DISCOVERING ASSOCIATIONS IN HETEROGENEOUS SOCIAL NETWORKS

LI HO LEUNG

M.Phil

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Computing

Discovering Associations in Heterogeneous Social Networks

Li Ho Leung

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Philosophy

March, 2014

Certificate of Originality

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

_____ (Signed)

Li Ho Leung (Name of student)

Abstract

The studying of association between behavioral patterns in online social network and influences outside the network is an emerging topic in social network analysis. Our work attempted to study whether it is possible to transfer the analysis model from popular social network sites to those sites which are less popular. If such transference is feasible, the existing social network analyses can be efficiently spread to numerous small and medium social networks in the world.

A set of abstract data models, which are named as Generic Networks Data Models (GNDMs), and one abstract association model, which is named as Generic Network Data Association (GNDA), are proposed in our work. The GNDMs conceptually solve the differences in contents among the heterogeneous social networks, while the GNDA defines an abstract model on the associations between online and offline social networks. The GNDA has two components. The first one is borrowed from the studies of popular online social network, while the second one is defined according to the features of the online social network. Therefore, our work is called "transplantation of association analysis" because the association analysis is contributed from one online social network and is used in other networks or applications. A service-based analytical framework (the D-Miner Service Framework) is proposed to implement the GNDA by integrating all relevant solutions proposed in our work. The framework uses a novel linking technique, which is called Generic Network Data Linking (GNDL), to connect the data in a form of the GNDMs. Networks Content Linkage (NWCL), which is based on the GNDL, is developed to automatically connect the news media with the online social network.

.

List of Publications

Journal Publication

Li Ho Leung; Ng, V.T.Y.; Discovering associations between news and contents in social network sites with the D-Miner service framework, Journal of Network and Computer Application (JNCA) Vol. 36, issue 6, pp. 1651-1659.

Conference and Workshop Paper

Li Ho Leung; Ng, V.T.Y.; Chen Chen, "Analyzing social networks with D-miner Cloud," *in Proceedings of the 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD),* 23-25 May 2012, pp.642,648.

Li Ho Leung; Ng, V.T.Y.; Shiu, S.C.K., "Predicting short interval tracking polls with online social media," *in Proceedings of the 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD),* 27-29 June 2013, pp.587-592,.

Li Ho Leung; Ng, V.T.Y.; Shiu, S.C.K., "Collaborative Discovering Influences of News in Social Network Sites," *in Proceedings of the International Conference on Systems, Man, and Cybernetics (SMC),* 13-16 Oct. 2013, pp.712-717.

Acknowledgments

I would like to take this opportunity to express thanks to all people who have supported me when I was studying my Master of Philosophy degree.

First for all, I would like to thank my chief supervisor Dr. Simon Shiu and co-supervisor Dr Vincent Ng for their support and guidance. They taught me the correct attitudes and techniques in doing my research. Also, I would like to thank Dr. Stephen Chan and Dr Grace Ngai for their warm encouragements.

I would also like to thank the research students I have met, especially Victor, Shelly, Kenneth, Will and Petrie. We were facing the similar challenge and had encouraged each other.

Thanks for the support from my family. I could never start studying this degree without their understandings.

Finally I would like to thanks Gillian. You always support and encourage me to complete my study, even when I started giving up.

Table of Contents

Abstract **3**
List of Publications **4**
Acknowledgments **5**
List of Figures **8**
List of Tables **9**
Chapter 1: Introduction **10**
1.1 Research Objective 15
1.2 Research Problems 16
1.2.1 Generalizing the contents in online and offline social networks 16
1.2.2 Linking the contents in online and offline social networks 17
1.2.3 Transferring association algorithms across networks 17
1.3 Thesis Outline 18
Chapter 2: Literature Review **19**
2.1 Correlation Analysis among Virtual and Real Communities 19
2.2 Detecting Influences from Real Communities 22
2.3 Framework of Analyzing Virtual Communities 25
Chapter 3: Heterogeneous Social Networks Analysis **27**
3.1 Classifying the hosts of online and offline social networks 27
3.2 Acquiring corpuses from the hosts 30
3.3 The D-Miner Service Framework 32
3.3.1 Virtual Communities Collection (VCC) Agent 35
3.3.2 Real Communities Collection (RCC) Agent 36
3.3.3 Contents Linkage (CL) Agent 37
3.3.4 Association Discovery (AD) Agent 37
3.3.5 User Management (UM) Agent 37
3.4 Generic Networks Data Model (GNDM) 39
3.5 Generic Network Data Linking (GNDL) 42
3.6 Generic Network Data Association (GNDA) 51
3.6.1 Transplanted Aggregation Model 52
3.6.2 Customized Aggregation Model 54

Chapter 4: Experiments 61
4.1 Evaluating the NWCL Algorithm 61
4.1.1 Dataset Characteristics 61
4.1.2 Pre-Experiment Evaluations 62
4.1.3 Delimited Words Selection 65
4.1.4 Experimental Setting 66
4.1.5 Experimental Results 68
4.2 Evaluating the GNDA in Polling 71
4.2.1 Dataset Characteristics 71
4.2.2 Experimental Results 72
4.2.3 Discussion 73
4.3 Evaluate the GNDA in News Media 74
4.3.1 Data set characteristic 74
4.3.2 Experiment Settings 75
4.3.3 Experiment Results 76
Chapter 5: Conclusion 78
Chapter 5.1 Conclusion 78
Chapter 5.2 Limitations and Future Work 79
Bibliography 81

List of Figures

Figure 1. The D-Miner Service Framework 33
Figure 2. A Service Unit in the D-Miner Service Framework 34
Figure 3. Diagram of Online Social Network 39
Figure 4. Diagram of Offline Social Network 41
Figure 5. NWCL algorithm, Procedure 1a 47
Figure 6. NWCL algorithm, Procedure 1b 48
Figure 7. NWCL algorithm, Procedure 2 50
Figure 8. External Influences in a Social Network 55
Figure 9. Histogram of Word Frequencies 62
Figure 10. Word Frequencies distributions for 3 sets of articles 63
Figure 11. Distribution of the “old age living allowance” News Articles 75

List of Tables

Table 1. The Four Types of Service Units in Framework 38
Table 2. Statistics of the Word-Frequency Distribution 63
Table 3. Statistics of Word-Frequency Distributions for the three datasets 64
Table 4. Trials of Building Delimited List and their Performance 65
Table 5. Statistics of Word-Frequency Distributions for the three datasets 66
Table 6. Performances of the Three Segmentation Techniques 68
Table 7. Query Association Rate of the Three Segmentation Techniques 69
Table 8. Comparison of PSR and PSNR 72
Table 9. Performance of the Seven Correlation Models 72
Table 10. Clusters formed from Distribution in Figure 11 76
Table 11. The Social Influences Calculated by the 3 Approaches 77

Chapter 1: Introduction

In the recent decade, there has been a good number of emerging online social networking services providers, such as Facebook, Twitter and various kinds of web bloggers. Those providers offer various kinds of social networking services to their registered members through hosting websites in the Internet. Members can establish and manage their own connections with other registered members of the same service. Based on the connections established, which are approved by both sides, individuals can communicate with each other [8]. The interpersonal networks formed in those sites are generally regarded as online social networks.

A lot of efforts have been devoted to study the connections and activities of individuals, which are commonly known as behavioral patterns, in social networks. The behavioral patterns of individuals in online social networks were found similar to those in other social networks. For example, the homophily [19], which suggests people have a tendency to connect to others who have a certain level of similarity with them, is also applicable in online social networks [26]. It is observed that individuals who share similar backgrounds, such as education, interest or expertise, were joined together and formed a number of communities. In online social networks, individuals can arbitrarily join or leave a community and concurrently participate in multiple communities. Therefore, the scales of the communities in online social network are varied. For instance, some communities are large in size and have over a thousand of members; while some are very small and only have tens or even several members. The communities formed in online social networks are called virtual communities in this thesis. They divide the individuals in online social networks in different number of groups, where members in a group are having certain level of similarity.

Conversely, the community is an effective indicator to identify individuals in online social network who are similar to each other. In many cases, the social network analyses will not examine every individual in the network but only focus on the individuals in designated virtual communities [4]. This is because of two reasons. First, the online social network may be too big so that it is not possible to examine all individuals but select communities as samples; second, according to analytical topics, not all virtual communities will be relevant to the analysis. For example, if the analysis is about the feedback of a recent movie, we can focus on the communities about movie-watching or the communities formed by director or performers. They are very likely related to the topic. On the other hand, the communities of football clubs are probably not relevant and therefore they can be excluded. The screening of virtual communities can reduce the analysis time and possible noises in network contents hence increase the efficiency. Our research also focused on such approach.

In additional to the homophily and the communities induced, the social influences, which states individuals' behavior is affected by the connected individuals, is another behavioral pattern that can be frequently observed in online social network. In our daily lives, it is common that some of our friends are having stronger point of views. They often raise their idea and persuade others to agree instead of following decisions drawn by others. On the other hand, some people have opposite behaviors. They frequently support the suggestions raised by others but seldom share they own idea. And the first group of people, who is more active, is likely to affect the behavior of the second group of people. Similar behavioral pattern is also observed in online social networks. For example, some connected to more individuals in the community and frequently interact with them. The idea or opinions of those individuals are more likely to propagate along the connections in the online social network and affect the

behaviors of other individuals. On the other hand, some individuals seldom initiate conversations but they often participate in the conversations initiated by others. The concept of social influence is frequently applied in the studies of online social networks as it is capable of deducing measureable results to describe the relationship between individuals in a virtual community and inspiring upcoming studies. Belák et al. [4] proposed that the virtual communities in an online social network are often overlapped as one individual can join multiple communities. The social inferences can therefore be propagated from one community to another through the overlapped individuals. Furthermore, recent survey demonstrated that the majority of individuals own two accounts of different social networking services providers on average [6]. The participations of individuals in multiple online social networks imply that the online social networks are often overlapped. Therefore, the theories of social influences are not only applicable within an online social network, but they can also be deployed across online social networks.

Another group of studies suggested that the influential power of social network is capable of crossing the boundary of the Internet [2, 10]. As the members of online social networks are also individuals in our society, according to the theory of social influences mentioned above, the influential power can be propagated from online social networks to outside social networks and vice versa. In this thesis, the term "offline social networks" will be used to represent the social networks outside the Internet in order to distinguish from the term "online social network". In order to examine the social influences across the online and offline social network, it is necessary to identify the overlapped individuals in online and offline social networks. However, this is frequently infeasible as online social networking services providers allow their users to remain anonymous. The personal information of the users, which

is the key to determine their identities in offline social network, will not be disclosed. Therefore, the measurements of social influences across the online and offline social networks are often conducted indirectly. Instead of determining the overlapped individuals for tracing the information propagation and the social influences, the associations of the behavioral patterns between two networks, which indicate how similar they behave, are being studied.

If the associations between online and offline social networks are strong enough, it will be possible to estimate or predict the social behaviors in one network by studying the behaviors of another network. For example, we can estimate the public opinions in our society by studying the public opinions in an online social network instead of conducting public polls. Therefore, the time-delay in polling results generation, which is a critical issue in social, commercial or even political decision-making, can be shortened. For instance, in the summer of 2010, Apple Inc. faced one of its most serious public relationship crises. A number of customers reported that the iPhone 4, which was the latest product of Apple Inc. at that time, had signal problems with its antenna. Apple Inc. responded to this issue within a week and claimed that it is a common flaw and it could also occur in other phones [11]. However, the explanations were not satisfactory and had induced negative reactions from customers and other manufacturers of smart phone. A week later, Apple Inc. announced a new explanation but it could not dispel the doubts. Eventually, the controversy lasted for three weeks until Apple Inc announced its final conclusions. Throughout the incident, various public polls had been conducted, and they tried to estimate the impacts of the antenna incident on the company's revenues and reputation [16, 24]. The polling results indicated that the public opinions are varied across countries and regions. Moreover, the results of both public polls were released two weeks after the incident.

It was not possible for Apple Inc. to observe them before it responded to the general public. Therefore, although public poll is a tool that capable of reflecting the customers' opinion, the values of the opinions cannot be used if the waiting or processing time of the polling is too long. If Apple Inc could instantly obtain the public opinion from contents in online social networks and trace it throughout the incident, so that it could better estimate the impacts of the antenna incident as well as the company's actions. Hence, its public-relations strategy could be adjusted to minimize or even avoid the unfavorable consequences happened in the case.

In the above example, the contents shared in online social networks were affected by the antenna incident, so that a number of discussions about the incidents were occurred and grown continuously. On the other hand, the discussions reflected the timely feedbacks from individuals and they were strongly related to the public opinions. As the public opinions need relatively longer time to collect and processed, the strongly associated opinions in social networks would be a good substitute. Nevertheless, the levels of association are varied among the online social networks and some social networks showed no associations with the public opinion. During the incident, a number of online votes were hosted in blogs and forums to independently collect opinions from individuals and their responses are inconsistent and sometimes contradicted with each other. The phenomenon highlights that not all online social networks can efficiently represent the public and they are sometimes biased.

Therefore, when one social network analysis performs significantly well (bad) in one online social network, it is not appropriate to deduce the analysis is outstanding (poor) in general. This is because the information in the social network could be biased and the analysis could be exceptionally outperformed (underperformed). However, the

generality of social network analyses are not the primary concern of most studies. A lot of social network analyses are often customized in order to resulting higher performances in accuracy or reliability for designated online social network.

Nevertheless, the research efforts devoted to the online social networks are in great difference. For example, the multi-national services providers like twitter.com and Facebook have enormous user bases, and they are the favorite focuses of study getting great attentions from the researchers. On the other hand, some small or medium platforms like the internal communication network of a college or discussion forum in a city have relatively fewer users and they failed to attract sufficient attentions from the scholars. In addition to the user base of the sites, the contents available also determined whether the site can attract sufficient attentions or not. For instance, the performances of stock market and results of political voting are typical examples which are capable of drawing great attentions from the scholars. As a consequence, most of the research efforts were input to studying the social network services providers where they have large user base and popular topics in academic research are being discussed. Multiple analytical approaches have been developed for those popular social network services providers, but their analytical standards and outcomes are often varied hence they are not comparable with each other. In contrast, the small and medium service providers are fail to receive sufficient attentions hence limited or no analytical standards have been derived for them.

1.1 Research Objective

Our work attempts to generalize the association analyses between online and offline social networks. In ideal situation, the association analyses of any arbitrarily selected pairs of online and offline social networks can be fitted into one generalized abstract

association model. As a result, the association analytical techniques developed for popular online social network sites can be first converted to such abstract model. The abstract model can be further converted and applied those topics which were not being concerned, or it can be applied to those small and medium social networks where limited studies have been done on them. The development efforts can therefore be greatly reduced and performances can be guaranteed. The proposed techniques are regarded on transferring of analytical analysis from popular social networks to unpopular social network and they were found similar to the transplantation of organs. Therefore the research objective is renamed to "transplantation of association analyses between social networks".

