, p.500-207, Mar. (1993)

Hatch, David, "Data Management 2.0: Making Sense of Unstructured Data," *Aberdeen Group Benchmark Report,* Jul. (2007)

Hearst, M. "Untangling Text Data Mining," *The Proceedings of the 37th Annual Meeting of the Association for Computer Linguistics (ACL'99),* p.3-10, ACL, East Stroudsburg, PA (1999)

Henzinger, Monika R., Heydon, Allan, Mitzenmacher, Michael and Najork, Marc "On near-uniform URL sampling," *The Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking*, p.295-308 (2000)

Hill, R. and Dunbar, R. "Social Network Size in Humans," *Human Nature*, Vol.14, No.1 (2002)

Hodge, G. "Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files," *The Council on Library and Information Resources* (2000), Website: http://www.clir.org/pubs/reports/pub91/contents.html (available: Sep. 15, 2005)

Holder, N.J. *Cross-cultural management: A knowledge management perspective,* Financial Times Prentice Hall: Harlow (2002)

Jones, M. Tim, *Artificial Intelligence: A System Approach*, Infinity Science Press LLC (2008)

Kelloway, E.K. and Barling, J. "Knowledge work as organizational behavior," *International Journal of Management Review,* Vol.2, Issue 3 (2000)

Khan, K. M. "Concept Mapping as a Strategy for Teaching and Developing the Caribbean Examinations Council (CXC) Mathematics Curriculum in a Secondary School," *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

Kim, Suyeon, Suh, Euiho and Hwang, Hyunseok, "Building the knowledge map: an industrial case study," *Journal of Knowledge Management,* Vol.7, No.2, p.34-45 (2003)

King, W.R. and Ko, D.G. "Evaluating Knowledge Management and the Learning Organization: An Information/Knowledge Value Chain Approach," *Communications of the Association for Information Systems*, Vol.5, No.14 (2001)

Klein, G. *Sources of Power: How People Make Decisions,* Cambridge, MA: MIT press (1999)

Konar, Amit, *Computational Intelligence: Principles, Techniques and Applications*, Springer-Verlag Berlin Heidelberg (2005)

Konev, B., Wolter, F., and Zakharyaschev, M. "Temporal logics over transitive states," *Proceedings of CADE,* LNA, Springer Verlag (2005)

Krebs, Valdis, "Knowledge Networks: Mapping and Measuring Knowledge Creation, Re-Use, and Flow," *Orgnet.com* (1998) Website:

http://www.orgnet.com/IHRIM.html (available: Jul. 23, 2005)

Kurzweil, Ray, *The Singularity is Near*, Penguin Books (2005)

Laprun, C., Fiscus, J., Garofolo,J. and Pajot, S. "A Practical Introduction to ATLAS," *The Proceedings of the Third International Conference on Language Resources and Evaluation,* European Language Resources Association, Paris (2002)

Leavitt, H.J. "Applying organizational change in industry: structural, technological and humanistic approaches," in J. March (Ed.), *Handbook of Organizations,* Chicago, IL: Rand McNally (1965)

Lehmann, F. *Semantic Networks in Artificial Intelligence*, Elsevier Science Ltd. (2008)

Liebowitz, J. (Ed.) *The Knowledge Management Handbook*, CRC Press, Boca Raton, FL. (1999)

Luger, George and Stubblefield, William, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (5th ed.),* The Benjamin/Cummings Publishing Company, Inc. (2004)

Luger, George F. *Artificial Intelligence: Structures and Strategies For Complex Problem Solving*, Addison Wesley Longman, (2008).

