Social Network Extraction: A Review of Automatic Techniques

Tasleem Arif
Department of IT,
BGSB University Rajouri,
J&K- India.

Rashid Ali
College of Computers & IT,
Taif University,
Taif-Saudi Arabia.

M. Asger
School of Mathematical Sci. & Engg,
BGSB University Rajouri,
J&K- India.

ABSTRACT

The advent of Web 2.0 has been instrumental in paradigm shift of how people communicate? These communications are a rich source of relationship data. Analyzing such a vast amount of relationship data is not a trivial task. Social Network Analysis is a promising field of research to take advantage of this huge pool of relationship data. But before this data is analyzed from Social Network Analysis perspective, *Social Networks* have to be extracted from this data. Social network extraction deals with the extraction of online social networks from a wide variety of online resources. These resources include web documents, e-mail communication, Internet relay chats, web usage logs, event logs, instant messenger logs, online blogs etc. Social network extraction is beneficial for many Web mining and social network applications such as expert finding for research guidance, potential speakers and contributors for conferences, journals, workshops, product recommendation, targeted advertising etc. In the last decade, many efforts have been made in the area of social network extraction. As a result, a good number of social network extraction methods have been proposed in the literature. These social network extraction methods use different sources for social network extraction. Some of these systems also use data from more than one resource. Although there are some social network extraction methods which construct a social network manually and as such cannot be considered in this work, as we deal with automatic methods only. In this paper, we classify automatic methods for social network extraction on the basis of information source they use. We also outline a general framework for social network extraction and give some future directions.

Keywords

Social Network, Social Networks Extraction, Data Mining, Link Analysis, Information Source.

1. INTRODUCTION

Recently, *online social networks* have gained significant popularity and are now among the most popular sites on the Web [1]. A social network is a structured representation of the social actors (nodes) and their interconnections (ties) and form social groups that share common interests [12]. Online social networks have emerged as a powerful tool for personal communication and interaction. Web based communities have become important places for people to seek and share knowledge and expertise [2]. Online social networks are organized around users in contrast to the Web which is largely organized around content [2].

Social networks have got a lot of focus from the research community long before the advent of the Web [3]. Social networks are formed by social interactions like co-authoring, advising, supervising, and serving on committees between academics; directing, acting, and producing between movie personnel; composing, and singing between musicians; trading and diplomatic relations between countries; sharing interests, making phone calls, and transmitting infections between people; hyper linking between Web pages; and citations between papers.

Social network extraction is an emerging field of research and the focus is to construct efficient systems for the identification of community structure in a generic network. Construction of the researcher network by automating information extraction from Web can benefit many Web mining and social network applications [5]. For example, in this case, *if all the profiles of researchers are correctly extracted, we will have a large collection of well-structured data about real-world researchers*. The profiles extracted can help in expert finding for *research guidance for new scholars*, *potential speakers* and *contributors for conferences, journals, workshops* etc. The extracted academic network may also be used for many applications like *trend detection/prediction*.

The challenges and issues concerning the researchers in the field of social networks are quite similar to those of the *WWW* [4]. An important and essential issue for social network extraction is the design of procedures and algorithms for profile extraction and name disambiguation which is a major challenge.

The automatic methods for social network extraction studied in this paper are classified based on the type of information source used to extract the social network. The rest of this paper is organized as follows: Section 2.1: presents overview of social network extraction and the challenges faced, 2.2: web mining techniques for social network extraction, 2.3: general framework for social network extraction, and 2.4: categories of information sources used. Section 3 presents the overview of different studies performed in the area of social network extraction. Finally we conclude the paper and outline some promising areas of future research.

2. SOCIAL NETWORK EXTRACTION

2.1 Overview

The *World Wide Web* (*WWW*) has become a popular medium to distribute information today. Data on the Web is rapidly increasing and is huge, diverse and dynamic so information users could encounter problems like: finding relevant information; creating new knowledge out of the information available on the web; personalization of information; learning about consumers or individual users, while interacting with the Web [7].