1.2 Research Problems

Our work is based on an assumption where the contents of those interested online and offline social networks have already been collected. And association relationship(s) between one or several pairs of online and offline social networks is proven to exist. This chapter will present the three main problems as well as their corresponding sub-problems in transplanting the correlation analyses.

1.2.1 Generalizing the contents in online and offline social networks

Online and offline social networks are various in structure and contents. Although it is assumed that contents in the networks are successfully collected, it cannot guarantee that all collected social networks contents are comparable with each other. This is because the contents are from heterogeneous sources and they are often incompatible with each other. Therefore, it is necessary to define general criteria to restrict the contents eligible to be generalized. In other words, if a social network site cannot

fulfill the criteria, it will be indicated as inappropriate to participate in the transplantation.

1.2.2 Linking the contents in online and offline social networks

The second problem is induced by the scalability of online and offline social networks. As too many contents are available in social networks, it is not possible to conduct association analysis for all contents. Moreover, different from other studies, where the associations of designated social networks are studied for designated purpose(s), our work aims to "transplant" the analysis mechanism from other applications to fit our needs. Therefore the contents generalization problem raised in the previous section can also simplify the comparison efforts in heterogeneous social networks. The next problem to solve is to connect the generalized contents of online and offline social networks where they are likely (or just more likely) to be associated. In other words, this problem is about how to preselect or scan the online and offline social network contents before they are proceed to the association analysis. Another problem about the deduced connection is the reusability. As mentioned above, our research attempts to cover multiple social networks. If the connection, which indicates those online and offline social contents where they are possible to be connected, is not reusable (i.e.: it only describes the relationships of the designated social network pair), a lot of connections have to be deduced. Their standards may be different hence induces noise in comparison and reduces the computation accuracy. Therefore, the deduced connection should be independent from the social networks or the contents.

1.2.3 Transferring association algorithms across networks

The next problem will be how to transplant the association analysis algorithm(s) to other social networks or applications. Similar to the transplantation organs, the first sub-problem will be matching a pair of a contributor and receiver. In our study, the

contributor is a social network site pair where one or more association analyses have been conducted on it. On the other hand, the receiver is a social network site pair where their association is rarely or even never studied and it needs to conduct association analysis. The contributor and the receiver should have a certain level of similarity in their contents hence the transplantation and be conducted smoothly. The second sub-problem will be re-building the same association analysis on the receiver side. This sub-problem has two phases. The first phase is to convert the analysis used by the contributor to the generic association model mention in the problems raised in the previous chapter. The second phase is to apply the model in the receiver. Since the contributor and receiver are different in most cases, customizations on the association model are expected to be performed.

1.3 Thesis Outline

The remaining part of this thesis is structured as follows. First, Chapter Two will review the theories in the related former studies and highlight those are strongly related to our research topic. Then, Chapter Three will describe the proposed solutions to the 3 main research problems. Those solutions are integrated into a service based SNSs analysis framework, which is named as D-Miner Service Framework (the D-Miner framework). Next, Chapter Four will present the experiments of the solutions presented in Chapter Three and demonstrate the corresponding evaluations. Afterward, Chapter Five will conclude our study and discuss its limitations and future work.

Chapter 2: Literature Review

A social network is often modeled as a social graph G = (V, E), where V is a collection of nodes and E is a collection of edges which connect the nodes. The nodes represent people in the social network and links represent their social relationships [23]. As demonstrated in former studies, the social graph is capable of modeling the social activities in social networks and studying the behavior patterns of the individuals. And based on the models, two kinds of measurement, the edge measures and the node measures, can be derived from the graphs. The edge measurements indicate the connections between two nodes and deduce their strength. A strong link indicates the two nodes are closer to each other and a larger number of their neighbor nodes are overlapping. Therefore, information or opinion can be effectively propagated between the two nodes hence they have a higher probability to influence each other. On the other hand, the node measures determine the social influences of a node by evaluating its importance in the network. Centrality measurement is one of the typical examples of node measure. It determines the number of neighbor nodes who are influenced by the given node. If a node has more influenced neighbor around, it has higher influential power in the network. In additional to the general definitions and usages of the social graph, Wang et al. [25] observed that a node in a social graph usually involved in multiple communities while a link is not. Therefore, the overlapped communities in a social graph are separated by clustering the edges and recovering their corresponding communities.

2.1 Correlation Analysis among Virtual and Real Communities

The correlation analyses among virtual and real communities had started drawing much attention in recent years when compared with those traditional knowledge

discovery studies in virtual communities, for instance, the deduction of people's opinions, communities and interactions.

The correlations between the people's behavioral patterns in online social media and people's behaviors in the society have been extensively discussed. Much efforts are focused on demonstrating whether the behavioral patterns observed in virtual communities is leading, lagging or irrelevant to the behavioral patterns in real communities. Gruhl et al. [10] studied a selected number of books on how related discussions in blogs and sales ranks in amazon.com are correlated. They observed that spikes in blog-discussions of a book led the spikes in its sales rank. Balog et al. [3] have continuously studied people's moods in blogs. They have defined different kinds of moods such as sad and excited; and identified them in the blogs by conducting sentiment analysis. The moods were aggregated and named as global mood. They discovered that the spikes of global mood level in blogs have been coincidenced with peaks of excited mood in the real world such as the release of the new volume of Harry Potter, one of the most famous fictions in the world at that time. These studies showed that people's behaviors in online social media and the society have close relationship with each other. Although the studies only focused on the issues of books and their` sales, they have still demonstrated the connections between online social media and the society are exist. Later, Asur et al. and Bollen et al. [2, 7] extended the applications of correlations in disparate areas. Asur et al. [2] studied the discussions about a selection of movies in Twitter. They aggregated the sentiments extracted from the tweets in a week by using the level of attention and positive-negative-sentiment ratio. The two aggregations were found correlated with box revenues of the corresponding movie. Hence they could be used to predict the movie revenue to be announced in coming weekend on a weekly basis. Bollen et al.

[7] attempted to forecast the performances of stock market in United States by tracking the mood of tweets in Twitter.com. They first defined disparate aspects of mood such as “Alert”, “Happy” and “Calm”, where they represent the mental behavioral pattern of individuals in the social network. The finding demonstrated that the combination of “Calm” and “Happy” moods was best correlated with the Dow Jones Industrial Index, which effectively represents the general performances of the stock in the America stock market. Those two moods can form an effective indicator which guides the future performance of the index.

The above studies provide sufficient evidences to support the existence of correlation among behavioral patterns in virtual and real communities. However, they do not conclude if the two communities are influencing each other. As shown by the regression models applied, both studies [2, 7] assumed the discussions in the virtual communities are independent from any influences. Nevertheless, some other studies which worked on studying the relationship between news and virtual communities demonstrated the assumption does not always hold [3 13, 15, 20]. Such controversy and its impacts will be discussed in the next subchapter. To summarize, the existence of correlations among virtual and real communities have been confirmed and examined by a number of studies; but the independency of behavioral patterns in virtual communities is still not clear. However, when compared with other evaluations of correlation analysis like accuracy and performance, little attention has been devoted to the formation of correlations like tracing how information is propagated within or across virtual and real communities.

Connor et al. [9] suggested the opinion from online social media is also correlated with the opinion from public polling. They studied this kind of relationship by comparing

the sentiments extracted from Twitter with the measurements about consumer confidences and political preferences in the United States. Those measurements included Index of Consumer Sentiment, Confidence Index and tracking polls during the 2008 United States presidential election cycle. The collection of data followed standard sampling techniques. As the sentiments obtained from Twitter messages were found rapidly changing, they have been aggregated by applying smoothing techniques in the experiment. Therefore, the aggregated sentiments would respond slowly to recent changes. The experimental results demonstrated that the aggregated sentiments in Twitter correlated well with consumer confidences index but they could not provide good prediction of political measurements at all times. It was suggested that high complexities of questions in political polls and lacking in representative in Twitter's population would be the two possible reasons leading the aggregated sentiments showed no correlation with the public

2.2 Detecting Influences from Real Communities

A number of scholars have studied the news in real communities and behavioral patterns in virtual communities. They discovered that the contents of news articles are often reflected in virtual communities. Java et al. [13] studied the user behaviors in Twitter and summarized "reporting latest news or commenting about current events" in real communities as one of the four main intentions of using Twtter. Kwak et al. [14] investigated the topological characteristics of Twitter and they found out that more than four-fifth of its trending topics, which referred to the most often mentioned phrases, words, and hashtags those identified by Twitter, were "headline news or persistent news (occurred in real communities) in nature". Mishne and Rijke [20] studied the searching habit of blogs and found out more than one-fifth of the top

ad-hoc queries, which are queries entered by users were related to news in real communities. Although the studies did not explicitly stress that virtual communities are being influenced by real communities, they demonstrated such influences do exist. On the other hand, the influences from the real communities introduce uncertainties to the correlation analysis because they are not being considered in the analysis models. This is because it is unclear that how the influences could affect the correlations among virtual and real communities. In the worst case, behavioral patterns in both communities are influenced by incidents in real communities and their correlations are totally relied on the incident but cannot exist by their own. The work presented in this thesis aims to cope with the possible uncertainties brought by influences from real communities by introducing a multi-agents framework. For instance, the crawler agents collect and archive the incidents in the real communities while the association agents detect the associations among virtual and real communities. The results will be further analyzed for the tracing of information-propagation. Here, we propose to use the information in news articles to represent the incidents occurred in the real communities. The articles will represent the recent incidents and be compared with the virtual communities and their possible associations will be identified.

Tsagkias et al. [29] studied the text-messages shared in virtual communities and news articles. They defined two types of connections. The first type is explicitly linked, which means the text-messages are referred to a particular news article via hyper-link or references. The second type is implicitly linked, where the text-messages "directly discusses the article's contents", but "not merely about the same topic as the source news article". Ikeda et al. [12] attempted to link news articles to virtual communities' messages by using word vectors, which are sequences of words extracted from the title and the first sentence of news article or the entire contents in virtual communities'

messages. The experiment compared the word vectors with different weighting approaches and similarity metrics. The results concluded that Inverse Document Frequency (IDF) is a better weighting approach when compared with Term Frequency – Inverse Document Frequency (TF-IDF), while inner-product was a better similarity metric when compared with cosine function. Phelan et al. [22] proposed a novel ranking system of emerging topics and breaking events which applies the linking technique. The system compares the similarity between news articles and the messages in virtual communities. For each article, its total number of similar messages is counted, and the articles with more similar messages will be ranked higher.

The time information is often involved in social influences analysis in two aspects. The first one focuses on the time delay of the social influences. In other words, after the emergence of an influence factor, it is interesting to know how long it will take to influence individuals. Myers et al. [21] modeled the behaviors of individuals in online social networks after they are exposed to external social influences. The second aspect focused on the life-cycle of social influences. It is interested in the changes of the social influences along with the time. Belák et al. [4] traced the cross community influences in an online social network over time and summarized the most influential communities and the communities which are easiest to be influenced.

The association to be presented in this thesis is different from the previous work in two aspects. First, we have an information definition. Instead of "linkage", which is pre-defined by other scholars and it required a message to reference or mention a news article, the definition to be deployed in our work will also include the messages which mention the same topic as the news article. We name such kind of relationship as

"association" in order to differentiate it from the "linkage" deployed in other studies. Second, the scope of comparison will be expanded. Instead of describing the relationship between news article and one message in virtual communities, the association will be applied on discussions, which consist of series of messages related to the same topic, in virtual communities. This is because the application of association in our work does not aim to study particular message(s) but interested in starting of discussion and propagation of information. The D-Miner Service framework to be introduced has been utilized and extended in order to identify the associated news articles and discussions in virtual communities. Detailed reviews and evaluations about the framework will be provided in next subchapter.

2.3 Framework of Analyzing Virtual Communities

Following to the increasing attention to the analysis of virtual communities, different forms of analytical frameworks and architectures have been proposed. Zhang et al. [31] designed the architecture of a Dark Web Forums Portal. It supports searching and browsing functions for almost thirty designated forums in five different languages. The architecture is capable of processing data from multiple virtual communities, which are probably different in forms and structures. On the other hand, the architecture does not further interpret or process the web content and just stores them as web pages. Moreover, although nearly thirty sites are supported by the architecture, all of them are forums therefore the architecture only supports homogeneous virtual communities. Ting et al. [17] proposed architecture of cloud-computing-based data warehouse and virtual communities' analysis system. It adopted crawling agents to collect data and store them in distributed environment. Cloud computing provides extensible services on servers and users can access these

services through the Internet anytime, anywhere, and through any platforms [6, 28, 30].

Although OSN sites analysis is a popular research topic, most of the studies focused on analyzing a designated OSN site. Many works focused on how to coordinate the data collection together with data analysis as a tightly-coupled application, which can be considered as a vertical integration approach [1]. The coordination of functions belonging to the same level, which can be considered as horizontal integration, is seldom discussed. Tang and Yang [27] presented a framework for social networks integration with privacy preservation, which is capable of combining two social networks for further analysis. The D-Miner Service framework proposed in our work is evolved from web-based application and aimed for analyzing virtual communities [6, 28]. The framework integrates heterogeneous virtual communities like Facebook, Twitter and Yahoo! Knowledge, by converting their contents into a common storage format. The framework is composed of multiple analytical, collection and scheduling agents which work cooperatively or independently to offer analysis services of virtual communities to framework users.

To conclude this chapter, we work under the assumption where virtual communities could be influenced by incidents in the real communities. The focuses our studies are about validating and evaluating the correlations identified among real and virtual communities. They are conducted by tracing the information propagation among the two communities and inspecting possible influences. Such idea is implemented by adding two agents to the service framework. The design, functions and performance of the agents will be discussed in detail in chapters below.

Chapter 3: Heterogeneous Social Networks Analysis

This chapter describes the methodology to solve the 3 main research problems mentioned in chapter 1.2 and the related algorithms are described. It will be started by defining the 2 different types of social networks, so that the scope of our work is deduced. Then, it is followed by describing the methodology to collect information from online and offline social media. Hence the assumption stated in the beginning of chapter 1.2 can be fulfilled. Afterward, it will be the introduction of the D-Miner Service Framework, where it integrates all the solutions for the research problems. Finally, the techniques suggested in our works will be introduced and explained one by one.

3.1 Classifying the hosts of online and offline social networks

Online social networks can easily be found in the world-wide-web as there are a number of online social networking services providers host such kind of services. Targeting on different chapters of customers, the providers offer disparate kinds of social networking services. Various terms such as Social Media, Social Network Site and Online Social Networking Site are often used to describe the hosts of the service providers. The meanings of those terms are similar but they are varied in detail. Therefore, it is necessary to properly elaborate the term(s) to be used in our studies before effective academic discussions can be conducted.

Kaplan and Haenlein [14] studied the challenges and opportunities of Social Media. According to their definition, social media is not just a collection of Internet-based applications which enables the creation and sharing of user generated content; it can be further categorized according to level of self-disclosure and media richness.