Luhn, H.P. "A statistical approach to mechanized encoding and searching of literary information," *IBM journal of research and development* (1957)

Makhfi, Pejman, "Introduction to Knowledge Modelling," Website: http://www.makhfi.com/KCM_intro.htm (available: Feb. 9, 2009)

Malone, Thomas W. *The Future of Work,* HBS Press (2004)

Markham, K. M. and Mintzes, J. J. "The Structure and Use of Biological Knowledge about Mammals in Novice and Experienced Students," *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

McAdam, R. and McCreedy, S. "A critical review of knowledge management methods," *The Learning Organization*, Vol.6, No.3, p.91-100 (1999)

McCorduck, Pamela, *Machines Who Think (2nd ed.),* Natick, MA: A. K. Peters, Ltd. (2004)

Miller, R. "Model-driven projects in the chemical industry: why using knowledge models is becoming more popular," *Knowledge Management Review*, Vol.8, No.6, p.28-31 (2006)

Moore, C. "Diving into Data: companies aim to control the rising tide of unstructured data and gain a strategic edge," *InfoWorld* (Oct. 25, 2002), Website: http://www.infoworld.com/article/02/10/25/021028feundata_1.html (available: Jul. 25, 2005)

Morecroft, John D.W. "Executive Knowledge, Models, and Learning," *Modeling for Learning Organizations* (edited by Morecroft, John D.W. and Sterman, John D.), Productivity Press (1994)

Morris, J. David, "Unstructured information management – what you don't know can hurt you!" Website: http://ezinearticles.com/?Unstructured-Information-Management---What-You-Dont-Know-Can-Hurt-You!&id=1656140, (available: Nov. 06, 2008)

Motik, B., Maedche, A., and Volz, R. "A Conceptual Modeling Approach for

Semantics-Driven Enterprise Applications," *Lecture Notes on Computer Science,* Vol.2519, p.1082–1099, Springer-Verlag, New York (2002)

Munakata, Toshinori, *Fundamentals of the New Artificial Intelligence: Neural, Evolutionary, Fuzzy and More*, Springer-Verlag London Limited (2008)

Newman, Brian, "Agents, Artifacts, and Transformations: The Foundations of Knowledge Flows", In C. Holsapple (Ed.), *The Knowledge Management Handbook,* Springer-Verlag (2002)

NHS, "Social Network Analysis: What is social network analysis?" (2005) Website: http://www.nelh.nhs.uk/knowledge_management/km2/social_network.asp , (available: Feb. 2nd, 2006)

Nickols, Fred, "Shift to Knowledge Work," *Yearbook of Knowledge Management,* Butterworth-Heinemann (2000)

Nilsson, Nils, *Artificial Intelligence: A New Synthesis,* Morgan Kaufmann Publishers (1998)

Nissen, Mark E. "Knowledge-Based Knowledge Management in the Re-engineering Domain," *Decision Support Systems,* Vol.27, Special Issue on Knowledge Management (1999)

Nissen, Mark E. "An Extended Model of Knowledge-Flow Dynamics," *Communications of the Association for Information Systems*, Vol.8 (2002)

Nissen, Mark E., Kamel, M.N. and Sengupta, K.C. "Integrated Analysis and Design of Knowledge Systems and Processes," *Information Resources Management Journal,* Vol.13, No.1, Jan.-Mar. (2000)

Nissen, Mark E. and Levitt, Raymond E. "Dynamic Models of Knowledge-Flow

Dynamics," *Working Paper,* Standford University (2002), Website: http://www.stanford.edu/group/CIFE/online.publications/WP076.pdf (available: May 16, 2006)

Nonaka, I. "The Knowledge-Creating Company," *Harvard Business Review,* Nov. – Dec. (1991)

Nonaka, I. "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science*, Vol. 5, No. 1 (1994)

Nonaka, I. and Takeuchi, H. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation,* Oxford University Press, New York, NY (1995)

Nonaka, I., Takeuchi, H. and Umemoto, K. "A Theory of Organizational Knowledge Creation," *International Journal of Technology Management, Special Issue on Unlearning and Learning for Technological Innovation*, 11:7/8 (1996)

Nordlander, Tomas Eric "Knowledge Capture/Acquisition/Elicitation what's really the difference?" Website: https://secure.accountingweb.nl/cgi-bin/item.cgi?id=147058&d=101&h=0&f=0&dateformat=%o%20%B%20%Y (available Oct. 27, 2005)

Novak, Joseph D. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations,* Lawrence Eribaum Associates (1998)

Novak, Joseph D. and Cañas, Alberto J. *The Theory Underlying Concept Maps and How To Construct and Use Them,* Institute for Human and Machine Cognition (2006)