Exponential growth of public information on the Web in a set of interlinked heterogeneous sources [8] has made the information search a challenging task. Search engines are the most widely used tools for searching information on the Web, but the general approaches to analyze this information cannot integrate different sources. The advent of Web2.0 has added another dimension to the way the data and information is being populated and shared. These new medium of interactions generate a lot of data about personal communications which when combined with search engine results can be used for extraction of social networks in more efficient ways.

2.2 Web Mining Techniques for Social Network Extraction

Web mining deals with the discovery and extraction of useful information from the Web. Web mining can be classified into Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). WCM analyzes the contents on the Web, WSM deals with the links and structure of websites, and WUM can be used to analyze how websites have been used.

Web mining is very useful in online social network analysis and extraction. WCM can be used for a number of purposes such as categorization or classification of documents on an online social networking website, analyzing users' reading interests, determining their favourite content, etc. WUM provides usage data and user communications logs on an on-line social networking website. This data can be transformed into relational data for social-networks construction [9]. WSM is very useful for extracting online social networks by extracting the links from WWW, e-mail or other sources. In addition it can also be used to analyze path length, reachability or to find structural holes. For most online social networks analyses, the three types of Web mining can't work alone and it is usually necessary to utilize all three types of Web mining techniques together.

The Web mining techniques discussed above can be used for online social networks analysis. For example, clustering can be used for finding the group of closest people in a network or cross networks, association rule mining can help discover the hidden relationships between nodes in a social network or even cross networks, recommendation and information filtering purposes. There are huge resources of interpersonal communications, relationships and behaviors data available for online social networks analysis. This data is produced by the incredible developments of on-line social networking websites and applications [10]. Web mining is considered to be the most suitable information technique for the analysis of online social networks [4].

Social network extraction is usually the main task in SNA. The network can be extracted from different information sources such as the Web, email communications, Internet relay chats, telephone communications, organization and business events, etc. [11].

2.3 General Framework for Social Network Extraction

In the proposed framework, as shown in Figure 1, we divide the process of social network extraction based on user input into five steps namely Data Extraction, Preprocessing, Algorithm Design, Network Extraction and Visualization. Required data is extracted from any of the resources like e-mail communications, Web pages, blogs, social networking services, instant messengers, citations, etc. Data may be extracted from one or more resources depending upon the requirements. The data extracted goes through preprocessing. Data is cleaned and formatted as per the requirements. For the purpose of integration, name disambiguation is performed to make the extracted data useable. Processed data is stored in the database for future reference. Algorithms for relation identification and extraction, expert user determination etc. are then developed as the case may be. Based on the extracted relationship data and profiles, a social network is extracted which can be visualized through a visualization package.

2.4 Web Information Sources

The relationship information for extracting of social networks is obtained from various online sources. The sources that have been used in these studies are Web pages, e-mail communications, instant messaging, Internet relay chats, blogs, online social networking sites, news, web albums, etc.

The different social network extraction techniques can be classified into following six categories on the basis of information source they use.

1. Web based Social Network Extraction.
2. E-mail Communication based Social Network Extraction.
3. Instant Messaging/ Chat based Social Network Extraction.
4. Blogs based Social Network Extraction.
5. Online Social Networking Sites Data based Social Network Extraction.
6. Multi-Source Data based Social Network Extraction.

3. SOCIAL NETWORK EXTRACTION METHODS

3.1 Web based Social Network Extraction

Several studies use a search engine to extract social networks from the Web [6, 11, 13, 14, 15]. Co-occurrence of names on the web, which is basically obtained by posing a query including two names to a search engine, is commonly used as proof of relational strength. Referral Web [11] was the first