Therefore, both collaborate projects in the Internet such as Wikipedia, where information is shared among individuals, and virtual game world, where communities or aligns among players can be formed, are regarded as social media. However, the services offered by the media are in great difference. On the other hand, according to the study conducted by Boyd and Ellison [8], where they reviewed the definitions and scholars of social network sites, both Social Network Site and Social Networking Site have similar meanings and both of them appear in public discourse. In fact, their applications are often interchangeable. However, the term "networking" emphasizes the connections established among the individuals and it cannot describe our work well. Therefore, the term Social Network Site (SNS) is adopted in this thesis as it better reflects positioning of the social networking services providers in our research. In order to identify the Social Network Sites (SNSs) from other sites which enable the sharing of user generated contents, the study suggested three criteria. First, the SNS should enable individuals to have their profile pages for sharing information about themselves. Second, it should allow individuals to announce a list of other users who share connections with them. Third, the connections recorded in the lists can be modified by their owners and the changes will be propagated to other's lists.

Similar to the observations raised by Boyd and Ellison, it was noticed that there was increasing number of websites, which merely provided the services of hosting user-generated contents in former years, had been enhanced in order to offer social network services to their users. As a consequence, there are still a number of sites which are regarded as SNS but they are offering totally different social network services. For example, according to the definition, both flickr.com and twitter.com are considered as instances of SNSs. However, the services offered by flickr.com are focused on photo sharing; hence more images about various topics can be captured

from the site. On the other hand, the focus of twitter.com is sharing of short-length text, and more text-messages about recent issues can be discovered in the site. The information available in the some SNSs can be in great difference and it is not possible for their analyses to be transplanted. Therefore, it is necessary to define some selection criteria. And only the SNSs which fulfilled the criteria, they can receive analysis from or transplant their analysis to others.

On the other hand, there are many influential factors occurring in the offline social networks every day and inducing social influences to the online social networks. For example, the release of a new model of popular smart phone is influential factor as this piece of information will be propagating rapidly in the corresponding virtual communities; the opening of a new restaurant next to my apartment is also an incidence because people may share their dinner in that new restaurant; even I got a "Fail" in my examination can also be considered as an influential factor, because I may receive responses from somebody who is in the same situation if I share it in the online social network. However, not all incidences can induce a discussion which is rich enough for analysis. Wrongly selecting an influential factor, where it is very unlikely to influence the SNSs, to analyze will waste time and resources. The influential factors interested in our study are those having high potential to or already induced discussions in the offline social networks. Therefore, it may also likely to induce discussions in the virtual communities in on line social networks. One of the typical hosts which broadcast the information in offline social network is the News Media. Therefore, Online Social Network Sites and the News Media are selected as the corresponding hosts of online and offline social networks.

3.2 Acquiring corpuses from the hosts

The next sub-problem faced in our study is to collect the user generated contents in those small and medium SNSs. In other studies, this problem is seldom discussed in detail. This is because most of the studies are targeted on those multinational SNSs which have a large user base and their discussions covered numbers of topics so that they have relatively higher academic value. Various kinds of techniques had already been built or customized to acquire users' messages from those SNSs, and little efforts are required to collect data from those sites. Besides, a number of those sites actively offered their corpuses, where personal information of the members is masked, for academic use. Therefore, the acquiring of users' messages is not an important problem in analyzing those popular SNSs. In contrast, those small and medium SNSs rarely provide their corpuses and academic studies are seldom conducted. Therefore, the user generated messages from small and medium SNSs, especially those being posted recently, are not handy to be used in our research. As a consequence, we have to collect the messages from the sites by our own. Nevertheless, even though it is just a small and medium SNS such as a discussion forum, over thousands of messages are being generated by its users every day. It is not feasible to collect the corpuses from the SNSs manually. Therefore, we develop our web crawlers, which are a kind of mature technique to collect web resources, in this study. It saves human resources by automatically collecting corpuses from designated small and medium SNSs on a regular basis.

Although the using of web crawlers can greatly reduce the efforts required to acquiring contents from the small and medium SNSs, one main problem needed to be solved before such technique can be applied in our study. As mentioned in the previous chapters, our work aims to generalize the available correlation analysis between

heterogeneous SNSs and the external incidences. Therefore, the text messages from the sites, where their structures and contents available are different, have to be collected for our study. The web crawler introduced above only offer an easier way to collect messages but the differences among the collected messages still remained. Therefore, the next problem in corpuses acquiring is to resolve the differences among messages collected from the heterogeneous sites. This problem can be further decomposed into two sub-problems. The first is to define the types of information, which are available in those SNSs and are extractable. Therefore, text messages from heterogeneous sources can be compared. The second is to convert the heterogeneous information into a common form for further analysis or comparison.

After the incidences interested in our study are defined, the second problem concerns how those incidences are being chosen and collected because it is not possible to search incidences in the real community without directions. Our study applies the incidences announced by third parties' channels. This is because when an incidence is being announced in a channel of a third party, it implies that such incidence had already attracted the attentions of the party members. And it will likely to attract more attentions from the audiences after it is being announced, hence it is relative easier to induce discussion in SNSs. Three sub-problems have to be solved when we adopt the incidences announced by the third party's channels in our study. The first sub-problem is about how the channels are being selected. There are various numbers of channels such as radio broadcastings, newspaper and public assembly, and they all discuss the incidences which are interested in our study. There are three main concerns when selecting channel(s) for our study. The first will be the transmission media adopted by the channels. Some channels discussed above use speech or body language to transmit messages to their audiences while some of them are using texts.

It is necessary to define a scope of transmission medium where the information carried by them can be effectively extracted and compared with those available in SNSs. The second concern in choosing the channels are the frequency which the channels announce information. Some channels rarely announced messages or they did it in extremely low frequency. Such kinds of channels are not suitable to be used in deducing correlation relations. This is because there are not enough of cases or instants to support the deduced relations. Therefore, those channels which regularly announce messages are more suitable for analysis because a number of cases or instants are released in a steady manner. The final problem will be resolving the difference among the heterogeneous information collected from the channels. Similar to the situation in SNSs, those channels are established by different parties and their data presentations and formats can be totally different. In order to enable a SNS message to be compared with multiple external incidences under a standard comparison mechanism, it is necessarily to generalize the information collected from the channels.

In order to better describe the solutions proposed in this thesis, a service based SNSs analysis framework, which is named as D-Miner Service Framework (the D-Miner framework), will be introduced in the next section. It contains various kinds of components where each of them solves corresponding problem mentioned in the previous chapter. The components are integrated in the framework to provide a complete solution for the research problems.

3.3 The D-Miner Service Framework

The D-Miner Service framework is an integration of solutions of the research problems raised in the previous chapter. The framework is based on cloud architecture and it

aims to deliver the social network analysis as web services to its users. The framework defines the protocols where the components have to be followed when they are interacting with each other and how system resources should be allocated.

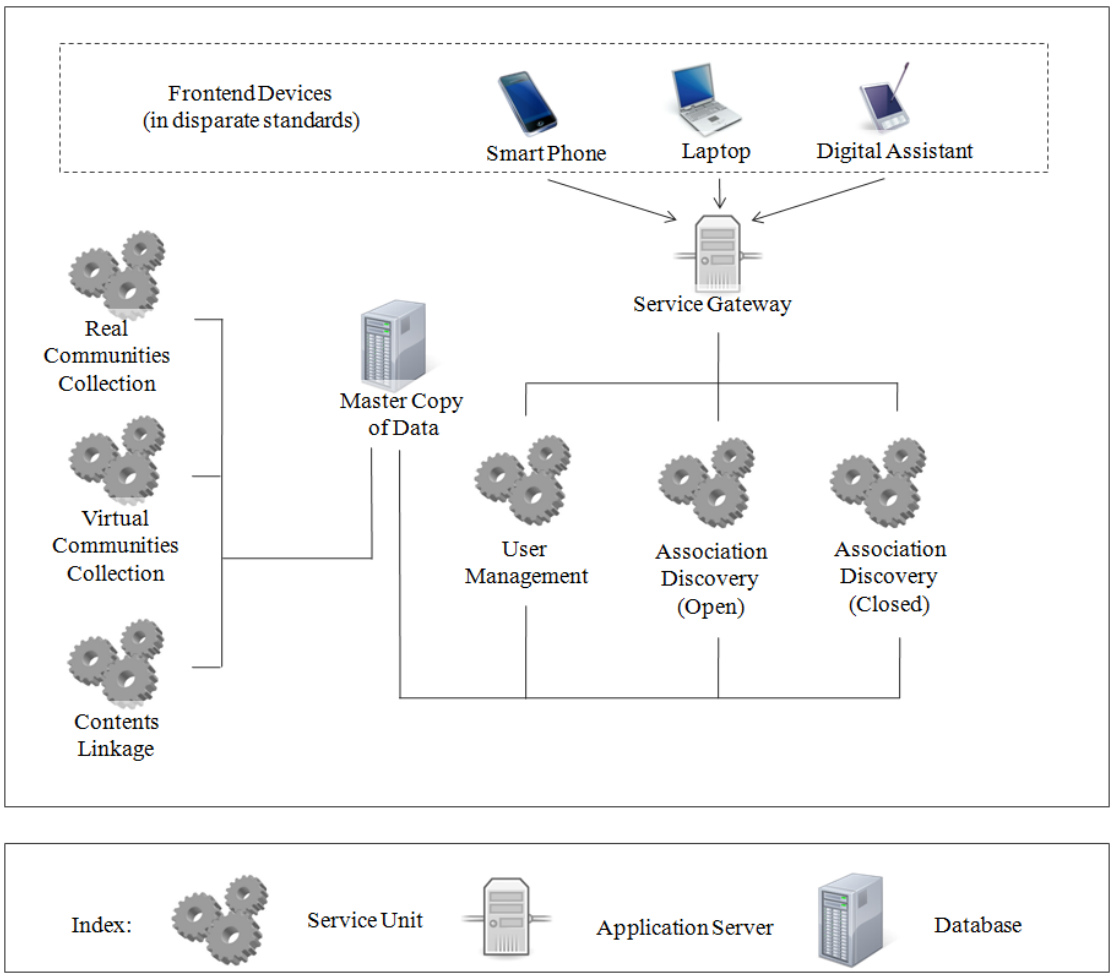

FIGURE 1. THE D-MINER SERVICE FRAMEWORK

Figure 1 illustrates the structure of the service framework as well as all its components. Users can start analyzing the virtual communities at any times and locations by sending out requests through their devices (i.e. the frontend devices presented in Figure 1) to the service framework via the service gateway. The functionality of the devices could be limited (like mobile phones) or relatively strong (like personal computers). The service gateway is an application server, which is responsible to deliver services to the frontend devices. It will interpret the requests and redirect

them to corresponding service units. After the requests are processed by the units, the gateway will forward their responses back to the frontend devices.

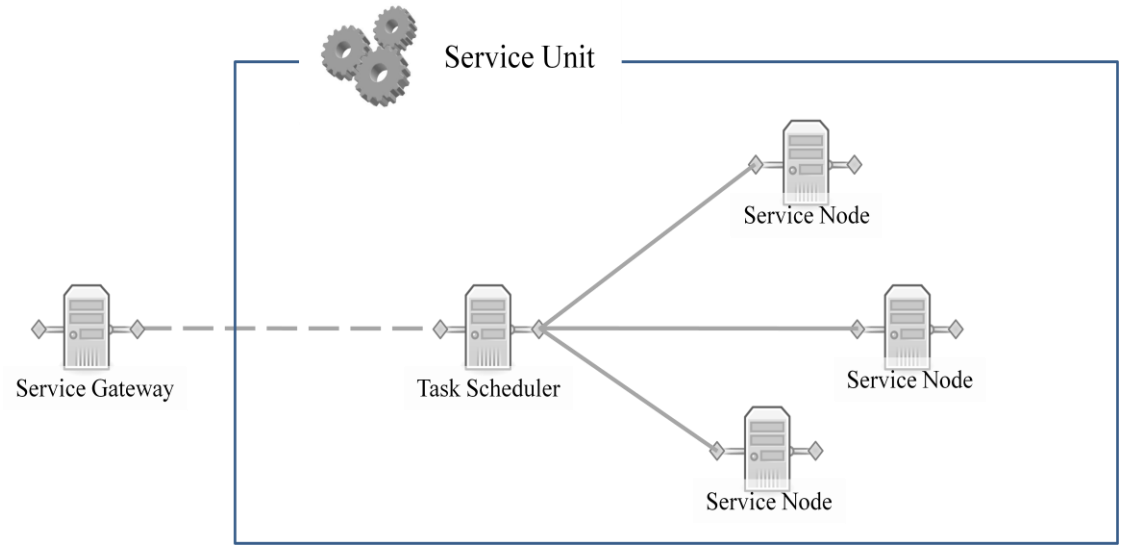

FIGURE 2. A SERVICE UNIT IN THE D-MINER SERVICE FRAMEWORK

A service unit is a small computer network which conducts the web services specified by the service gateway to the frontend devices. They are transparent to frontend devices but they determine the performances of the entire framework. As shown in Figure 2, each unit is composed of several application servers, where they are named as Task Scheduler and Service Nodes. The Task Scheduler is responsible for assigning incoming requests to the service nodes behind. The assignment is based on selecting the correct nodes for the tasks and balancing their workload. A Service Node is a server that conceptually contains a processor and a database. The processor is capable of interpreting the requests from the scheduler, executing the services, forwarding the request to other services nodes if it is needed and sending back responses; while local database stores analytical data, results and logs generated. The number of service nodes deployed in a service unit will significantly affect the unit's performances. Therefore, it will be regularly reviewed and adjusted according to the workloads and performances of the service unit.

According to their intentions, the six service units in the D-Miner service framework can be classified into five kinds of agents, where they offer different types of service to the users. Those agents are known as: Virtual Communities Collection (VCC) Agent, Real Communities Collection (RCC) Agent, Contents Association (CA) Agent, Association Discovery (AD) Agent and User Management (UM) Agent. The first four agents offer core functions and the last one provides utility function. In the remaining part of this chapter, the five agents will be discussed one by one.

3.3.1 Virtual Communities Collection (VCC) Agent

The VCC Agent attempts to solve the classification problem specified in chapter 3.2. It is responsible for acquiring corpus from those virtual communities. The collection process is being controlled and only designated kinds of information, where they are related to the analysis, will be subject to be collected. The candidate virtual community can be an open community like discussion forum, where most or nearly all information is publicly accessible; or it can be a semi-closed community like Facebook, where the accessibility of information is limited and the available information will be varied according to the identity of the login user. The agent employs a series of web crawlers to collect the information from open communities. Those crawlers are customized according to the structures of SNSs where the open communities are located. They are regularly executed to collect communities' information from the sites where no users' authentications are required. The data collection is automatically triggered and it needs not to notify the end users. As a result, the data archived are being grown steadily and silently. On the other hand, the data collection of those semi-closed communities is different. Since the information in those sites in delegated for certain group(s) of users, the information available will be varied according to the identity of the logged-in user. Therefore, the accessing of the

achieved data will be restricted. Those data will be encrypted and only the users who downloaded the data can access or analysis them. The contents collected from the VCC will be converted into a predefined common data format. The data collected from the VCC are followed to the Generic Networks Data Models (GNDMs), which will be discussed in the next chapter.

