Analyze Online Social Network Data for Extracting and Mining Social Networks

CH.R.Vinodkumar
Asst.Professor, Department of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh, India.

Abstract – In recent years, social computing has become a very popular application in the Web, and therefore large amount of social (communication) data has been collected in different social computing application. This paper will introduce a methodology to collect and analyze multi-source social, and by this for extracting social networks from the data. This paper to show how the data can be collected pre-processed, analyzed. Hence this paper gives an idea about the key topics of using data mining in Online Social Networks which will help the researchers to solve those challenges that still exist in mining Online Social Networks.

Index Terms – Social Netowrk, Extracting, Mining, Web, Communication.

1. INTRODUCTION

With the rapid growth of Internet and communication technologies, there are many communication and social activities of people have been transferred to Internet-based platform, e.g. e-mail communication, instant messaging software and social networking websites (such as Blog and web albums), etc. Under this background, large amount of personal communication and social data has been aggregated and stored in different locations. However, these valuable data have not been well organized, treated and used. Thus, it is an interesting research issue about how to use current information techniques to process and analyze these data, such as artificial intelligence, data mining or visualization technique.

Social network analysis and construction are originally in the research fields of Sociology. In recent years, many research issues of information science and social networking have been concerned due to the development of information techniques and the requirements of data processing ability. The target of social network analysis and construction is relationship data and it is therefore suitable to process and analyze communication and social data that discussed previously.

Since the communication data, such as e-mails and the logs of instant messenger, are very common data in our daily life. However, there is less work focusing on how to organize, process and analyze these data. In this paper, we therefore will propose method to analyze the common social data and discuss how to extract social networks from multi-sources social data dynamically. Furthermore, this paper will propose system architecture to use the three techniques of social network analysis, social network construction and visualization to process and analyze those valuable data. The system will allow user to input tasks for dynamic social network analysis and construction and the final results will be presented by visualized mean and interface for decision support.

The structure of this paper is organized as below: In section 1, the background and introduction will be introduced. Some related literatures of social network extraction, social network and data mining and social network analysis will be reviewed in section 2. A system architecture about how to extract dynamic social networks from multi-sources data will be proposed in section 3 as well as the introduction of the components in the system. In section 4, we will focus on how to extract social networks from social data and how to use data mining techniques for decision support. In section 5, this paper will be concluded with the suggestions for future research.

2. LITERATURE REVIEW

In this section, related literatures will be reviewed, including social networks analysis, social networks extraction and social networking for decision support.

Social Networks Analysis: The research methodology of social network analysis is developed to understand the relationship between "actors", and the term actor can be a person, an organization, an event or an object. In a social network, each actor is presented as a node and each pair of nodes can be connected by lines to show the relationships. The social network structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology that used to understand the graph and the relationships and actors in the social network

There are three important elements that included in a social network: actors, ties, and relationships. Actors are the essential elements in the social network to define the people, events or objects. Ties are used to construct the relationship between actors by using a mean of path to establish the relationship directly or indirectly. Ties can also be divided into strong tie and weak tie according to the strength of the relationships; they are also useful for discovering the subgroups of the social network. Relationships are used to illustrate the interactions and relationship between two actors. Furthermore, different relationships may cause the network to reflect different characteristics.

The most important measurements of SNA include network size, diameter, density, centrality and structure holes. Size is a measurement to measure the amount of nodes or links in a network, and the measurement of diameter is to measure the amount of nodes between two nodes in a network. Density is used to calculate the closeness of a network.

Traditionally, researches about SNA are mainly focus on small group of actors and are process manually in most cases. However, with the rapid growth of Internet and web techniques, more and more data have been collected and it has become a hard task to process these data by only the mean of manually. Therefore, the scholars of information technology and computer science are starting to devote related researches to deal with these research issues. Currently, the researches of computer science in SNA can be divided into four main topics, including social networks construction, social networks extraction, social networks analysis and visualization.

Social Networks Extraction:

In the research field of information technology and computer science in social networking, social networks extraction is a subfield focusing on extract social networks from large amount of communication data. With the rapid growth of Internet and WWW, there are various kinds of data have been generated due to communication purpose. The common used communication data such as email communication data, web usage logs, event logs, instant messenger logs, logs of telecommunication…, etc.