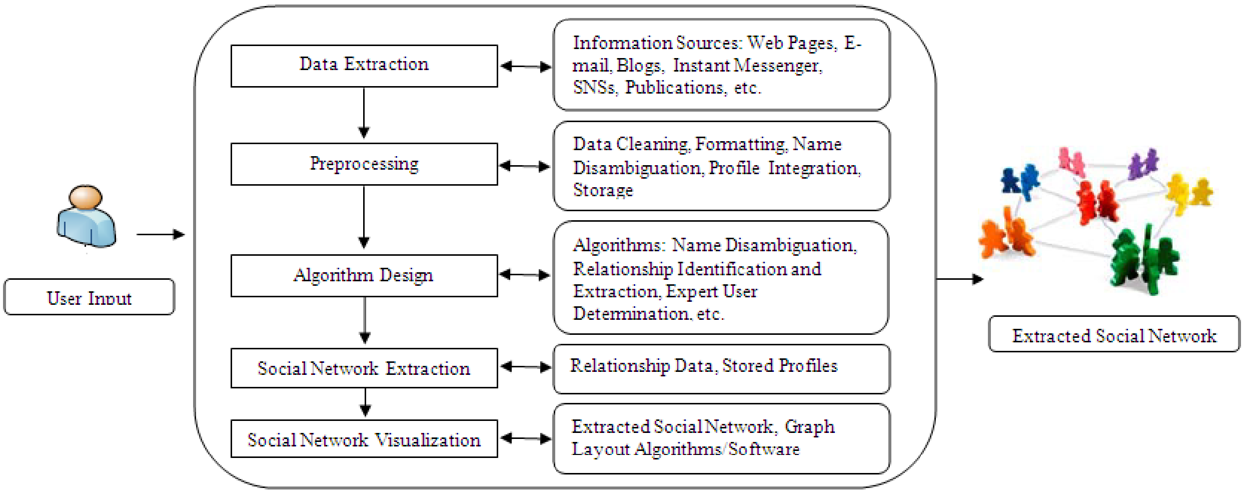
Data Extraction
Information Sources: Web Pages, E-mail, Blogs, Instant Messenger, SNSs, Publications, etc.
Preprocessing
Data Cleaning, Formatting, Name Disambiguation, Profile Integration, Storage
User Input
Algorithm Design
Algorithms: Name Disambiguation, Relationship Identification and Extraction, Expert User Determination, etc.
Social Network Extraction
Relationship Data, Stored Profiles
Social Network Visualization
Extracted Social Network, Graph Layout Algorithms/Software
Extracted Social Network

Figure1: General Framework for Social Network Extraction.

attempt of this kind to develop an automated interactive tool for social network extraction and finding shortest referral chains to experts. It uses *Altavista* to extract social networks through co-occurrence of names in close proximity in any document e.g. personal homepages, lists of co-authors in technical papers, citations, and organizational charts publicly available on the WWW taken as evidence of a direct relationship. The network obtained is an egocentric network. With increasing usage of Internet and development of WWW large amount of information about our daily lives is available online, making automatic extraction of social relations more demanding than when Referral Web was developed.

Tombe et al. [6] proposed a system for social network extraction of predefined conference participants from the Web by using participants attributes like *Name*, *E-mail*, *Affiliation*, etc. The relationships between any two participants are determined using the Web information gathered by posing a query to a search engine in a similar fashion of [11] by using Jaccard Co-efficient. P. Mika developed *Flink* [14], a system for extraction, aggregation, and visualization of online Semantic Web community. The Web mining module of *Flink* obtains as in [11] hit count from *Google* for both the persons *X* and *Y* individually as well as hit count for co-occurrence of these two names and calculates strength of their ties using Jaccard coefficient. It also performs the additional task of associating a researcher with a given topic of interest. Matsuo et al. developed *POLYPHONET* [13], which also uses *Google* to measure the co-occurrence of names. In their study, several co-occurrence measures [16] have been used, including the matching coefficient, mutual information, Dice coefficient, Jaccard coefficient, and overlap coefficient. The overlap coefficient performs best according to their experiments.

The fundamental idea behind [6, 11, 13, and 14] is that *the strength of a relation between two entities can be estimated by co-occurrence of their names on the web*. The criteria to recognize a relation, such as the measure of co-occurrence and a threshold, are determined beforehand. Although the approach is effective for extracting a social network of researchers, studies [6] indicate that it does not perform well for various entities on the Web. Co-occurrence-based methods become ineffective when two entities co-occur universally on numerous Web pages and function ineffectively when applied to inhomogeneous communities which mean that co-occurrence of names on the Web not always represent precisely the relational strength of two entities and for inhomogeneous entities it is difficult to precisely recognize the relation using a single criterion [6].