3.3.2 Real Communities Collection (RCC) Agent

Similar to the VCC Agent, the tasks assigned to the RCC Agent is also about the collection corpus. Nevertheless, the RCC Agent is targeted on collecting the incidences of the real communities. As discussed in the chapter above, the RCC agent will collect those incidences announced by third-parties' channels. According to the designs of those channels, there will be small variances in the collection approaches. The RCC Agent also follows the Generic Networks Data Models (GNDMs), which will be introduced in next subchapter, so that all required data will be collected. The data collection where be executed twice a day, the retrieved data will be automatically archived and the duplicated data will be removed. Both the VCC Agent and the RCC Agent work together to response to the corpus acquiring problem mentioned in chapter 3.2.2. The two agents collect the source data from the Internet and more importantly, generalize the heterogeneous data by converting them into a common format. Without these two agents, the CL Agent and the AD agent are not able to function. However, the techniques adopted by these two agents are either mature or it is offered by third parties. Therefore, the techniques adopted will not be discussed in details in the following chapters but the source data collected by them will be review and discussed in detail in the experiment chapter.

3.3.3 Contents Linkage (CL) Agent

The CL Agent solves to the generic connections problem specified in chapter 1.2.2. It attempts to connect the archived SNS messages and the incidences collected from the third parties' channels, where they are possible to be associated. In other words, the CL Agents filtered the SNS messages (incidences) which are not possible to be correlated with the incidences (SNS messages). The advantage of adopting this agent is that we can quickly identify the SNS messages (incidences) to be analysis once the incidences (SNS messages) to be analyzed are determined.

3.3.4 Association Discovery (AD) Agent

The AD Agent solves to the generic connections problem specified in chapter 1.2.3. This agent is capable of deducing the association between social networks. In our design, the AD Agent will be frequently introduced when different disparate association analyses are being compared. The framework design enables the AD Agents to be plug-in and plug-off easily hence the AD Agents and the association algorithm inside can be added, tested and modified in short period of time.

3.3.5 User Management (UM) Agent

The UM Agent manages the users' authorities and executes the predefined access restrictions of end users on the service framework. For example, it determines what kinds or agent(s) will be available to the connected frontend devices. For instance, a group of end users can only access the VCC or RCC agents but they are not allowed to access the AD agents; while another group of users can access all agents in the framework, but they can only send analysis request to the AD agents five times per day in order to control the system usages. This agent is considered as utility functions in the framework because it does not offer any functions which are critical in the analysis. Therefore, this agent will not be discussed in detail in this thesis.

Table 1 summarizes the six agents in the framework according to their invoking mechanisms and connectivity to web services offered by third-parties. Among the six agents, the SNSs Collection agent has two variations. The Closed Virtual Communities Collection agent targeted on communities like Facebook and Twitter, where their data are user-specific and user authentications are required; the Opened Virtual Communities Collection agent targeted on those communities, where their data are freely accessed by any users in the Web.

	On-User-Demand	System Routine
Externally Connected	✧ Closed Virtual Communities Collection	✧ Opened Virtual Communities Collection ✧ Real Communities Collection
Internally Connected	✧ Association Discovery ✧ User Management	✧ Contents Association

TABLE 1. THE FOUR TYPES OF SERVICE UNITS IN FRAMEWORK

In the service framework, executions of web services can be initiated by a frontend device or the internal system. In the prior case, an external request is sent to the gateway for retrieving designated resources. The gateway will first checks the user's authority from the UM Agent. If the user has such authority, the gateway interprets the request and decides which service unit is responsible for offering the required service and transmits the request to that unit. The request will first arrive to the task scheduler in the unit. The scheduler interprets the request, selects the service node(s) that is most suitable for processing such request and sends a request to that service node(s). The service node either processes the request or forwards the request to corresponding SNS site(s) via the APIs offered by the sites. Here, this step is most likely to be a bottleneck as it involves many service invocations and transmissions.

After the process is completed or a response is transferred from the site(s), the service node forms the corresponding responses, and they will be propagating back to the frontend device through the gateway. In the latter case, internal requests are generated and they are sent to corresponding service units. Although the rest of procedures are similar to the prior case; frontend devices will not be involved. Therefore, this type of service model is used by internal routine processes such as the collections and associations of news articles conducted by RCC Agent and CL Agent.

3.4 Generic Networks Data Model (GNDM)

This subchapter describes the data models defined in our work. They are used to describe the contents in virtual and real communities. The GNDMs are generic, they are not only be used in the service framework, but can also be used in other applications.

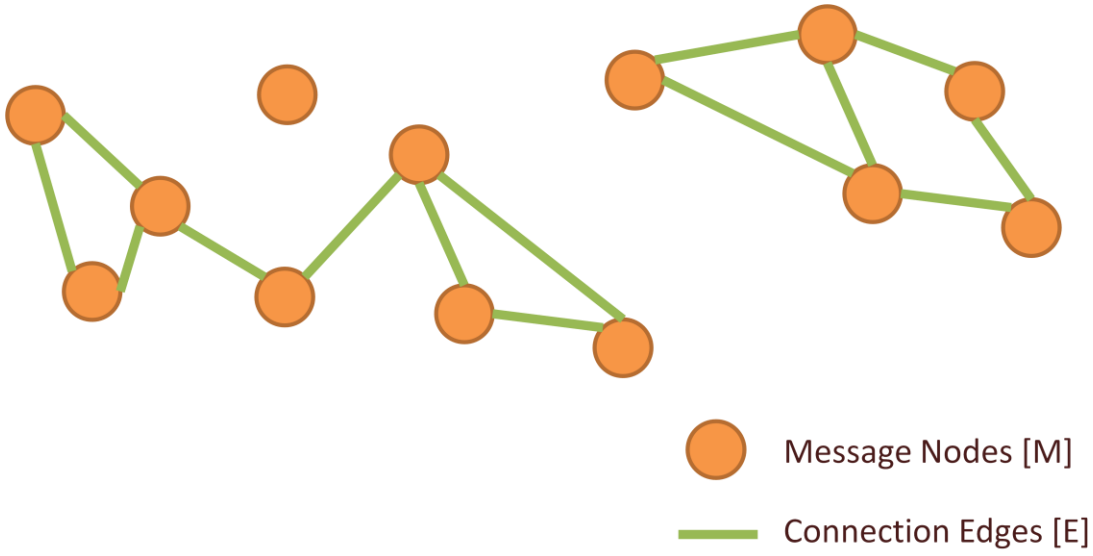

FIGURE 3. DIAGRAM OF ONLINE SOCIAL NETWORK

Figure 3 illustrates a simplified online social network which is composited of nodes and edges. Each node represents a messages shared in the network, while each edge

represent the connections among messages. Connections is being established when one message is being replied, referenced or quoted by another message. Nodes are connected by edges thus forms a discussion and individuals participated a discussion by sharing composed messages. Three discussions are formed in fig. and they are varied in size. One of those discussions only has one message, and this implies that such message is isolated in the network as it is being replied, referenced or quoted by other messages.

The graph in Figure 3 can be represented as an equation: G = (M, E) where M is a collection of nodes and E is a collection of edges which connect the nodes. Each node represents a message while an edge refers to the connection between two messages such as reply, reference and quotation. Let M be a node (message) collection and $M_k = \{m_{k1}, m_{k2}, \ldots m_{kn}\}$, where n is the total number of messages in online social network k and n >= 1. Each message m contains three elements: an author tag α which can uniquely identify an author within the online social network (k); a message content c records the texts shared by the author, where the contents contributed by other authors, such as quotations, are removed; a posting time t which indicates when the message is being shared to the network. Hence $m_{ki} = \{\alpha_{ki}, c_{ki}, t_{ki}\}$. On the other hand, we have $E = \{e_1, e_2, \ldots e_y\}$ edges, where y is the total number of messages connections and y >= 1. Each connection e contains three elements: a source s record the message ms which replies, quotes or takes reference of other messages; a destination d record the message m_d which is referred by the source message; and a reference format $F = (f_1, f_2, \ldots f_b)$, where b>=0. It is a vector which indicates the kind(s) of interaction that the connection e has. The value of b is the total number of referrals supported by the data model, and an edge is represented as $e_i = \{s_i, d_i, F_i\}$.

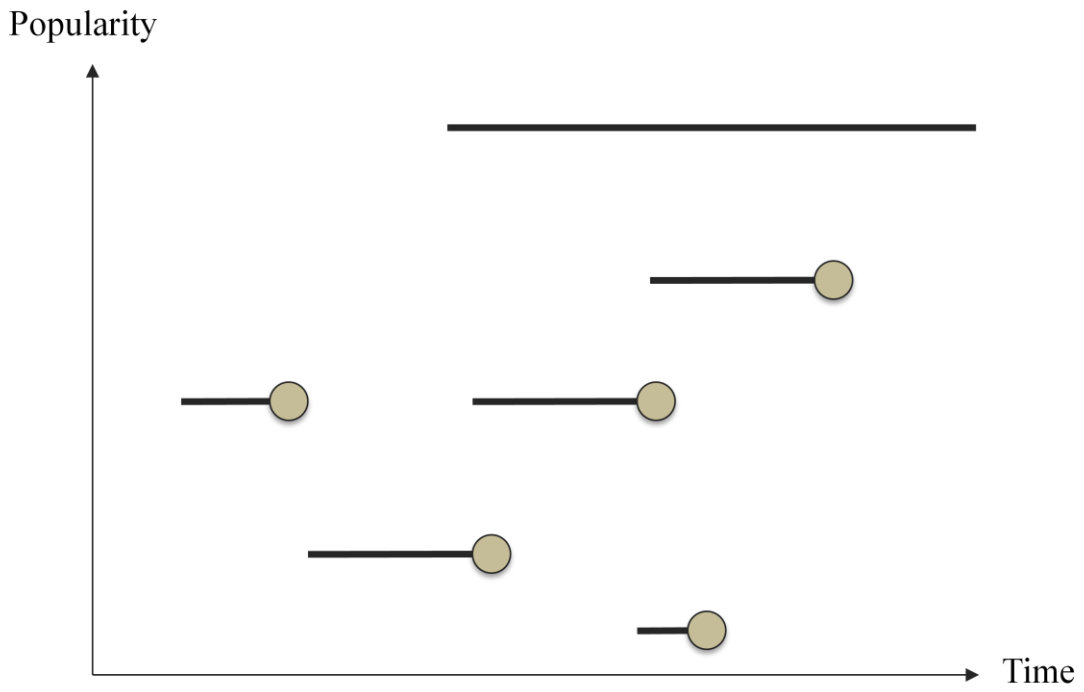

FIGURE 4. DIAGRAM OF OFFLINE SOCIAL NETWORK

Figure 4 is a graph that illustrates the behaviors of offline social network. Although it is also constituted of nodes and edges, their meanings are not the same. The nodes represent message which are released in the offline social network. The edges represent the data collection period of the messages. The position of the node implied how many individuals are involved in the data collections. In the graph, there is an edge without nodes, and it implies the social messages are still in data collection stage and it is not yet released.

Let R_q be the contents collection from the offline social network q, where $R_q = \{r_{q1}, r_{q2}, \ldots r_{qn}\}$, where n is the total number of messages in the network and n >= 1. Four attributes are extracted from every content r_{qi} where $r_{qi} = \{\varphi_{qi}, \tau_{qi}, \eta_{qi}, \mu_{qi,} p_{qi}\}$. The first attribute is an identifier φ, which can uniquely identify the content; the second attribute is a time-span (τ) which indicates the data collection period of the contents by recording the start time (τ_s) and end time (τ_e). For some contents which are directly

published and no explicit data collection time is provided, their time span will be as short as an instant. In other words, the start time (τ_s) will be equal to the end time (τ_e); the third attribute is a summary sentence (η) which briefly describes the contents. It can be a headline of news article, a survey question in polling or a topic sentence of any documents. The forth attribute is the free-format text-information carried by the contents. It can be a decimal number which indicates the responses of a survey question, or it can be an article collected from news media or social network sites. The last attribute is a population (p). It represents the number of individuals represented by the contents. If the content is polling or a co-signed announcement, the number will be equal to the number of participants. If the content is an individual article, the number will be equal to one.

The GNDM unifies the contents and structures of various online and offline social network hence one type of analytical approach is applicable in multiple networks.

3.5 Generic Network Data Linking (GNDL)

The GNDMs presented in the previous subchapter generalizes the contents collected from heterogeneous online and offline social media. However, the above process only conducted format conversion and it not yet attempted to reduce the size of collected data. Therefore, data selection has to be executed in every analysis and such action could be costly. The GNDL to be presented in this subchapter attempts to connect all offline social network contents to a collection of online social network messages, where they are candidates to be associated. In other words, it filters out the online social network messages which are irrelevant to the provided offline social network content hence the following data association process becomes scalable.

The procedure employs the time and entity information as the selection criteria and it can be divided into two phases. They are time-based linking and entity-based linking. In the time-based linking, a pair of time intervals is derived from every offline social network content. And only the online social network messages which are posted within the time intervals are connected to the offline social network contents hence can be preceded to the next phase. If the offline social network content is extracted from polling, the time-span, which indicates the data collection time period, will be adopted as the time interval. Otherwise the time interval will be set to ensure the date range is within the same calendar day.

$$h\,(M_k, R_{q,} \Delta t_{0,} \Delta t_1) = \{(\, r_{q1}, M_{k1}'), (r_{q2}, M_{k2}') \ldots (r_{qn}, M_{kn}')\} \qquad (1)$$

In (1), the linking procedure h matches the contents in offline social network (Rq) with the message obtained from the online social network k (M_k). The matching is controlled by the time differences Δt_0 and Δt_1. Each content in online social network (r_{qi}) will be connected to a message collection M_k', where the messages are being posted after (τ_{qi} - Δt_0) and before (τ_{qi} + $\Delta t_{1)}$. The matching conducted in (1) adopts time as the only criterion to link the contents in offline social media with the messages in online social media. Therefore, it is possible that the same message in an online social network is matched to the contents from disparate offline social network. If the association analysis is related to the behaviors about certain entities, the results of the time-base linking will be proceed to the entity-based linking.

In the entity-based linking, all entities interested in the analysis are extracted. The aliases of each entity are deduced and an indexing table, which maps aliases with their corresponding entities, is maintained. The message content of every online message which is chosen by the time-based selection will then be scanned with the use of the

indexing table and each message can be deduced and assigned to corresponding message-group.