Currently, there are some researches which are focusing on the extraction of these social data. For example, Bird et al. propose a method to extract social networks from e-mail communications, Agrawal et. al using web mining techniques to understand the behavior of users in newsgroup. Web is considered as the biggest database in the world, so that various social networks can be extracted from this resource such as Furukawa et al. were trying to identify social networks from blogspace Jin et al. and Matsuo et al.

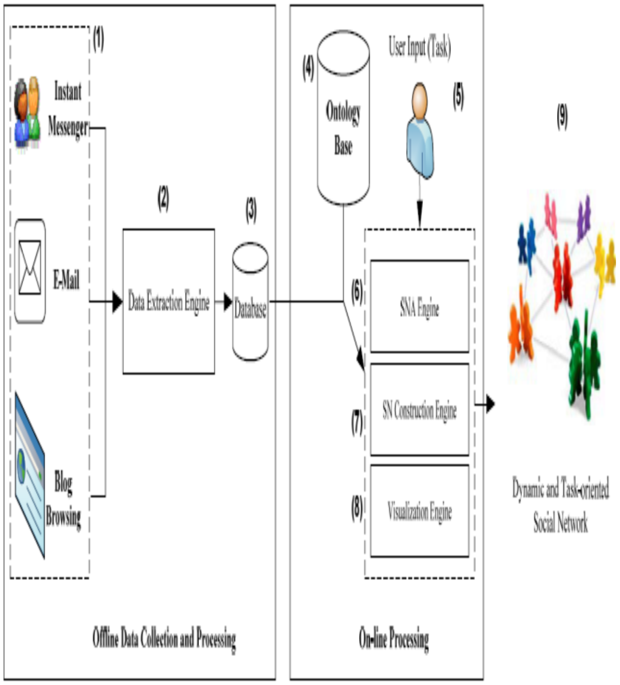
Most of the researches that discussed above are focusing on a single source for social network extraction. However, the issue of how to extract social networks from different sources has not been discussed well in related literatures. It is also a hard task about how to integrate multi-source data for social networking extraction. In addition to the problem of multisource data, instant messenger is a very popular and hot software for people to send message and communication recently. However, it has not been seen in recent research about how to extract social networks from the data. These research issues will be discussed in this paper.

Web Mining Techniques for Social Networking

According to different analysis targets and resources, the web mining techniques can be divided into three different types, which are Web Content Mining, Web Structure Mining and Web Usage Mining.

Web content mining is a web mining technique to analyze the contents in the web, such as texts, graphs, graphics, etc. Recently, most of web content mining researches are focused on the text data processing and few are focused on other multimedia data. Natural language process is therefore the main technology that used in this area. The concept and techniques of Semantic Web and Ontology also have to be studied.

Web usage mining is a web mining technique that can be used to analyze how the websites have been used, such as the navigation behavior of users. The server-side Click stream data (logs file) is the main sources that used for web usage mining. Client-side data (such as client-side logs file, cookies) is sometimes to be used due to some research concerns, such as in order to record more complete behavior of users. Different web usage mining analyses include basic statistical analysis of the navigation behavior of users in a website, such as how many times the website has been browsed, where the users comes from, etc. Furthermore, advanced web usage mining analyses can also be provided, such as more complex analysis for understand the navigation history of users in a website or cross website analysis.



Figuare1: social Network Extraction System Architecture

3. SYSTEM ARCHITECTURE

According to the research background and motivation of this paper, we have designed a system architecture to addressed the raised issues. The system will allowed us to collect social data from different sources, such as e-mail, instant messenger and blog. The multi-source social data will then be pre-process and analyze. The processed data can be used to extract social networks automatically and dynamically. The system then can be further developed to a decision support system. However, this paper will not focus on the decision support system and only the methodology to collect multi-source social data, preprocessing and how to integrate the data for generating social networks, which is the main contribution of this paper. The architecture of the social network extraction system is presented in figure 1.

As shown in figure 1, the system can be divided to two major phases according to the characteristic of processing. The two phases are offline data collection and processing and online process. The elements and process of the two phases will be introduced in detail below.

Offline data collection and processing: The first phase of the system is mainly working offline, and there are three elements in this phase including multi-source social data collection, data extraction engine and a database.