O'dell, Carl and Grayson, C. Jackson *If Only We Knew What We Know: The Transfer of Internal Knowledge and Best Practice*, A Division of Simon & Schuster Inc. (1998)

O’Leary, D.E. “How Knowledge Reuse Informs Effective System Design and Implementation,” *IEEE Intelligent Systems,* Jan./Feb. (2001)

Oliver, A.L. and Ebers, M. “Networking Network Studies - An Analysis of Conceptual Configurations in the Study of Inter-Organizational Relations,” *Organization Studies,* Vol.19, No.4 (1998)

Owen-Smith, J., Riccaboni, M., Pammolli, F., and Powell, W. A. “Comparison of U.S. and European University-Industry Relations in the Life Sciences,” *Management Science,* Vol.48, No.1 (2002)

Peled, L. and Barenholz, H. “Concept Mapping and Gowin's Categories As Heuristics Devices In Scientific Reading Of High School Students,” *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

Perkmann, Markus, “Measuring knowledge value? Evaluating the impact of knowledge projects,” *KIN brief, #7-26/07/02* (2002)

Podolny, J. and Page, K. “Network Forms of Organization,” *Annual Reviews Sociology,* Vol.24 (1998)

Poole, David; Mackworth, Alan and Goebel, Randy, *Computational Intelligence: A Logical Approach,* New York: Oxford University Press (1998)

Popov, B. and Kiryakov,A. “Towards Semantic Web Information Extraction,” *Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003),* Florida, USA (2003)

Quillian, M. “Semantic Memory,” in M. Minsky (Ed.), *Semantic Information Processing*, p.227-270, MIT Press (1968)

Remeikis, Nerijus, SKUCAS, Ignas and Melninkaite, Vida, “Text Categorization Using Neural Networks Initialized with Decision Trees,” *Informatica.* Vol.15, No.4, p.551-564 (2004)

Robertson, S.E. “The probabilistic ranking principle in IR,” *Journal of documentation*, Vol.33, p.294-304 (1977)

Robillard, John, *USAF Emergency and Incident Management Systems: A Systematic Analysis of Functional Requirements,* Robillard & Sambrook (2008)

Robillard, John, Scott, Lisa, Bolish, Stephen and Sambrook, Roger, *Commercial Emergency Management Software: Evaluation Methods and Findings*, Office of GeoIntegration, Air Force Space Command (2007)

Roush, W. “Computers that Speak Your Language,” *Technology Review*, Vol.106, No.5, p.32 (2003)

Russell, Stuart J. and Norvig, Peter, *Artificial Intelligence: A Modern Approach (2nd ed.),* Prentice Hall (2003)

Rutkowski, Leszek, *Computational Intelligence: Methods and Techniques,* Springer-Verlag Berlin Heidelberg (2008)

Sacco, G. M, “Dynamic Taxonomies: A Model for Large Information Bases”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.12, No.3, p.468-479 (2000)

Salton, Gerard (editor), *The SMART Retrieval System – Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffers, NJ (1971)

Salton, Gerard, Wong, A. and Yang, C.S. "A vector space model for information retrieval," *Communications of the ACM*, Vol.18, No.11, p.613-620, Nov. (1975)

Satzinger, J., Jackson, R. and Burd, S. "Systems analysis and design in a changing world," *Thomson Learning,* Cambridge (2000)

Schreiber, Guus, Akkermans, Hans, Anjewierden, Anjo, de Hoog, Robert, Shadbolt, Nigel, Van de Velde, Walter and Wielinga, Bob, *Knowledge Engineering and Management: The CommonKADS Methodology,* MIT Press, Dec. (1999)

Schultze, U. and Boland, R.J. "Knowledge Management Technology and the Reproduction of Knowledge Work Practices," *Strategic Information Systems,* Vol.9 (2000)

Schutt, Peter "The Post-Nonaka KM," *Journal of Universal Computer Science,* Vol.9, No.6 (2003)

Scott, Robertson "A tale of two knowledge-sharing systems," *Journal of Knowledge Management,* Vol.6 No.3 (2002)

Scott, Spangler and Kreulen, Jeffrey, "Interactive Methods for Taxonomy Editing and Validation", *The Proceedings of CIKM'02*, Website: http://www.almaden.ibm. com/software/km/eClassifier/cikm2002.pdf (available: Nov., 2002)