Each subset of messages is associated with a unique combination of offline social network content and entity information to from a message-group. Similar to the time-base linking, since an individual message can describe multiple entities, hence a message can be present in more than one message-group.

$$H\,(E, h\,(M_k, R_{q,}\,\Delta t_{0,}\,\Delta t_1)) = \{(\,r_{q1}, e_1, M_{k11}'), (r_{q1}, e_2, M_{k12}')\ldots(r_{qn}, e_s, M_{kns}')\,\} \qquad (2a)$$

$$M_k{}' = \begin{bmatrix} M_{k11}{}' & \cdots & M_{k1s}{}' \\ \vdots & \ddots & \vdots \\ M_{kn1}{}' & \cdots & M_{kns}{}' \end{bmatrix} \qquad (2b)$$

The entity linker H applies entity matching on every time-base linking procedure h, which is generally designed for all offline social networks. It also aggregates the results of the executed linking procedures so that all entities in every offline social network will be associated to the corresponding group of messages unless the entity was never mentioned. The data entity can be expressed as $H\,(M_k, R_q) = M_k'$. The two inputs are messages in online social network k, which is represented as M_k, and offline social network q, which is represented as R_q. The output is $M_k' = \{(r_1, e_1, M_{k11}), (r_1, e_2, M_{k12}) \ldots (r_1, e_s, M_{k1s}), (r_2, e_1, M_{k21}) \ldots (r_k, e_s, M_{kns})\}$. It associates the online social media messages to the offline social network and entities. The M' can also be presented as a n × s two-dimensional matrix, n is the total number of contents in the offline social network relevant to the entity, where the message can be duplicated; s is the total number of interested entities involved in the offline social network and elements in the matrix are the messages corresponding to the combination of the network and entity.

The scalability of the above matrix can be regarded as O(n). Comparing these two dimensions, the number of entities being studied should be far fewer than the number of contents in social network. The number of entities will then be insignificant when calculating the complexity of the matrix. As a result, the complexity of the matrix O(n × s) can be assumed as O(n).

As expressed in (2b), element M_{kns} in matrix M_k' is the collection of online messages which is relevant to entity s in the n^{th} individual content. When a combination of entity and offline social network content is provided, a unique message-group, where its messages are relevant to such combination, can be selected. Such message-group can be applied by other analyses which regard on the same entity and the online and offline social network pair.

By using the time-based matching, the CL Agent in the D-Miner Service framework is capable of linking the contents in real and virtual communities which are released within a certain period of time. This is conducted through associating the text-contents in the two communities, which are news articles from real communities and SNS messages from virtual communities. They are collected by corresponding collection agents in the framework and transferred to the CL Agent. The CL Agent first duplicates the collected news articles and converts the copy into inverse document index for later usage. Then it adopts the novel **N**et**w**orks **C**ontent **L**inkage (NWCL) algorithm to conducts the associations. The NWCL algorithm takes two procedures to complete an association: (1) forming a query from a news article; (2) comparing the similarity among the article and the corresponding online social network contents via the query.

In many studies, various features in a news article like its title, first sentence and entire passage have been used to form queries. Such kind of queries represents the article in information retrieval or content analysis. The NWCL algorithm selects the titles of new articles to form queries because they represent critical information of news articles' contents in minimum length. It is observed that the structure of news articles' titles is often repeated and some words are frequently appeared across titles about different incidents. Those words contain little information to distinguish the titles hence they are chosen by the NWCL algorithm as one of the two indicators to split a title of a news article into two or more fragments. Each fragment is composed of two or more words; the fragments, regardless of their meanings or correctness in grammar, can be used to form a query. As those frequently appeared words are used to divide the news titles, they are named as delimited words and such term will be used in the rest of this paper. The performance of a delimited word list is measured by its segmentation rate, which is the percentage of news articles' titles which contained at least one delimited word in the list (i.e.: the percentage of titles which can be segmented by the list).

To form a delimited word list, we identified all distinct words from the collection of news articles' titles. For each word, its accumulated frequency of appearance across titles, which is named as word-frequency in the rest of this thesis, is calculated. Given a collection of news articles' titles, a threshold of segmentation rate and a table which contains all distinct words with their corresponding word-frequencies, Figure 5 illustrates the first part of the NWCL algorithm about the formation of delimited word list. For each word c in table H, if its character-frequency is greater than the word-frequency threshold k, it will be added to delimited word list L. After all words in table H are examined, if the segmentation rate of L cannot reach the threshold R, L

will be discarded and the comparison between c and H will be started over with a new k, where its value will be stepped down by one. Theoretically, the word-frequency threshold k has to initialize as the maximum word-frequency among the collection of words. This setting ensures the delimited word list to be generated will have minimum number of words. In practice, k is not started from the highest word-frequency but the plus one standard deviation is selected. This arrangement aims to improve the efficiency in building the optimum L and it is proven that delimited word list with minimum number of words can still be obtained.

NWCL Procedure 1a: Preparation of delimited words

Input: H (table of words and their corresponding word-frequencies), A (article title list),
R (threshold of segmentation rate)

Output: L (delimited word list)

Definition: *max_frequency* (H) - the highest frequency in word-frequency table H
seg_rate (L, A) - the segmentation rate of delimited word list L for article title list A
get_frequency (*c, H*) - the word-frequency of word c in table H

```
k ← max_frequency (H), L ← Ø
while seg_rate (L, A) < R do
    L ← Ø
    for c in H do
        if get_frequency (c, H) >= k then
        L ← c
        end if
    end for
    k ← (k-1)
end while
```

FIGURE 5. NWCL ALGORITHM, PROCEDURE 1A

Stop word is a widely-adopted indicator for splitting titles. Different from delimited words, stop words are well-defined and frequently appearing in many kinds of web documents. Such difference indicates that stop words are negligible in article-message association while delimited words are not. Therefore, two segmentation approaches, which named as inclusive and exclusive segmentation, are

designed for the delimited words and stop words, respectively. The differences between inclusive and exclusive segmentation are illustrated in the following example. Supposed a title of news article has n characters and is expressed as [1, 2, … k, k+1, … n], where each index represents a word. If the inclusive segmentation is conducted on the k^{th} character, both segmented fragments will contain the k^{th} character, and they are represented as [1, 2, … k] and [k, k+1, … n]. If the exclusive segmentation is conducted on the k^{th} character, the k^{th} character will not exist in both fragments and they are represented as [1, 2, … k) and (k, k+1, … n]. Figure 6 illustrates the segmentation process of a news title in the NWCL algorithm. The corresponding segmentation method is executed for every word *w* in a title when *w* is a stop word, delimited word or the last word (to prevent missing queries) of the title. If *w* is a stop word and delimited word, it will be processed as stop word.

NWCL Procedure 1b: Segmentation of news title

Input: T (title of a news article), S(stop word list), D (delimited word list)

Output: Q (query term list)

Definition: *isStopWord* (*w, S*) - identify whether word w is a stop word defined in S

isDelimitedWord(*w, D*) - identify whether word w is a delimited word defined in D

isLastWord (*w*, T) - identify whether word w is the last word in T

ex_segment (T, *w*) - exclusive segmentation of T by word w

in_segment (T, *w*) - inclusive segmentation of T by word w

```
for w in T do
    if isStopWord (w, S) then
        Q ← ex_segment (T, w)
    elseif isDelimitedWord(w, D) then
        Q ← in_segment (T, w)
    elseif isLastWord (w, T) then
        Q ← in_segment (T, w)
    end if
end for
```

FIGURE 6. NWCL ALGORITHM, PROCEDURE 1B

Since the number of news articles is continuously growing, the distribution of delimited words is likely to change over time. The NWCL algorithm suggests the delimited word list should be dynamically generated whenever it is needed. In our implementation, the list is updated once a week in the D-Miner Service Framework. The selection is based on the dataset characteristics which will be demonstrated in chapter 4.1.1.

The NWCL algorithm applies information retrieval techniques to compare the similarity levels between a query segmented from title of a news article and SNS messages, which has been converted into inverse document indices by the CL Agent. The comparison has two concerns: selection of weighting approach and similarity metric. Inverse Document Frequency (IDF) and Term Frequency–Inverse Document Frequency (TF-IDF) are two classic weighting approaches in information retrieval [9, 23]. Since the objective of NWCL algorithm is to select SNS messages which discuss similar topic as the news article, multiple matches of the same query term will not increase the relevance. Therefore, IDF is adopted as the weighting approach in the algorithm. Both inner-product and cosine similarity metrics have been used by different scholars to calculate the similarities between news articles and SNS messages. After referencing their performance and comparing it with our work, we selected inner-product as the similarity metric in NWCL algorithm.

NWCL Procedure 2: Query-Message Comparison

Input: Q (query term list), D (inverse document indices), T (time span of the association)

Output: H (table of sns messages and their corresponding accumulated similarity levels)

Definition: *get_doc_indices* (D, w) - get document indices of word w in D

get_idf (D, w) - IDF of word w in D

accu_idf (v, f, T) - update sns message v and its accumulated similarity level f to in T

desc_similarity_sort (T) – arrange T in descending order of accumulated similarity level

```
f ← 0, m ← 0, s ← Ø
for w in Q do
    f ← get_idf (D, w)
    s ← get_doc_ indices (D, w, T)
    for v in s do
        H ← accu_idf (v, f, H)
    end for
end for
H ← desc_similarity_sort (H)
```

FIGURE 7. NWCL ALGORITHM, PROCEDURE 2

Figure 7 illustrates the comparison process between a query, which is segmented from a news article, and the inverse document index formed by the CL Agent. Time and contents are the two criteria in the comparison. Nevertheless, the NWCL algorithm does not require time information of the queries (news article). This is because the host of the algorithm, which is the CL Agent, is invoked daily. Therefore, the queries are usually formed by recent new articles which are published within 24 hours. Instead, time span T is deployed to specify a time range before the analysis, and only the SNS messages posted within the range are being processed by the algorithm. At the beginning, the IDF for every query term *w* in the query was calculated. Then the list of document index is computed. Every index refers to a unique SNS message, where it contains query term *w* and is posted within time span T. The indices in the list and their corresponding IDFs will be added to table H. If the message has been

already existed in H, the IDF value will be accumulated. After all query terms have been examined, H will have the SNS messages which are similar to the query. The elements in H will be sorted in a descending order according to their similarity levels (i.e.: IDF values) hence the first document is most similar with the news article represented by the query. Depending on situation, the list may be sorted in chronological order but not levels of similarity to indicate the first SNS message that discusses the same topic as the news article.

3.6 Generic Network Data Association (GNDA)

$$S_t \approx P_t \tag{3}$$

In (3), both S and P represent the individuals' behavioral patterns aggregated from an online and offline social networks pair at time t. It is assumed that the pattern S and P are associated hence P can reflect S. The equation can only have one online and offline social network pair but their causal relationships are not restricted. In other words, the same equation can be used to project the behaviors in online social networks which induced by those in offline social network, and vice versa.

$$P_t = \sum_i w_i F_{Ti}(t) + \sum_j w_j F_{Cj}(t) + \varepsilon \tag{4}$$

The behavioral pattern P_t is represented by a linear regression model as shown in (4). When deciding the mathematical model of P_t, the compatibility of the model was the primarily concern. This was because the behavioral pattern P_t aimed to be applied in social networks which are not being analyzed before. Linear regression model is a traditional model and it has been successfully adopted in various types of social network analysis. It is expected that when such model is applied in new type(s) of data set, less effort will be required to refining the model.

The regression model can be decomposed into two major components. The F_T represents the existing analytical components. They are transplanted from other analyzes and their performances were already proven in other social network(s). The F_C represents the analytical components deduced from our work. The F_C is customized for particular type(s) of social analysis and aim to support or enhance the performances of F_T.

The proposed association model will be applied in online social network which are seldom or never being analyzed. In many cases, there are no existing association results to be compared with those deduced by the proposed model. Nevertheless, since the performance of F_T has been proven in other social network, such performances can be referenced in deducing the benchmark of the association model.The remaining parts of this subchapter will introduce the transplanted and customized aggregation models one by one. Since the model can be applied in both online and offline social networks, the term text-datum (plural: text-data ($\mathcal{M}$)), will be used to replace the term message in online social network (M) and content in offline social network (R). In other words, $\mathcal{M}_k = M_k$ when referring to online social networks and $\mathcal{M}_q = R_q$ when referring to offline social networks.

3.6.1 Transplanted Aggregation Model

$$\mathrm{PNSR_i} = \frac{|\mathrm{pos}(\mathcal{M}_i)|}{|\mathrm{neg}(\mathcal{M}_i)|} \tag{5a}$$

The Positive Negative Sentiment Ratio (PNSR) summarizes the sentiment in a social network. Sentiments are extracted from each text-datum in the network, where it is selected by the linking algorithm as described in the previous chapters. The text-data will then be classified as containing positive, negative, neutral or both positive and negative sentiments. As shown in (5a), the PNSR is calculated by dividing the

number of text-data with positive sentiments by the number of text-data with negative sentiments. The PSNR will be greater than one if the number of text-data which contain positive sentiments is more than those which contain negative sentiments and vice versa. The PSNR has no upper bounds. And in extreme scenario, when the online social media has no negative text-data regarding the entity, the calculation of PSNR will have problem of "divided by zero". The scenario can be further decomposed to two cases. In case one, the number of text-data with positive sentiment is also equal to zero; while in case two, the number of text-data with positive sentiment is equal to a positive integer. The first case indicates that all people in the online social media are either having neutral attitudes or ignoring the entity. The PSNR will become meaningless and it will be set to zero. In case two, it is suggested that the PSNR should be replaced by the Positive Sentiment Ratio (PSR) in (5b).

$$\mathrm{PSR_i} = \frac{|\mathrm{pos}(\mathcal{M}_\mathrm{i})|}{|\mathrm{pos}(\mathcal{M}_\mathrm{i}) \cup \mathrm{neg}(\mathcal{M}_\mathrm{i})|} \tag{5b}$$

Similar to the PSNR, the Positive Sentiment Ratio (PSR) also measures the weighting of positive sentiment in the social network. It is calculated by dividing the number of positive text-data by the union of positive and negative text-data. The PSR will never have the "divided by zero" problem as presented in the first case, but it is less preferred. This is because the PSR is bounded by zero and one and it is less sensitive to the differences of positive and negative sentiments. Since the two ratios describe similar issues, they are substitute to each other and will not appear together in one equation.

$$\mathrm{EA_i} = |\mathcal{M}_\mathrm{i}| \tag{6}$$

The Entity Attention ($\mathrm{EA_i}$) calculates the total number of messages which mentioned or discussed the entity i. Different from the PNSR, which only considers the message

with positive or negative sentiments, the EA also includes those messages with neutral sentiments on the entity. However, those messages which contain more than one kind of sentiments will only be counted once in the calculation of EA. Therefore, the value of EA_i is inclusively bounded by zero and N, where N is equal to the total number of text data in the social network. This is because in the extreme scenario, all messages in the network are describing the entity i.

$$EP_i = \frac{EA_i}{\Delta t} \tag{7}$$

Different from the EA, the Entity Popularity (EP) in (7) measures how frequent the given entity e is being mentioned in the social media within a certain period of time Δt. The calculation of EP is to divide the total number of messages related to the entity in the social network (the entity attentions) by the predefined duration. As mentioned in the chapter 3.3, the time interval will be equal to the data collection period for polling or voting, otherwise, it will be set to ensure the date range is within the same calendar day.

3.6.2 Customized Aggregation Model

In most correlation analysis such as the estimation of the movie box revenue or stock market performance, the number of multiple messages posted from the same user in the social network is not modeled. This is because similar behaviors such as purchasing multiple tickets of one movie or holding multiple units of stocks of a company could also happen. This Author-Message Ratio (AMR) is specially designed for analyzing those social networks which only allows individuals to share their opinions once hence reflecting the differences between the two behavioral patterns.

$$\mathrm{AMR_i} = \frac{\mathrm{EA_i}}{\mathrm{disA}(\mathcal{M}_\mathrm{i})} \tag{8}$$

As presented in (8), the calculation of AMR is to divide the total number of messages related to the entity in poll by the distinct number of authors who shared those messages.