Multi-source social data collection: This is the first step of the system. In this paper, we intend to collect social data which are most related to personal daily communication. Thus, three types of data will be collected including instant messenger data, email data and blog browsing data. The messaging history of MSN or other messengers will be recorded in a structural format, such as xml. The detail of the message contents and communication target and time will be stored in the file. In the paper, the history file of MSN messenger is used as the social data of instant messenger.

About data collection: a data collection system will be introduced in section 4 of the paper. The system is developed by web-based concept and it allows the users to upload email and the history file of MSN messenger by either automatically or manually mean. About the collection of blog browsing data, we use a client side agent to collect the navigation history of a use when using particular blog. With the ability of client side logging, the complete browsing data will be recorded without missing, such as the behavior of browsing, posting a message and responding to a message.

Data extraction engine: The second step of the system is data extraction engine. The engine will firstly used to process the data that collected from previous step. Then, useful data will be extracted and filtered out from the raw data. The detail of how the data will be processed and extracted will be discussed in section 4 of the paper.

Database: After data collection and extraction, the output of the data extraction enginer will then be stored in a database. The database is designed according to the characteristics of different sources of social data.

Online processing:

The second phase of the system is a possible application of the paper in the future. The works in this phase are mainly processed online according to the data that collected offline. The elements of this phase will be introduced as follow even they are not the main focus of the paper.

Ontology-base In the system, the user can use the collected social data for personal decision support. It will allow the user to input a keyword and other parameters to ask for decision support. A social network for dealing with the problem will then be generated, which provides possible solutions for the user.

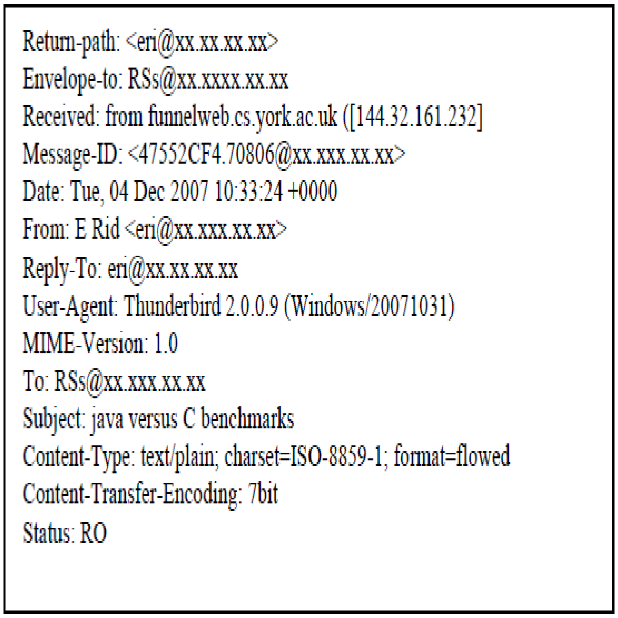
User input: There are two essential user input of the system, including a keyword and other parameters. The parameters are limitation and condition for the system to scale down the extracted social network.

4. DATA COLLECTION SYSTEM AND EXTRACTION ENGINE

In this section, the paper will focus on three main tasks of the system architecture that introduced previously. The three tasks include the data collection system, data extraction methodology and relationship calculation methods for social networks construction.