The quality of social network extraction method depends primarily on whether it has been able to address well the problems associated with profile extraction and name disambiguation. *Arnetminer* [5, 17] focuses primarily on *profile extraction* and *name disambiguation* for academic researchers and proposes a unified approach based on *Conditional Random Fields* (CRFs) [18] to extract researcher profiles from the Web using *Google* and integrating the extracted researcher profiles and the crawled publication data from the online digital libraries. A unified probabilistic framework for dealing with the name ambiguity problem has been proposed for integration. The draw back in this case is that *k* (actual number of researchers having same name, say *'a'*) has to be provided manually.

The study in [19] uses the co-occurrence of named entities in news articles to infer relationships and extract social networks from the relationship data. The proposed system extracts social network from unstructured data i.e. multilingual news from about 1500 websites in 40 languages. Most of the methods discussed in this section used co-occurrence based metrics to compute the weight of the extracted relations among entities but few of them have examined how to weigh each relation among entities beyond the co-occurrence based metrics [13]. Oka and Matuso [20] propose a method for weighting the relation among entities based on the weight of relations through the keyword, overcoming the shortcomings of the co-occurrence based metrics. The method receives a pair of entities and various relations that exist between entities as input. The output is the weight value for the pair of entities according to the generality of the keyword as a measure of its web hit counts.

3.2 E-mail Communication based Social Network Extraction

Email is one of the primary ways that people use to communicate and access their widespread social networks. Because of its inherent properties it is considered as a highly relevant area for research on communities and social networks [21]. It is the number one online activity for most users and there are few advanced email technologies that take advantage of the large amount of information present in a user's inbox [22]. Email data becomes a powerful information source for studying social networks because of a number of advantages: availability of large amount of data on personal communications in a standard electronic format; ubiquity of email usage; frequency, longevity, and reciprocity of email communications; type (content) of communication; temporal data; and availability on both sender and receiver side. In addition to the advantages, accessing email communications has certain issues as well. Privacy issues like compromising personal privacy and organizational confidentiality concerns are the biggest barriers for email related social research which can be alleviated by accessing only header information but ignoring information carried in the message significantly limits the potential of using email as source of information for analyzing social relationship. Although the format of email messages is relatively standard and it is easy to generate communication links from email archives, automatic extraction of social network is not easy because of issues like: multiple identities of same person; spam and group aliases; categorization of social relations by email content; weighting ties by different indicators such as reciprocity, frequency, and longevity of discussion; and aggregation of "*To, Cc, Bcc*" [23].

Several studies [21, 22, 23 and 24] have used email communication as data source for social network extraction and tried to leverage the associated benefits and address the issues concerning its usage.

Automatic construction of a network of correspondences and community detection from email data keeping privacy concerns in mind using only "to:" and "from:" fields from each email has been discussed in [21]. The study in [21] automatically identifies communities within an organization in two basic steps: (a) uses the headers of email logs to construct a graph where the vertices are senders or recipients of email messages and the edges denote an email communication between the nodes they connect, and (b) finds the communities embedded in the graph using the concept of betweenness centrality [25] to partition the graph obtained in first step into discrete communities of nodes. The authors claim that the method was able to identify small communities within a 400-person organization (HP Labs.) in a matter of hours, running on a standard Linux desktop PC and identifies leaders within these communities through network of correspondences.

In [21] population size is predefined and corporate directory (of HP Labs.) is used to remove name ambiguity. This is much easier as compared to an informal network where membership is not clearly defined and similar names pose ambiguity problem. In [21] only header information is used to extract the link structure ignoring information carried in the message which significantly limits the potential of using email as a research proxy for social relationship.

EmailNet [23] uses information both from the header and message to extract the link structure and addresses the concerns of privacy and confidentiality by hashing each email to make the messages unreadable to the human. It extracts mails both from personal email clients as well as organizational mail servers, uses filters to handle the issues such as, spammers, duplicate identities, etc., and employs a text clustering technology based email categorization function to categorize emails into several given types of social connections. Email usage pattern analysis functions in [23] help user investigate email interactions in detail, such as time distribution across hours and days, response thread visualization. An email oriented network visualization interface helps users explore the email social network. *R*-Social Network Analysis package (www.sna.stanford.edu/rlabs.php) is used as the social network analysis engine.