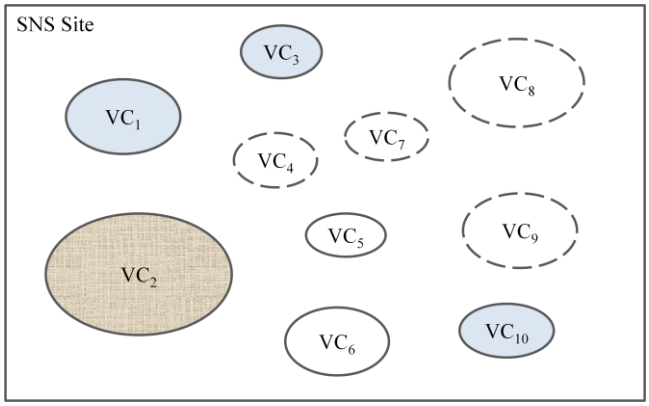

FIGURE 8. EXTERNAL INFLUENCES IN A SOCIAL NETWORK

The absolute social influences can reflect the contrasts of influences linking with individuals while relative social influences normalize the variations of disparate social environments. Both of them are estimated in our work for evaluating and tracing the social influences of news media across virtual communities in online social networks. Figure 8 illustrates an example where two text-data from the same external source are shared in one social network. It is suggested that only four communities are formed by the individuals in the network for simplicity. Every eclipse in Figure 8 represents a virtual community. Those having solid boundaries and shaded are affected by the text-data, where $\{VC_1, VC_3, VC_{10}\}$ and $\{VC_2\}$ are affected by different text-datum, source A and B, for example. Also, the eclipses have solid boundaries but not shaded are the communities which do not induces by the text-data but also concerning the same issue mentioned by those text-data. In other words, they are self-initiated or influenced by unspecified sources. The eclipses with dotted boundaries are the

communities neither affected by the articles nor discussing the same issues. To better present the influences in one social network, the attention drawn by entity i (EA_i) mentioned in (6) is applied and further customized.

$$EA_i = EA_i(\lambda) + EA_i(\gamma) \quad (9)$$

As mentioned above, the EA_i is the number of messages which mentioned or discussed the entity i. Therefore, the EA_i equals to the number of messages in $\{VC_1, VC_2, VC_3, VC_6, VC_{10}\}$ as presented in Figure 8. The determination of the $EA_i(\lambda)$ is dependent on which external source is being studied. In the example in Figure 8, there are two external sources A and B. Therefore, the $EA_i(\lambda)$ can be the number of messages in $\{VC_1, VC_3, VC_{10}\}$ (source A) or the number messages in $\{VC_2\}$ (source B). The $EA_i(\gamma)$ can be determined once the $EA_i(\lambda)$ is decided as $EA_i(\gamma) = EA_i - EA_i(\lambda)$.

The influential measurements are divided into 3 levels according to their scopes. The primary level only focuses on the effect of single article, hence either $\{VC_1, VC_3, VC_{10}\}$ or $\{VC_2\}$ as presented in Figure 8 is evaluated in one time. The secondary level measures the influences of multiple news articles by combining the results obtained from the primary-level measurements. According to the selection of news articles, the results could represent the influences of the medium or its influences on particular topic(s). The tertiary level applies multiple secondary-level measurements where they are about disparate news media. Collaboratively, their results are integrated for deducing the resultant social influences of multiple news media in online social networks. The absolute, relative and combined measurements to be introduced below are capable of extending to those three levels.

The absolute social influences of one influence source j about the entity i can be expressed as EA_{ij} in (10a).

$$EA_{ij}(\lambda) = \sum_{j=0}^{n} w_{ij} m_{ij} \tag{10a}$$

$$EA_{ij}(\lambda)' = \frac{EA_{ij}(\lambda)}{EA_i} \tag{10b}$$

In most cases, the value of $EA_{ij}(\lambda)'$ is inclusively bounded by zero and one. However, the calculation will not be capable of computing meaningful information when EA_i is equal to zero. In such scenario, the influencing text datum was not shared in the online social network and the issues mentioned by the text datum were not being discussed by any individuals in the social network as well. The calculation of the social influences will become meaningless because both the text datum and the issues carried by that text datum are negligible to the social media.

As mentioned in previous chapters, the information from external data sources was propagating across social network by an overlapped group of individuals. The text-data induced by the text datum from another social network are considered as influenced by that social network as a whole. Conversely, the social influences of any social network can be estimated by integrating the social influences of all of its induced text data. This concept can be further extended to various applications. For example, we can deduce the social influences of a news media in England soccer by only integrating the social influences of all news articles in the media, which are related to England soccer. However, the integration does not guarantee the integrity of the results. It is because the mechanism only assembles the social influences of a

given set of news articles to form an overall influence. If the selected articles have flaws in data integrity, the integration results will be unreliable or even meaningless.

Suppose there is an arbitrary group of text data collection and it is expressed as $\mathcal{M}$ = [m_1, m_2, …α_n]. The absolute social influences of such collection are defined as $EA_i(\lambda)$ in (11a).

$$EA_i(\lambda) = \sum_{j=0}^{n} EA_{ij}(\lambda) \tag{11a}$$

The function sums up the absolute social influences of every individual news article and the accumulated value represents the social influences of such news collection. The value is inclusively bounded by zero and γN, where N is equal to the total number of messages in the online social network and γ is the number of news articles in the collection. On the other hand, the relative social influences of the collection are defined as $EA_i(\lambda)'$ in (11b).

$$EA_i(\lambda)' = \frac{1}{n}\sum_{j=0}^{n} EA_{ij}(\lambda) \tag{11b}$$

Instead of simple accumulation, the function calculates the average relative social influences of all news articles in the collection. Its value is inclusively bounded by zero and one. The measurements (11a) and (11b) extend the social influences measures from individual text data to an arbitrary group of text data in the same social network, where the group can be altered to fit various analytical purposes. It is expected that the differences of social influences between text data will be further increased when data across news media are involved in the calculations.

Note that the two measurements in (11) are customized for different focuses and they cannot resolve the differences independently. It is suggested to have a new measurement where it is named as combined measurement, and it can consider the magnitude and coverage of social influences. The new measurement is expressed as WEA in (12a).

$$\mathrm{WEA_i} = \omega_{ij} \sum_{j=0}^{n} \mathrm{EA_{ij}}' \tag{12a}$$

$$\omega_{ij} = \frac{\mathrm{EA_{ij}}}{\mathrm{EA_i}\ (\lambda)} \tag{12b}$$

Similar to (11b), the calculation summated the relative social influences of all news articles. Nevertheless, a weight is dedicated to the influences of every text datum. The weight for a text datum is presented in as ω_i in (12b). It is calculated by dividing the absolute social influences of the article, which is defined in (10a), by the overall absolute social influences of the articles collection as defined in (11a). Therefore, the values of ωi and $\mathrm{WEA_i}$ are inclusively bounded by zero and one.

In order to effectively trace the changes in social influences of a news media, it is necessary to consider the behaviors of the external text data. This is because the contents of news articles are invoked by the incidents in our society and guided by the news media, so that they are unpredictable in most of the time. As a result, the social influences sometimes violently fluctuated in a few days and this induced noises in the influences measures. Our work attempts to form similar text data into groups and compares the social influences between those groups. It is observed that the text data articles, where they are published in disparate epoch, are distributed along with the time. The distribution of the text data is one-dimensional so that it can be clustered

by classic approaches likes K-means algorithm according to their sharing times. Suppose R text data are formed into m clusters. In order to compare the clusters formed, our work deduces two attributes from every cluster. The first attribute is the population. It indicates how many news articles are involved in a cluster and its value is inclusively bounded by one and R-(m-1). The second attribute is the population density. It represents on average, how many news articles will be appeared in a designated period of time. Since news articles are published daily in convention, the population density measures the average number of news articles in a day. On the other hand, since the overall social influences can be formed by arbitrary text data collection, the overall social influences of the clusters can be deduced by using (11) and (12) introduced above. Through comparing the social influences of the clusters and the time intervals among the clusters, the trends of social influences of the targeted social network can be traced.

Despite the small variances in definition, the positive negative sentiment ratio and entity popularity are often employed by other correlation models. For example, the analysis about revenue of movie box or the further direction of stock market index. A part of our work adopted the models (with essential modifications) to analyze the results of consecutive polling sessions of a tracking poll. It is shown that the model successfully correlated the results of polling sessions and aggregated data from online social media. The proposed AMR extended the correlation model and outperformed previous models. Although the improvement was not significant, it demonstrated a promising direction to extend the correlation model and adopt the differences of online social media.

Chapter 4: Experiments

4.1 Evaluating the NWCL Algorithm

This chapter describes the experiments conducted on the CL Agent in the service framework. The experiments evaluated the NWCL algorithm about its performance in associating the news articles with the SNS messages. Also, they demonstrated how the algorithm traces the information propagations and how it inspires the correlation analysis to be conducted among real and virtual communities. The experimental details to be presented in this chapter are constituted of five components. First, it is started with describing the real and virtual communities studied in the experiments. Then, it evaluates whether the assumptions of the NWCL algorithm are valid in the selected virtual community, so the community's contents can be used in the experiments. Next, it presents the formation of delimited words from the contents of virtual communities. After that, it describes how experiments are conducted. Last, it discusses the results of the experiments.

4.1.1 Dataset Characteristics

Two datasets, which are news articles from real communities and SNS messages from virtual communities, were collected by the corresponding agents in the proposed framework before the experiment. The news articles were written in traditional Chinese. They were "local news" in Hong Kong and collected from website of Mingpao[1], which is a newspaper in Hong Kong. The SNS messages were public dialogs obtained from Uwants[3], which is a popular discussion forum in Hong Kong. From 9th July, 2012 to 30th July, 2012 (inclusive), 569 pieces of news articles were collected. Their titles, leading sentence and the Uniform Resource Identifiers (URIs) were extracted. In the collection, the titles had 6306 characters together and there

were 1276 unique characters. Among the 1276 characters, 1256 of them were Chinese characters while the other 20 were whitespaces, symbols, numbers or English letters. During the period of 9th July, 2012 to 06th August, 2012 (inclusive), 12690 SNS messages from 998 distinct threads were obtained from the "Current Affairs Forum" in Uwants. The information in all collected messages included forum information and author identification, collection time of the message, message-titles and contents. The number of messages seems not significant when compared with related studies. As the selected virtual community does not have global but only regional scope of population, and such community is one of the most popular forums in Hong Kong, The collected messages reflect a typical behavioral pattern of a regional virtual community.

4.1.2 Pre-Experiment Evaluations

The query formation mechanism proposed in the NWCL algorithm was based on two assumptions. (1) Some characters extracted from the titles of news articles appeared more frequently; (2) the delimited words defined in chapter 3.5 are capable of segmenting most or all titles of news articles into queries. In order to testify for the validity of assumption (1), we studied the words in the titles of news articles. Their word-frequencies ranged between of 1 to 83 (inclusive) as shown in Figure 9.

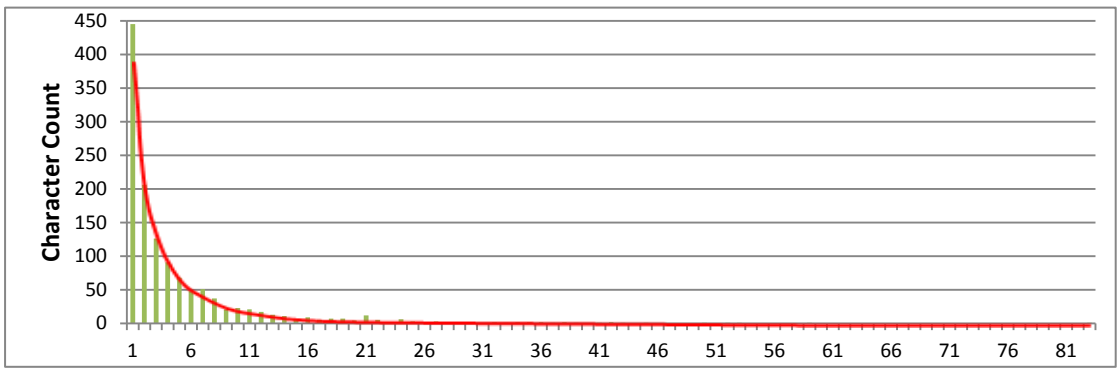

FIGURE 9. HISTOGRAM OF WORD FREQUENCIES

Range	Mode	Median	Mean	Standard Deviation
82	1	2	5.1458	7.3178

TABLE 2. STATISTICS OF THE WORD-FREQUENCY DISTRIBUTION

The trend of the distribution was exponentially decreasing, indicating that when the word-frequency increased, the character count was dropped significantly. The histogram and statistics presented in data characteristic section could not fully support the first assumption because it only showed the experimental data followed the assumption while the assumption may be invalid when the total number of distinct characters continues to increase. In order to further investigate the trend, we derived three sets of articles from the collection of news articles and labeled them as S_A, S_B and S_C accordingly. Among the three sets of articles, S_A was collected from the first 10 days; S_B was collected from the first 16 days and S_C was collected from the first 22 days, which is also the entire set of data. Therefore, $S_A \subset S_B \subset S_C = U$, where U is the complete collection. Moreover, the three sets of articles (S_A, S_B and S_C) represent the data of the universe (U) in different time frame.

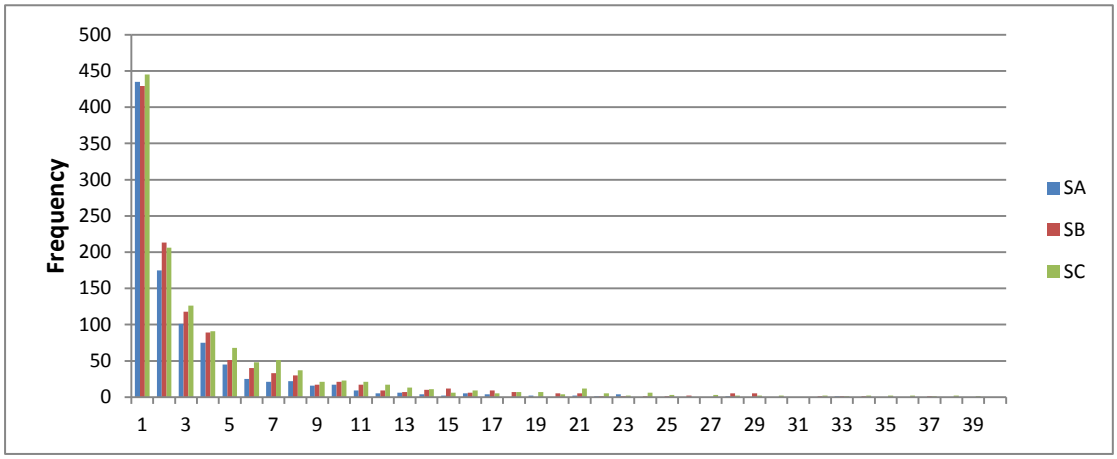

FIGURE 10. WORD FREQUENCIES DISTRIBUTIONS FOR 3 SETS OF ARTICLES

By comparing the differences of distributions presented in Figure 10, we can study the changes of word-frequency distribution in the collected news articles along the time.