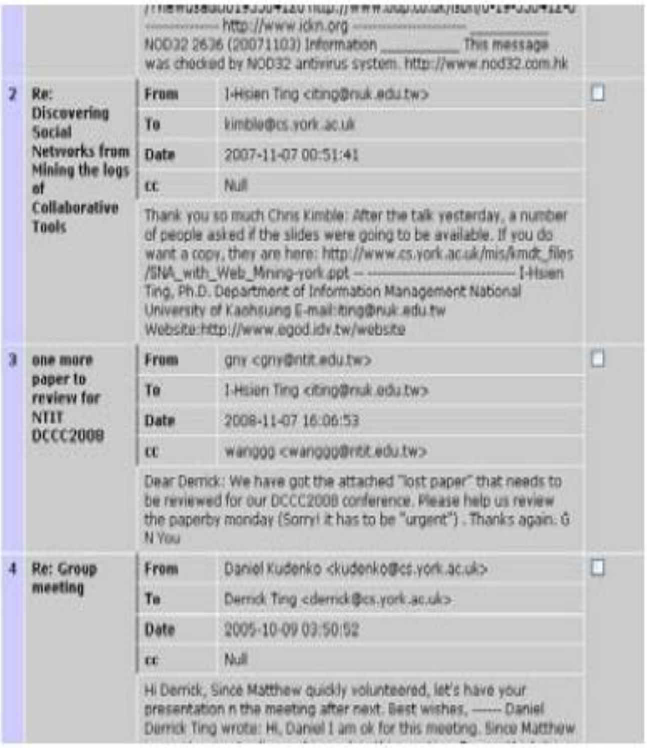
The data collection system: Data collection is the first step of the system, and therefore we design a sub-system for uploading related data. The system is a web-based system, and it allows the user to upload email file (in *.eml or text format), MSN history data (in *.xml format) and client side logging data (in *.log or text format). The format and file sample of email and MSN history data will be introduced below. About email file, the system allows users to upload email file manually or automatically. Although different email agents or servers may produce various email file, the system will accept any email file and there are some common fields in different email file format. Fields extraction of the uploaded email file will be discussed later in this paper to show which fields are useful for the research. Figure 2 shows a sample email file which is saved by an email agent and figure 3 shows the collected mail in the system.

In this research, only the instant messenger history file of MSN is accepted. The MSN history file is stored in a .xml file and based on the format of xml. Each contactor in the MSN contactor list has an independent history file, and the information that stored in the file include session ID, Date and Time, from, to and message content.



```
Return-path: <eri@xx.xx.xx.xx>
Envelope-to: RSs@xx.xxxx.xx.xx
Received: from funnelweb.cs.york.ac.uk ([144.32.161.232]
Message-ID: <47552CF4.70806@xx.xxx.xx.xx>
Date: Tue, 04 Dec 2007 10:33:24 +0000
From: E Rid <eri@xx.xxx.xx.xx>
Reply-To: eri@xx.xx.xx.xx
User-Agent: Thunderbird 2.0.0.9 (Windows/20071031)
MIME-Version: 1.0
To: RSs@xx.xxx.xx.xx
Subject: java versus C benchmarks
Content-Type: text/plain; charset=ISO-8859-1; format=flowed
Content-Transfer-Encoding: 7bit
Status: RO
```

Figure 2: Sample Email source file



			[illegible] ------------ http://www.idkn.org ----------------- ________ NOD32 2636 (20071103) Information __________ This message was checked by NOD32 antivirus system. http://www.nod32.com.hk	
2	Re: Discovering Social Networks from Mining the logs of Collaborative Tools	From	I-Hsien Ting <iting@nuk.edu.tw>	□
		To	kimble@cs.york.ac.uk	
		Date	2007-11-07 00:51:41	
		cc	Null	
			Thank you so much Chris Kimble: After the talk yesterday, a number of people asked if the slides were going to be available. If you do want a copy, they are here: http://www.cs.york.ac.uk/mis/kmdt_files/SNA_with_Web_Mining-york.ppt -- ------------------------ I-Hsien Ting, Ph.D. Department of Information Management National University of Kaohsuing E-mail:iting@nuk.edu.tw Website:http://www.egod.idv.tw/website	
3	one more paper to review for NTIT DCCC2008	From	gny <gny@ntit.edu.tw>	□
		To	I-Hsien Ting <iting@nuk.edu.tw>	
		Date	2008-11-07 16:06:53	
		cc	wanggg <wanggg@ntit.edu.tw>	
			Dear Derrick: We have got the attached "lost paper" that needs to be reviewed for our DCCC2008 conference. Please help us review the paperby monday (Sorry! it has to be "urgent") . Thanks again. G N You	
4	Re: Group meeting	From	Daniel Kudenko <kudenko@cs.york.ac.uk>	□
		To	Derrick Ting <derrick@cs.york.ac.uk>	
		Date	2005-10-09 03:50:52	
		cc	Null	
			Hi Derrick, Since Matthew quickly volunteered, let's have your presentation n the meeting after next. Best wishes, ----- Daniel Derrick Ting wrote: Hi, Daniel I am ok for this meeting. Since Matthew	

Figure 3:E-mail Collection system

Data extraction methodology:

In order to filter-out unnecessary data from the collected email and MSN history file, we developed a data extraction methodology to extract useful data for constructing social networks. In section 4.B, we have introduced a sample email file format. However, different email agents and servers may have different email format. Thus, we have selected some important fields in the email file which are useful for us to calculate the social relationship between the communicators of the collected emails and MSN history.