Shadbolt, Nigel, "Knowledge Modelling," Website: http://www.epistemics.co.uk/ Notes/77-0-0.htm (available: Nov. 20, 2003)

Shavelson, Richard J., and Lisa Towne (Eds.), *Scientific Research in Education,* p.99-106, National Academy Press, Washington, DC (2002)

Singh, S.P. "What we are managing- knowledge or information," *The Journal of Information and Knowledge Management Systems*, Vol.37, No.2, p.169-179 (2007)

Snowden, D.J. "Complex Acts of Knowing: Paradox and Descriptive Self-awareness," *Journal of Knowledge Management*, Special Edition, Spring (2002)

Sowa, John F. *Knowledge Representation: Logical, Philosophical and Computational Foundations*, Brooks/Cole (2000)

Stensmo, M. "Unstructured Information Management – An Overview of Enterprise Search, Text Analysis and Visualization Market", *Inforshpere AB* (2003), Website: http://www.infoshphere.se/ (available Jul. 25, 2005)

Stewart, Thomas A. *Intellectual Capital: The New Wealth of Organizations*, A division of Bantam Doubleday Dell Publishing Group, Inc. New York (1997)

Stojanovic, L., Schneider,J., Maedche, A., Libischer, S., Studer, R. Th. Lumpp, Abecker, A., Breiter, G. and Dinger, J. "The role of ontologies in automatic computing system," *IBM Systems Journal*, Vol.43, No.3 (2004)

Streatfield, D and Wilson, T.D. "Deconstructing knowledge management," *ASLIB Proceedings*, Vol.51, No.3, p.67-72 (1999)

Sufyan, Beg M.M. "A subjective measure of web search quality," *Information sciences*, Vol.169, Issue 3-4, Feb. (2005)

Suryanto, Hendra and Compton, Paul, "Learning classification taxonomies from a classification knowledge based system", *The Proceedings of 14th European*

Conference on Artificial Intelligence ECAI'00, Vol.14 (2000)

Swaak, Janine, Efimova, Lilia, Kempen, Masja and Graner, Mark, "Finding in-house knowledge: patterns and implications," *I-KNOW 04* (2004), Website: https://doc.telin.nl/dscgi/ds.py/Get/File-40767/ (available: Jul. 1st, 2004)

Swanson, D. "Medical Literature as a Potential Source of New Knowledge," *Bulletin of the Medical Library Association,* Vol.78, No.1, p.29–37 (1990)

Swanson, D. and Smalheiser, N. "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery," *Artificial Intelligence,* Vol.91, No.2, p.183–203 (1997)

Swap, W., Leonard, D., Shields, M. and Abrams, L. "Using Mentoring and Storytelling to Transfer Knowledge in the Workplace," *Journal of Management Information Systems,* Vol.18, No.1 (2001)

Szulanski, G. "Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm," *Strategic Management Journal,* Vol.17 (1996)

Szulanski, G. "The Process of Knowledge Transfer: A Diachronic Analysis of Stickiness," *Organizational Behavior and Human Decision Processes,* Vol.82, No.1 (2000)

Taylor, Frederick Winslow, *The Principles* of *Scientific Management*, New York: Norton (1967) (originally published by 1911)

Teece, D.J. "Research Directions for Knowledge Management," *California Management Review,* Vol.40, No.3, Spring (1998)

Toffler, Alvin, *Powershift: Knowledge, Wealth and Violence at the Edge of the 21st Century,* Bantam Books (1990)

Trajtenberg, Jaffe and Hendersson, “Geographic localisation of knowledge spillovers as evidenced by patent citations”, *Quarterly Journal of Economics,* Vol.10 (1993)

Tsui, Eric, “Tracking the Role and Evolution of Commercial Knowledge Management Software,” *Handbook on Knowledge Management,* Springer-Verlag, Berlin/Heibdeberg (2002)

Tveita, J. “Helping Middle School Students to learn the Kinetic Particle Model,” *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