Extracting social networks and contact information from email and the Web and combining this information is discussed in [22]. The input to the system [22] is the set of email messages in a user's inbox and the output is an automatically-filled address book of people and their contact information, with keywords describing each person, and links between people defining the user's social network. The six modules of the system are *person name extraction, name co-reference, homepage retrieval, contact information and person name extraction, expertise keyword extraction,* and *social network analysis.* After extracting people names from email messages, it [22] finds each person's Web presence, and then extracts contact information from these pages using a probabilistic model (CRFs). In addition, the system uses an information-theoretic approach to extract keywords for each person that act as a descriptor of his or her expertise. It then obtains social links by extracting mentions of people from Web pages and creating a link between the owner of the page and the extracted person. The entire system is called recursively on these newly extracted people, thus building a larger network containing "friends of friends of friends". The network so obtained contains a significantly wider array of expertise and influence, and represents the contacts that the user could efficiently make by relying on current acquaintances to provide introductions, perform expert finding, and make new relevant connections.

Bird et. al [24] introduces the problem of identifying email users' aliases. The authors in [24] construct social networks of Open Source Software (OSS) developers through the email archives of OSS projects which provide a useful trace of the communication and co-ordination activities of the participants and consider the developers related if there is evidence of email communication between them. It employs a hybrid (automated/manual) approach to resolving aliases (name disambiguation). The automated approach executes in two steps: (a) automatically crawling messages and extracting all message headers to produce a list of <*name, email*> identifiers (IDs) and (b) executing a clustering algorithm that measures the similarity between every pair of IDs. IDs that are sufficiently similar are placed into the same cluster. Once clusters are formed, they are manually post-processed. Communication links between pairs of individuals are extracted from message headers, the sender, the receiver, the sent time, and the identifier of the message (if any) to which this message was a reply. Three measures, *in-degree*, *out-degree* and *betweenness* are taken as indicators of the importance of an individual in a network. Apache developer (www.apache.org) mailing list with 2544 separate IDs was used to empirically study the proposed techniques. It concludes that the most active developers play the strongest role of communicators, brokers, and gatekeepers.

3.3 Instant Messaging/ Chat based Social Network Extraction

Instant messaging (IM) or Internet Relay Chat is a popular form of real time computer-based communications service. Relationship extraction/identification is a central problem in the analysis of such large-scale social networks in their study as social networks as there is no clear measure of relationship strength. In [26] several such measures, obtained from the status log of an IM user, have been proposed that describe the link information between any pair of members. Resig et al. show [27] that, in spite of their simplicity, status logs contain a great deal of structure and relationship identification from this data is not a trivial task. The problem can be alleviated by obtaining acquaintances (e.g. buddies in AOL) list for each user but unfortunately, such lists are not published, so in order to obtain a collection of them one has to contact each author of each list making it an impractical solution. The solution lies in constantly tracking the status (online, busy, away, offline etc.) of each user relative to the IM service. This status data, along with the time at which a given client transitions from one state to another, is published electronically, making it possible to track the state of a population of IM users over a period of time. In [26] these status logs are used as a measure of the degree to which any two AOL IM (www.aim.com) users are related on the IMSCAN framework. The IMSCAN framework does not have content monitoring capabilities.

As pointed in [26], there are two major types of link data related to instant messaging networks; buddy lists: most useful but hardest to acquire, and 3rd party social-networking websites: readily available and easily accessible. [26] uses two types of link-based discovery mechanisms viz. co-relation and clustering have been employed. For co-relation based link discovery two experiments were performed in [26]. *First*: for each user, in the instant messaging activity log, the amount of time in seconds the user was online and the number of times the user changed his or her status to online from some other state is counted. *Second*: the degree to which each pair of users is, according to the IM status log, online at the same time is measured. [26] concludes that further efforts are required to accurately model the relationship of two users being linked if they are online at the same time and clustering techniques for recovery of information about social communities in IM networks.