From the email file, some necessary fields will be extracted, including "deliver-to", "receive-id", "date", "to", "from", "subject", "msg-id", "priority", "reply-to", "mailer (agent)", "encode". "content-type", "content", "cc". These fields will be extracted from the original. Some of the extracted fields are used to identify emails and some are important for relationship measurement.

In addition to the extraction of email fields, we also extract useful fields from the MSN history file. The fields will be extracted from the file, including "msn-from", "msn-to", "msncontent", "msn-datetime", "msn-id", "msn-sessionid", "msntotage".mong all of the extracted fields, the "msn-sessionid" field is used to record the session number and "msn-totage" field is used to identify a communication with multi-users.

5. CONCLUSION AND FUTURE RESEARCH

Social and communication data are very common data in our daily life; however these data have not been used well for use to make decision. In this paper, we firstly provide an overview about the characteristics of these data and to illustrate how to use the concept of social networking and web mining to analyze the data. System architecture is then providing to give the reader a picture about how to use the multi-source social data to generate dynamic and task-oriented social networks and by this to assist the decision making. More detail process of data collection, data extraction and relationship measurement in the system are also provided. We will try to study how to use the techniques of web mining to get better analysis results and to enhance the accuracy of the decision support system. In addition, we will also try to understand more social data sources which are useful for including in the system and our future research.

REFERENCES

[1] Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003) "Mining Newsgroup Using Networks Arising From Social Behavior" In Proceedings of World Wide Web 2003 Conference, Budapest, Hungary, pp. 529-535

[2] Bird, C., Gourley, A., Devanbu, P., Gertz, M. and Swaminathan, A. (2006) "Mining Email Social Networks" In Proceedings of MSR 2006, May 22-23, 2006, Shanghai, China.

[3] Furukawa, T., Matsuo, Y., Ohmukai, I., Uchiyama, K., Ishizuka, M. (2007) "Social Networks and Reading Behavior in the Blogosphere" In Proceedings of ICWSM 2007, Boulder, Colorado, USA, pp. 51-58

[4] Garton, L., Haythornthwaite, C., and Wellman, B.(1997) "Studying Online Social Networks," Journal of Computer Mediated Communication (3:1).

[5] Godbole, N., Srinivasaiah, M., Skiena, S.: Large-Scale Sentiment Analysis for News and Blogs. In: Proceedings of ICWSM 2007, Boulder,Colorado, USA (2007)

[6] Jin, Y. Z., Matsuo, Y., and Ishizuka, M. (2007) “Extracting Social Networks among Various Entities on the Web” In Proceedings of the Fourth European Semantic Web Conference, 2007

[7] Kumar, R., Novak, J., and Tomkins, A. (2006) “Structure and Evolution of Online Social Networks” In Proceedings of KDD 2006 Conference, August 20-23 2006, Philadelphia, Pennsylvania, USA, pp. 611-617

[8] Joe, M. Milton, and Dr B. Ramakrishnan. "A survey of various security issues in online social networks." International Journal of Computer Networks and Applications 1.1 (2014): 11-14.

[9] Joe, M. Milton, and B. Ramakrishan. "Enhancing security module to prevent data hacking in online social networks." Journal of Emerging Technologies in Web Intelligence 6.2 (2014): 184-191.

[10] Joe, M. Milton, B. Ramakrishnan, and R. S. Shaji. "Prevention of losing user account by enhancing security module: A facebook case." journal of emerging technologies in web intelligence 5.3 (2013): 247-256.

[11] Joe, M. Milton, and B. Ramakrishnan. "Novel authentication procedures for preventing unauthorized access in social networks." Peer-to-Peer Networking and Applications (2016): 1-11.

[12] Joe, M. Milton, R. S. Shaji, and F. Ramesh Dhanaseelan. "Detection of M-worm to provide secure computing in social networks." Elixir International Journal–September-50 (2012): 10363-10365.

[13] Altunbey, Feyza, and Bilal Alatas. "Overlapping community detection in social networks using parliamentary optimization algorithm." International Journal of Computer Networks and Applications 2.1 (2015): 12-19.