It is observed that S_A has a positive skew distribution as well as an exponentially decreasing trend, while S_B and S_C also have similar patterns. Therefore, even for different time frames, the first assumption is valid.

When the standard deviation was continuously increased, the distribution of word-frequency became more dispersed as shown in Table 3. Provided that mode and median remained unchanged and had small values, when the number of distinct characters increased, few characters move to right (i.e. has higher word-frequency).

	S_A	S_B	S_C
Total Characters	3168	4885	6582
Distinct Characters	979	1151	1276
Range	41	55	82
Mean	3.2360	4.2441	5.1458
Mode	1	1	1
Median	2	2	2
Standard Deviation	3.8881	5.5479	7.3178

TABLE 3. STATISTICS OF WORD-FREQUENCY DISTRIBUTIONS FOR THE THREE DATASETS

In order to verify the validity of assumption (2), we formed a sample of delimited word list and evaluated its performance by calculating the segmentation rate. The list was formed by those characters which have word-frequency greater than one standard deviation (12.46). In the ideal case (normal distribution), the size of the delimited word list should around 15% of the total number of distinct characters. In our experimental data, the percentage was smaller (9.56%) as the distribution was positively skewed. Surprisingly, nearly all (99.33%) of the titles can be segmented with the short list and such behavior supports the second assumption.

4.1.3 Delimited Words Selection

This subchapter describes how the selection of delimited words mentioned in NWCL algorithm that presented in previous chapter is executed in the experiment. A list of delimited words was chosen to segment the titles of news articles. Its threshold of segmentation rate was set to 1.0, so all the news articles' titles have to be segmented by the list. If the performance of the selected list cannot pass the threshold, more characters will be introduced to the list to increase its segmentation rate until the threshold is reached. Table 4 presents the trials attempted for building the optimum delimited word list. The initial list was formed by characters with word-frequency greater than one standard deviation (12.46). The number of words in list was increased stepwise to obtain better segmentation rate. In the experiment, the process was stopped in the third trial.

Trial	Word-frequency	Word Count	Percentage of delimited words	Segmentation Rate
1	13 or above	122	9.56%	0.9933
2	12 or above	139	10.89%	0.9933
3	**11 or above**	**160**	**12.53%**	**1.0**

TABLE 4. TRIALS OF BUILDING DELIMITED LIST AND THEIR PERFORMANCE

The list of delimited words obtained in the third trial was considered as optimum; because it reached the preset performance threshold with minimum number of words. The list originally contained 161 characters, but one of them was identified as invalid and was removed. The invalid characters had such high word-frequency because multiple characters in the titles were not supported by the encoding method used in the framework hence they converted into the same invalid character. After reviewing the raw data, we confirmed there were no invalid characters satisfied the criteria to be included in the delimited word list.

A list of stop words was collected from the Web[4] and compared with the delimited word list formed in the experiment. The performance of delimited word list and stop word list are presented in Table 5.

	Delimited Word List	Stop Word List
Number of Characters	160	125
Percentage Overlapped	6.875%	8.8%
Segmentation Rate	1.0	0.6077

TABLE 5. STATISTICS OF WORD-FREQUENCY DISTRIBUTIONS FOR THE THREE DATASETS

The statistics in table 5 illustrate that the delimited words and stop words are of different sets of words as it is shown in table their overlapping rates were less than 10%. According to the experimental results, only 60.77% of the news articles (361) contained stop words, while 100% of the news articles (594) contained delimited words. As presented in table 4, it shows the delimited word list with 122 words already capable of obtaining more than 0.99 segmentation rates when compared with the 0.6077 contributed by 125 stop words. Therefore it is reasonable to conclude that in the experimental data, the delimited word list outperformed the stop word list in coverage of news titles.

4.1.4 Experimental Setting

There were two main objectives experimental objectives. One was to evaluate the performance of the NWCL algorithm, and the other was to trace back the associations identified by the algorithm. However, the tracing of associations could be meaningless if the news articles are not interrelated. For example, there is a sequence of news article and SNS message pairs and it observed that the number of pairs was decreased steadily. Nevertheless, it did not imply fewer people concerned the news or the media. This is because the articles were not from the same topic or category,

so that the differences could be contributed by the variances of the topics or categories. Therefore, it is necessary to define a theme to constraint the news articles involved in the experiment in order to ensure they are interrelated in contents. In our experiment, the promotion of "national education" in Hong Kong, which was a controversial social topic in both real and virtual communities for months, was selected as the theme of this experiment. The 569 archived news articles were reviewed manually and 29 of them, which were published from 9th July, 2012 to 31st July, 2012, were identified as relevant to the theme. On the other hand, because of limitation of time, a collection of SNS messages was also sampled from the 12690 archived messages and adopted in the experiment. The sampling process was loosely supervised in order to ensure the extracted samples will not be composed by the messages in one or few days. From 9th July, 2012 to 06th August, 2012, 252 threads, which were constituted of 2686 SNS messages, were sampled from the archive. Therefore, not all archived news articles and SNS messages described in Chapter 4.1 were involved in the experiment.

A few trials were conducted before the experiments. It was observed that if no threshold was applied, where all retrieved article and message pairs are determined as associated, the precisions of the retrieval will be very low. This is because the query generations from the news articles are not being supervised. The generated queries often have one query term which matches to a number of unrelated SNS messages with low similarity levels. The mismatched query terms are considered as noises in the experimental environment and are inevitable. If a similarity threshold is applied to remove the noise, the precision of the retrieval will rise significantly. However, if the threshold is set too high, the recall of the retrieval will drop sharply. After a sequence of evaluations, we concluded that the query term threshold, which contains the number of matched query terms in a query, is a more effective screening criterion when

compared with the similarity threshold. The experiment adopted the query term threshold. The retrieved article and message pairs are considered to be associated if and only if half or above query terms are matched.

4.1.5 Experimental Results

In order to evaluate the performance of the NWCL algorithm, which adopts both inclusive and exclusive segmentations, it is compared with the two segmentation techniques individually. Therefore, the 29 news articles were triplicated and each copy was converted to a set of queries by corresponding segmentation techniques. There were 87 queries in total.

	Exclusive Segmentation Only	Inclusive Segmentation Only	Both (Adopted by NWCL algorithm)
Number of Associated Messages	91	79	82
Precision	0.8681	0.9747	0.9390
Recall	0.9405	0.9167	0.9167
Number of Connections	265	478	1115
Segmentation Rate	0.7931	1.0	1.0

TABLE 6. PERFORMANCES OF THE THREE SEGMENTATION TECHNIQUES

As presented in table 6, the three segmentation techniques were evaluated by comparing their performances in associating the new articles with SNS messages. If the segmentation only relied on stop word (i.e.: the exclusive segmentation), 91 threads could be identified. The precision of the retrieval was 0.8681, which was the lowest among the others. Nevertheless, the recall of the results was 0.9405, which was the highest value in the three techniques. If the segmentation only applied delimited word (i.e.: the inclusive segmentation), 79 threads could be identified, which is the

fewest number among the other techniques. However, the retrieval had the highest precision, where its value was 0.9747, and the recall of the retrieval was 0.9167. If the segmentation applied both stop words and delimited words (i.e.: the NWCL algorithm), 82 threads could be identified. The results demonstrated that neither the precision nor the recall of the retrieval outperformed the other two techniques. It seemed that the NWCL algorithm performed worse than the other two segmentation techniques in the experiment. However, the NWCL identified the largest number of connections when compared with the other two techniques. Since the number of connections indicates the strength of the associations among news articles and SNS messages, the great differences in the number of connections inspired us to study the reliabilities of the three techniques. The reliability study evaluated the queries generated by the three techniques and identified the query association rate, which is the proportion of queries that is successfully associated with a SNS message.

	Exclusive Segmentation Only	Inclusive Segmentation Only	Both (Adopted by NWCL algorithm)
Query Association Rate	65.51%	96.55%	100%

TABLE 7. QUERY ASSOCIATION RATE OF THE THREE SEGMENTATION TECHNIQUES

As presented in table 7, the NWCL algorithm successfully associated all news articles to a message in the SNS hence all queries were occupied. Also, if only inclusive segmentation was applied to the news title, over 95% of the articles can be associated to a SNS message. Last, if exclusive segmentations were applied to the news titles, only around 65% of the articles can be associated to a SNS message. The two experiments presented above indicated that the queries formed by exclusive segmentation could associate with more distinct SNS messages when compared with

inclusive segmentation or NWCL algorithm. However, the number of associations discovered by the exclusive segment is the fewest and they are contributed by relatively small proportion of queries. In other words, the performance of associations is likely to be fluctuated when a new collection of news articles is used. This is because exclusive segmentation may not successfully to form queries from the articles which can be associated with a SNS message. On the other hand, although the NWCL algorithm did not achieve the highest precision or recall, it is capable of discovering the largest number of associations when compared with the other two segmentation techniques. The NWCL algorithm offers higher level of reliability so that it is more likely to convert a new headline to a query.

The tracing of the promotion of "national education" focused on whether the real communities, the news article, or the virtual communities, SNS messages, initiate the discussion. Although no correlation analysis has been conducted, it is believed that behavioral patterns in virtual communities led behavioral patterns in real communities. This is because the virtual communities were used as platforms to announce the information about the coming assemblies in the real communities. The CL Agent used the NWCL algorithm to associate the news articles with the SNS messages. Surprisingly, according to the results of the associations, the first incident of the "national education" in real communities was occurred on 9th July, 2012, which is one day before the occurrence of the first SNS discussion. A sharp raise of SNS discussions about the "national education" was recorded on 11th July, 2012 and the number of discussion was steadily increased afterward. On the other hand, more incidents were observed after the boom of the SNS discussions. In that period of time, two incidents were recorded in one day and the occurrence of incidents became more frequent. The case study demonstrated that the real communities initiated the

discussion, and the incidents in real communities and discussions in virtual communities are very likely to influence each other. If correlation analysis about such topic is conducted, additional data processing such as including the incidents in the correlation model or removing the discussion which are strongly influenced by the incident has to be executed according to the addition information provided by the CL Agent.

4.2 Evaluating the GNDA in Polling

4.2.1 Dataset Characteristics

In this experiment, a rolling poll conducted by Public Opinion Programme (POP), the University of Hong Kong[1] was adopted to represent the public opinions in offline social network (which is our society). The poll studied the public supporting rates of the three candidates in the Hong Kong Chief Executive Election 2012. The rolling poll was from 27th February, 2012 to 23rd March, 2012 and it was constituted by eleven individual polls where over 10,000 people were questioned via phone interviews in total. The poll assumed that the election would be held on tomorrow and asked interviewees to vote for a candidate. Interviewees may also select none of the three candidates or abstain so that there were five mutually exclusive options. The poll results aggregated the decision of the responded individuals and summarized the popularity (in percentage) of the five options. Those results are available in the POP website and are publicly accessible. On the other hand, the online social media messages involved in the experiments were collected from Uwants.com[2], a popular discussion forum in Hong Kong. It was selected because it has a predefined community for the Chief Executive Election 2012 which centralized the corresponding discussions. A web crawler was tailor-made to download the discussions from the

site and extract their information specified by the data models. The crawler was executed three-time daily.

Correlation Variables	Adjusted R^2	p-value
PSR	0.611565	0.040794
PSNR	0.691297	0.025068

TABLE 8. COMPARISON OF PSR AND PSNR

Correlation Variables	Adjusted R^2	p-value
PSNR	0.691297	0.025068
EP	0.001876	0.371894
AMR	0.041112	0.332315
EP + AMR	0.386895	0.223115
PSNR + EP	0.885998	0.017889
PSNR + AMR	**0.894140**	0.016008
PSNR + EP + AMR	0.885228	0.068067

TABLE 9. PERFORMANCE OF THE SEVEN CORRELATION MODELS

4.2.2 Experimental Results

The performances of positive negative sentiment ratio (PSNR) and Positive Sentiment Ratio (PSR) were compared and presented in Table I. It is observed that the PSNR has stronger correlation with the results in tracking polls and has a higher confident level. Therefore, PSNR was selected and formed the seven combinations with the EP and AMR as introduced in previous chapter. Those combinations were evaluated and the corresponding results are presented in Table II. It shows that the PSNR is correlated with the results in the rolling poll (R^2 = 0.6913). On the other hand, neither the entity popularity (EP) nor the author-message ratio (AMR) shows correlations with polling results (R^2 = 0.0019 and 0.0411 respectively). Even when

the EP and AMR were put in one correlation model, the correlation was still weaker than the one formed by PSNR only. Whereas the correlation models formed by EP and/or AMR do not have significant performances, the two aggregations performed well with the PSNR. The model formed by PSNR and AMR had the highest level of correlation (R^2 = 0.8941), while the model formed by PSNR and PE also had significant performance (R^2 = 0.8860). The experimental results indicated that although the EP and AMR are only weakly correlated with the results in rolling poll, they can significantly strengthen the correlation between PSNR and results in the poll. However, when one more variable (EP or AMR, depends on the original combination) is introduced to the correlation model, the correlation and confident level of the model were dropped.

4.2.3 Discussion

If a correlation model uses PSNR as the only variable, significant satisfactory correlation level can be obtained. However, the relation may not be strong enough to predict the future trends, and other variables have to be added. The correlation levels derived from EP and/or AMR are weaker than the one derived from PSNR. Nevertheless, the EP and AMR can be used to increase the correlation significance of PSNR. However, when all the three aggregated data are used in one correlation model, its performance and confident level will be dropped. This is because both EP and AMR describe the message distribution in the online social media but using different aspects, which are time and number of distinct authors. If both of them are present in the same correlation model, the importance of data distribution to the model will be overestimated and the correlation level will be affected. Although the correlation model formed by PSNR and AMR outperformed the one formed by PSNR and EP in the experiment, this does not imply that AMR is a better aggregation of data

in all time. This is because the data collection periods are consistent for all individual polling applied in the experiment. Therefore, EP had not been utilized in the experiment and this may affect its performance. The experimental results demonstrated that the combination of PNR and EP, which are transplanted from existing online social network analyses, are strongly associated with the polling results. However, when replacing the EP with the Author-Message Ratio (AMR), which is a customized analytical components deduced in this work, stronger association was deduced.

4.3 Evaluate the GNDA in News Media

This chapter presents the preliminary experiment that was conducted on March, 2013. Such experiment was worked on real data which were collected from a small social network site. It compared the performances of the social influences measurements, which were suggested in the previous chapter, by applying them and evaluating their results. This chapter will first introduce of the site and news media observed in the experiment. It will then by describe and discuss the results deduced from the experiments.