In [27] an Internet Relay Chat (IRC) bot called PieSpy [28] is used to monitor channels and infer the social network structure. Measures like, direct addressing of users, temporal proximity, temporal density, and private message monitoring have been used to infer relationship strength. After inferring social relations, the authors in [27] have used modified spring embedder force model based on [29] for connected network components and m-limited force model [27] for disconnected components networks for extracting social networks.

3.4 Blogs based Social Network Extraction

Blogosphere is the network of social media sites, in which individuals express and discuss their opinions, facts, events, and ideas pertaining to their lives or society at large. This environment stands out as a rich source of social information. The dramatic increase in their size, diversity and popularity in the recent years has made it a ripe field for automatic extraction of underlying social networks.

Mesquita et.al. [30] proposed a system called SONEX [SOcial Network EXtraction] that extracts information networks from the blogosphere by identifying named entities and relations among them. It works by extracting and clustering entity pairs from the posts in the blogosphere. The clustering step groups together entities with similar context, each cluster is analyzed and a common label is assigned by inspecting the contexts of the pairs within the cluster. Unlike co-occurrence based methods, SONEX assigns a type for each edge by finding relations among entity pairs and uses sentences instead of web pages as the unit for co-occurrence resolution.

SONEX uses this idea behind Open Relation Extraction (ORE) [31] for clustering entity pairs extracted from the blogosphere. The system [30] comprises of three modules viz. Server, Document Parser and Relation Extractor. For the purpose of efficiency each module is implemented in parallel as a separate multi-threaded process. Of the four main Server components, Loader is responsible for reading documents, Sender for sending them to the Document Parser module, Receiver for receiving the parsed posts back and Storage for storing all processed data in a local database. The Server and the Document Parser modules work in a pipeline fashion. To avoid synchronization overhead and achieve parallelism all communication between the modules is buffered in local queues. The Relation Extractor module of SONEX runs after all blog posts are processed and all entities are identified. It works in three steps: identifying entity pairs, clustering such pairs, and labeling the resulting clusters. The Relation Extraction module considers two named entities to form a pair if they appear within the same sentence and are separated by at most 5 intervening words. The ultimate goal here is to cluster entity pairs that belong to the same relation. Hierarchical Agglomerative Clustering (HAC) [32] is used here to cluster the entity pairs since it does not require the number of clusters in advance.

Blogs have become an important means of information diffusion and the hidden social structures have significant influence on the rate and scope of information flow. The study in [33] proposed a new approach to measure the influence of social structure on information flow. Information flow in a network is tracked by using strength of relationships. Information flow is tracked by tracking an arbitrary topic on blogs cruising from the blogs infected by this topic to its neighbors. New topics on the blogs are discovered by change of high frequency words and using tags. All tags are added to the dictionary during blogs segmentation. Two words W1 and W2 in [33] are considered to belong to a same topic if they appear in same entries for a number of times. The experimental results indicate that the social structures have influence on the scope of the diffusion networks of "interest" topics and that the influence of social networks on information diffusion is related to the characteristic of the information itself.

3.5 Online Social Networking Sites Data based Social Network Extraction

The advent of Web 2.0 has evolved as the most preferred way of communication in a short span of time. The online social networking sites have taken full advantage of the paradigm shift and have become popular in the recent years and large amount of personal data is available on these sites. Online social networking sites contain huge number of user profiles containing semi-structured personal data [34]. There are some researches which have tried to extract social networks out of the dynamic data available on social networking sites.

The study in [34] extracts profile data from the deep web using the approach adopted in [35]. HTML content is parsed and a vector of tokens is produced in the Data pre-processing phase. Using Java IO methods HTML contents from profile's URL address are extracted and stored in character array. The text in the string after removal of HTML tags is broken down into tokens and the tokens are placed in a vector. Breadth First Search is then used to visit the specified profile webpage if it has not been visited before and the extracted personal details are placed in a repository. Friends list and their profile addresses are extracted and inserted into the repository if they have not been stored before. From the data in the repository an online social network graph is generated and analyzed.