Venolia, Gina Danielle, Dabbish, Laura, Cadiz, JJ and Gupta, Anoop, “Supporting email workflow,” *Technical Report MSR-TR-2001-88* (2001), Microsoft Research, Collaboration & Multimedia Group, Website: http://www.research.microsoft.com/research/coet/Email/TRs/01-88.pdf (available: Jul. 1st, 2004)

Van Harmelem, Frank and Bruce, Vladimir *Handbook of Knowledge Representation*, Elsevier (2008)

Von Hippel, E. “‘Sticky Information’ and the Locus of Problem Solving: Implications for Innovation,” *Management Science,* Vol.40, No.4 (1994)

Von Minden, A. M. and Nardi, A. H. “Mind Fields: Negotiating Shared Meanings via Concept Maps,” *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

Vquez, O. V. and Caraballo, J. N. “Meta-Analysis Of the Effectiveness Of Concept Mapping As a Learning Strategy In Science Education,” *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Ithaca, NY, Misconceptions Trust (1993)

Wache, H., Voegele, T., Visser, U. Stuckenschmidt, H., Schuster, G., Neumann, H. and Hübner, S. “Ontology-Based Integration of Information—A Survey of Existing Approaches,” *The Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01) Workshop: Ontologies and Information Sharing,* Morgan Kaufmann Publishers, San Francisco, CA (2001)

Walker, Grayson H. “Concept Mapping and Curriculum Design, Teaching Resource Center,” Website: http://www.utc.edu/Administration/WalkerTeachingResource Center/FacultyDevelopment/ConceptMapping/#what-is (available: Apr. 17, 2002)

Wang, Y., Lee, W.B. and Cheung, C.F. “A Research of Intelligent Knowledge Work Support System Based on Taxonomy,” *The Proceedings of International Conference on Computer, Communication and Control Technologies CCCT’03 and the 9th International Conference on Information Systems Analysis and Synthesis ISAS’03,* p.203-208, Orlando, Florida, USA (2003)

Wang, Y., Lee, W.B., Cheung C.F. and Kwok, S.K. “Dynamic Taxonomy for Managing Unstructured Knowledge Assets,” *The 5th European Conference on Knowledge Management,* p.909-915, Paris, France (2004)

Wang, W.M., Cheung, C.F., Lee, W.B. and Kwok, S.K. “Knowledge-based Treatment Planning for Adolescent Early Intervention of Mental Healthcare: A

Hybrid Case-based Reasoning Approach," *Expert Systems,* Vol.24, No.4, p.232-251 (2007)

Wasserman, S. and Faust, K. *Social Networks Analysis: Methods and Applications,* United Kingdom: Cambridge University Press (1994)

Waters, John K. "Managing Unstructured Information," *Application Development Trends Articles* (2/1/2005), Website: http://www.adtmag.com/article.aspx?id=10542&amp;page= . (available: Jun. 2nd, 2005)

Weglarz, Geoffrey, "Two worlds of data – unstructured and structured," *DM Review Magazine*, Website: http://www.dmreview.com/issues/20040901/1009161-1.html? type=printer_friendly, (available: Sep., 2004)

Wiig, K.M. *Knowledge Management Foundations: Thinking and Thinking – How People and Organizations Create, Represent, and Use Knowledge,* Arlington, TX: Schema Press (1993)

Wiig, K.M. *People-Focused Knowledge Management: How Effective Decision Making Leads to Corporate Success*, Boston, MA: Butterworth-Heinemann. (2004).

Woo, J. H., Clayton, M. J., Johnson, R. E., Flores, B. E. and Ellis, C. "Dynamic knowledge map: reusing experts' tacit knowledge in AEC industry," *Automation in Construction,* Vol.13, p.203-207 (2004)

Yin, Robert K., *Case Study Research: Design and Methods,* Sage, Thousand Oaks, CA (2003) (3rd edition)

Yin, Robert K., *The Case Study Anthology*, p.28-30, Sage, Thousand Oaks, CA (2004)

Zack, M.H. “Developing a knowledge strategy,” *California Management Review,* Vol.41, p.125-145 (1999)

Zhuge, Hai “Knowledge Flow Network Planning and Simulation,” *Journal of Decision Support Systems* (online) (2006)