4.3.1 Data set characteristic

In the experiment, Ming Pao Newspaper Limited, a news paper publisher in Hong Kong was selected as the interested news medium. The latest news articles released in its website (http://inews.mingpao.com/rss/INews/gb.xml) were collected three times a day. From July 2012 to February 2013, 5965 pieces of local news were collected and archived. On the other hand, uwants.com (http://www.uwants.com), which is a popular local discussion forum in Hong Kong, is chosen as the interested social network site. The site predefines a number of communities according to the

discussion topics, and the communities of "Current Affairs and News Reports Discussion" were chosen. From July 2012 to February 2013, 167354 messages were collected in the community, and they were contributed by 7959 members.

4.3.2 Experiment Settings

The experiment focused on the social influences of the news medium about the "old age living allowance" issue, which was a recent controversial social topic. And 94 news articles, which reported such issue, were manually selected from the archived news articles collected by a customized web crawler. The reporting of such issue was lasted from 209 days from 17th July, 2012 to 10th February, 2013. Figure 11 presents a histogram of news articles with respect to the days, where the day 0 represents the 17th July, 2012. Also, 6 groups of news articles can be formed according to the time intervals between the publications of news articles. The detail descriptions of those 6 groups including the population and population density which were introduced in Chapter 3.3 are presented in Table 1.

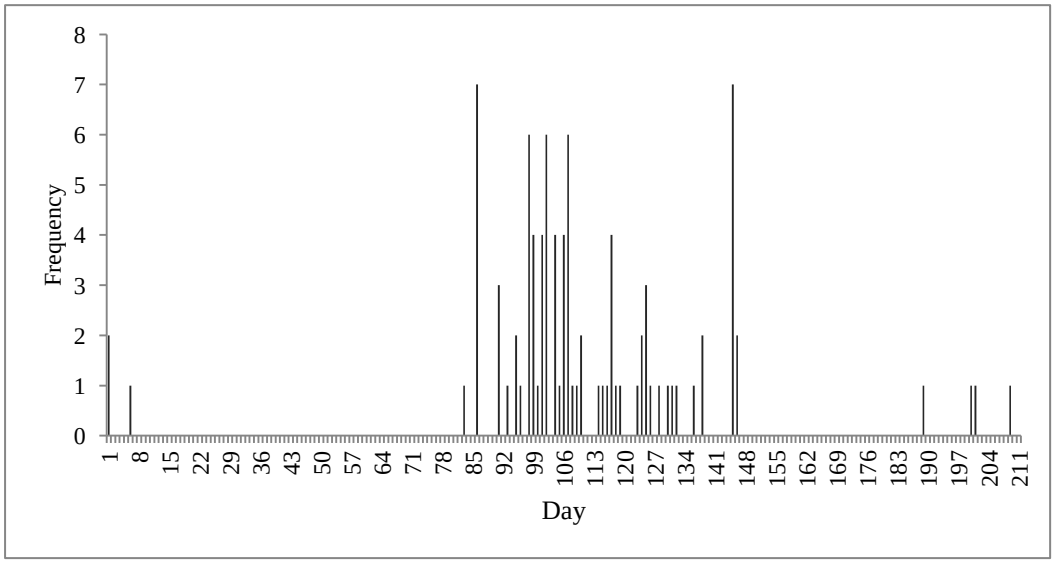

FIGURE 11. DISTRIBUTION OF THE "OLD AGE LIVING ALLOWANCE" NEWS ARTICLES

Article Group	Epoch	Num of Days	Population	Population Density
A1	2012-07-17 to 2012-07-22	2	3	0.6
A2	2012-10-07 to 2012-12-01	34	78	1.42
A3	2012-12-08 to 2012-12-09	2	9	4.5
A4	2013-01-21 to 2013-01-21	1	1	1.0
A5	2013-02-01 to 2013-02-02	2	2	1.0
A6	2013-02-10 to 2013-02-10	1	1	1.0

TABLE 10. CLUSTERS FORMED FROM DISTRIBUTION IN FIGURE 11

4.3.3 Experiment Results

Table 10 presents the social influences of the 6 groups of news articles. They were measured by the 3 social influences measurements which were introduced in Chapter 3. It is observed that if the absolute social influences (represented by the $EA_i(\lambda)$) are applied, A2 will have the greatest influences while A3 only has one-tenth influences as A2. And the influences of A1 will be the same as A4, A5 and A6. On the other hand, if the relative social influences (represented by the $EA_i(\lambda)'$) is applied, it successfully distinguished A1 from A4 to A6 and deduced the changes of influences: the news medium had no influential power at the beginning (the A1 group), and then the medium successfully influenced the online social network (from A2 to A3), finally the issued reported by the news was faded out hence the influences cannot be determined (from A4 to A6). However, the social influences of A2 and A3 were conflicted to those obtained by absolute social influences measurements. The data in

Table 11 indicated that the relative social influences of the two articles are almost the same while their absolute social influences have ten-time differences. The proposed measurement WEA_i adopted the advantages of absolute and relative social influences measurements. On one hand, it preserves the trends of changes of social influences of the news medium; on the other hand, it is capable of showing the differences of social influences between A2 and A3.

Article Group	**Number of Articles**	**Article Density**	**$EA_i(\lambda)$**	**$EA_i(\lambda)'$**	**WEA_i**
A1	3	0.6	0	0.0	0.0
A2	**78**	**1.42**	**40**	**0.97**	**0.88**
A3	**9**	**4.5**	**4**	**1.0**	**0.09**
A4	1	1.0	0	n/a	n/a
A5	2	1.0	0	n/a	n/a
A6	1	1.0	0	n/a	n/a

TABLE 11. THE SOCIAL INFLUENCES CALCULATED BY THE 3 APPROACHES

Chapter 5: Conclusion

Chapter 5.1 Conclusion

As mentioned in the beginning of this thesis, our work aims to transplanting the social network analysis approaches from popular social networks to other social networks, especially those has low popularity and seldom being discussed. It is suggested that the analytical techniques can be migrated from one network to another with minimal level of modifications while the performances can be preserved.

However, the first problem faced was that social networks are different from each other. The features of the networks like the contacts being shared and communication approaches will not be the same. Therefore, work first proposed a generalized data model, the Generic Networks Data Model (GNDM), to conceptually represent the data from heterogeneous network so that they can be compared and evaluated. Then, a data selection approach, the Generic Network Data Linking (GNDL), is suggested hence the contents from arbitrarily selected social network pair can be matched according to their posting time and subjects being studied in analyses. In addition, the Networks Content Linkage (NWCL) algorithm was proposed. Such algorithm is capable of connecting two documents by forming a query from the title or topic sentence from one of the documents. Based on the linked contents, an abstract association model, the Generic Network Data Association (GNDA), is developed. The GNDA conceptually models the possible association between online and offline social networks. The GNDA is constituted of two components: the transplant component which is a selection of association models those borrowed from the previous studies; the customized component which is a collection of data model specially designed for particular kinds of sites.

Three separate experiments were conducted to evaluate our work. The first experiment evaluated the performance of the NWCL algorithm by matching news articles with the messages in discussion forums. It is proven that queries can be generated from every news article if NWCL algorithm was used. It outperformed the separation by stop word, where over one-third articles failed to generate queries. However, such matching is not totally automatic and human-supervision was needed in the experiment in order to maintain the performance of the NWCL algorithm. The second and the third experiment evaluated the GNDA. The GNDA was separately applied to two online and offline social network pairs to evaluate their association. The second experiment demonstrated that the transplanted analysis models can still achieved satisfactory performances, but the performances can be slightly improved when the customized analysis models are added. The third experiment demonstrated the directionless of the GNDA. In addition to modeling the influences from social network to the external network, such as predicting the polling results, the influences from external network to the social network, such as responding the news, can also be fitted in the same model.

Chapter 5.2 Limitations and Future Work

Although the transplantations were completed and it was proven to be feasible in application, there are several outstanding issues which needed to be solved in order to strengthen our theory.

In the data collection process, it is assumed that news articles in the news channel are the first-hand information which is directly shared from the channel. However, in many cases, the articles about foreign news being shared in the sites are directly translated from various news media in foreign countries. In other words, some

contents in the news media are already influenced by foreign news media. It is expected that interesting and unexpected results will be found if the influences of foreign news media are being studied in future. Second, the query matching in the NWCL algorithm also has rooms of improvement. The existing matching approach often has relatively low precise. Therefore, it was restricted that more than a half of the query terms have to be matched in a query. Although such approach significantly improved the precise to the result, it has no effects on the queries which only have two query terms. However, such tuning is not generic enough as it may need to be adjusted in other cases in future. Therefore, the matching algorithm needed to be improved by also considering the concept of location and order of matched query terms. The Generic Network Data Associations (GNDAs) proposed in our work was focused on single-topic analysis. Additions or modifications of analytical models are needed to analyze multi-topics, where those topics might be correlated with each other. The two experiments about the GNDAs proved that the transplantation of association analysis is feasible. However, the number of experiments conducted was not large enough to support the transplantation can perform well in general. Extensive experiments have to be conducted to summarize the limited condition of the theory.

Bibliography

[1] R. Ackland, "Social Network Services as Data Sources and Platforms for e-researching SocialNetworks", Social Science Computer Review Special Issue on e-Social Science, vol.27, no.4, 2009, pp.481-492.

[2] S. Asur and B.A. Huberman, "Predicting the Future with Social Media, Media", international conferences on Web Intelligence and Intelligent Agent Technology (WIIAT), 2010, pp.492-499.

[3] K. Balog, G. Mishne and M.D. Rijke, "Why are They Excited? Identifying and Explaining Spikes in Blog Mood Levels", conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL), 2006, pp.207–210.

[4] V. Belák, S. Lam and C.Hayes, "Cross-Community Influence in Discussion Fora" in Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2012, pp.34-41.

[5] S.V.Belleghem, D.Thijs and T.D.Ruyck, "Social Media around the World 2012", Internet: http://www.slideshare.net/InSitesConsulting/social-media-around-the-world-2012-by-insites-consulting, Sep. 24, 2012 [Jun. 23, 2014].

[6] K. Birman, G. Chockler and R. V. Renesse, "Toward a Cloud Computing Research Agenda", Newsletter ACM SIGACT News archive, vol.40, issue 2, 2009, pp.68–80.

[7] J. Bollen, H. Mao and X. Zeng, "Twitter Mood Predicts the Stock Market", Journal of Computational Science, vol.2, issue 1, 2011, pp.1–8.

[8] D.M. Boyd and N.B. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, vol.13, issue 1, 2007, 210-230.

[9] B. O. Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From Tweets to Polls: Linkin Text Sentiment to Public Opinion Time Series", in Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2010, pp. 122–129.

[10] D. Gruhl, R. Guha, R. Kumar, J. Novak and A. Romkins, "The Predictive Power of Online Chatter", in Proceedings of the International Conference on Knowledge discovery in data mining (SIGKDD), 2005, pp.78–87.

[11] M. Helft and N. Bilton, "Design flaw in iPhone4, testers say", in The New York Times. Internet: http://www.nytimes.com/2010/07/13/technology/13apple.html, Jul. 12, 2010 [Jun. 23, 2014].

[12] D. Ikeda, T. Fujiki and M. Okumura, "Automatically Linking News Articles to Blog Entries", AAAI Spring Symposium, 2006, pp.78-82.

[13] A. Java, X. Song, T. Finin and B. Tseng, "Why we twitter: understanding microblogging usage and communities", proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007, pp.56–65.

[14] A.M. Kaplan and M. Haenlein, "Users of the World, Unite! The Challengers and Opportunities of Social Media", Business Horizons, vol. 53, issue 1, 2010, 59-68.

[15] H. Kwak, C. Lee, H. Park and S. Moon, "What is twitter, a social network or a news media? ", international conference on World Wide Web (WWW), 2010, pp.591–600.

[16] S. Lerman and J. Diviney, "Apple's antenna-gate casts doubt over millions of consumers" Internet: news.opinium.co.uk/survey-results/apples-antenna-gate-casts-doubt-over-millions-consumers, Jul. 27, 2010 [last retrieved on Mar. 15, 2013].

[17] H.L. Li and V. Ng, "Discovering Potential Drug Abuse with Fuzzy Sets", international conference on Systems Man and Cybernetics (SMC), 2010, pp. 2656-2662.

[18] H.L. Li, V. Ng and C. Chen, "Analyzing Social Networks with D-Miner Cloud", international conference on Computer Supported Cooperative Work in Design (CSCWD), 2012, pp.642-648.

[19] M. McPherson, L. Smith-Lovin and J. M.Cook, "Birds of a Feather: Homophily in Social Networks", Ann. Rev. of Sociology, vol. 27, 2001, pp.415-444.

[20] G. Mishne and M.D. Rijke, "A Study of Blog Search", European conference on information retrieval, Lecture Notes in Computer Science (LNCS), vol. 3936, 2006, pp. 289–301.

[21] S. Myers, C. Zhu, and J. Leskovec, "Information Diffusion and External Influence in Networks" in Proceedings of the international conference on Knowledge discovery and data mining (SIGKDD), 2012, pp.33-41.

[22] O. Phelan, K. McCarthy and B. Smyth, "Using twitter to recommend real-time topical news", ACM Conference on Recommender Systems (RecSys), 2009, pp.385–388.

[23] H. Sayyadi, M. Hurst and A. Maykov, "Event Detection and Tracking in Social Streams", in Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2009, pp.311–314.

[24] W. Stofega and M. Shirer "IDC Survey Finds the iPhone 4 Antenna Issue Sending Mixed Signals" Internet: http://news.opinium.co.uk/survey-results/apples-antenna-gate-casts-doubt-over-millions-consumers, Jul. 16, 2010 [Jun. 23, 2014].

[25] J. Sun and J. Tang, "A Survey of Models and Algorithms for social influence analysis", in Proceeding of the international AAAI conference on Weblogs and Social Media (ICWSM), 2010, pp.34-41.

[26] L. Tang and H. Liu, "Toward Predicting Collective Behavior via Social Dimension Extraction" IEEE Intelligent Systems, vol. 25, no. 4, 2010, pp. 19-25.

[27] X. Tang and C.C. Yang, "Generalizing terrorist social networks with K-nearest neighbor and edge betweenness for social network integration and privacy preservation", Intelligence and Security Informatics (ISI), 2010 , pp.49,54, 23-26.

[28] I. H. Ting, C.H. Lin and C.S. Wang, "Constructing a Cloud Computing Based Social Networks Data Warehousing and Analyzing System", international conference on Advances in Social Networks Analysis and Mining (ASONAM), 2011, pp.735-740.

[29] M. Tsagkias, M.D. Rijke and W. Weerkamp, "Linking Online News and Social Media", international conference on Web search and data mining (WSDM), 2011 pp.565-574.

[30] X. Yu, A. Pan, L. Tang, Z. Li and J. Han, "Geo-Friends Recommendation in GPS-Based Cyber Physical Social Network", international conference on Advances in Social Networks Analysis and Mining (ASONAM), 2011, pp.361 – 368.

[31] Y. Zhang, S. Zeng, C.H. Huang, L. Fan, X. Yu, Y. Dang, C.A. Larson, D. Denning, N. Roberts and H. Chen, "Developing a Dark Web Collection and Infrastructure for Computational and Social Sciences", international conference on Intelligence and Security Informatics (ISI), 2010, pp.59-64.