The study in [36] deals with the possibility of extracting relevant information about relationships among subscribed users from Facebook (www.facebook.com). For the purpose of acquiring data [36] employs an agent that simulates the behavior of real users. It visits each publicly available profile on Facebook and extracts relationship information from them. The agent uses Facebook filters to get relationship information from real users only, keeping fan pages and companies away from friendship results. It [36] obtained an undirected graph comprising of 547,302 vertices and 836,468 edges and produced several graphs for SNA and visualization purposes.

Most of the social network extraction methods for online social networking services use explicit links but few studies have tried to analyze the relations within message threads. Social networks can be extracted by extracting the social relations hidden behind message threads [37]. The study in [37] proposes a method to extract the latent social relationship from a social networking service by analyzing the users' activities. Conversation intensity and the relationship are estimated from users' writing patterns and uses Facebook data for evaluation purposes.

It [37] uses a modification of frequent set mining techniques [38] to extract hidden relationships between message threads by discovering sets of users that appeared frequently together by examining users' writing patterns. To capture conversation intensity in a message thread, user's occurrence frequency in a message thread is used as the weight for each user. User groups whose occurrences are equally high and appear frequently together are extracted by adding weight to the basic frequent set mining algorithm and the harmonic mean was applied to the weight of each user. Firstly, in order to calculate relation score between two users in a message thread, the harmonic mean of two users' weights in the message thread is calculated and then overall score among all the transactions is calculated. It then uses weighted harmonic rule mining as a modification of association rule mining algorithm [39] to extract the relationship among the users within message threads.

Very few studies have addressed social network extraction from multimedia data. The study in [40] constructs social networks from contents of photo albums from Facebook data using unsupervised face recognition. It first determines the owner of the album by determining the frequency of the most occurring image (face) in the album. If two persons appear in the same photo, an edge is added between two corresponding albums.

3.6 Multi-Source Data based Social Network Extraction

The Web considered as the biggest database in the world has been used in various studies to extract social networks. Most of the researches mentioned above focused on a single source and the issue of social network extraction from different sources on the Web has not been discussed well in literature [41]. Ting et. al. [42] proposed a system to extract social networks from instant messages and e-mails. It has two major components: one for offline data collection; and the other for online data processing. Related communication data from e-mails and instant messenger is collected, the data so extracted is filtered by the data extraction engine and relevant data is stored in the database. Data collected, processed and stored by the offline data collection module is used in online processing module for social network construction and visualization.

The relationship (communication frequency) is the most important element to form a social network. In [41] the strength of a relationship (*Ri*) from a specific node to a node '*i*' is calculated using a weighted combination of e-mail, messaging and blogging relationships respectively.

Ansari et.al. [43] propose a system for mining social networks from Web server log files by considering average time spent on each page by every user. The Web server log files goes through a data preparation phase consisting of five processes viz. data collection, data cleaning, session identification, user identification and data summarization. The data prepared goes to the clustering architecture. The clustering architecture consists of five sequential steps viz. site structure mining, outliers deletion, user interest discovery, user clustering and compression. The importance rate of each page in [43] is obtained based on the average time spent on each page by every user. User's virtual communities are constructed from log files using Web mining techniques.

The study in [44] proposed a generic model for multidimensional social network extraction. It captures information about different types of activities and interactions between users along with the dynamics of user's behaviour. The proposed model utilizes information about different relations and the groups that exist within a given relation layer in a specific time window. It analyzes the social network on the basis of three dimensions viz. relations, time, and groups. The relations in the *layer dimension* describe all the relationships between the users of a system. They may be direct relationship between users via e-mail or phone or indirect relationships like sharing and co-editing documents in business intranet. A single layer represents a simple social network of all the users of the system connected to each other by that relationship only. Aggregation of different layers gives multi-layered social network of the same set of nodes connected by more than one relationship. *Group dimension* contains all the social groups that can be obtained in the clustering process. The *time dimension* may represent a snapshot at given time stamp, i.e. relation existing at that time, and also relations extracted for a given period, i.e. based on human activities within a particular time-window.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we classified and discussed various automatic methods for social network extraction. These methods have an edge over the manual methods where extracting attributes involved a lot of interaction with the profile owner, e.g., questionnaires and interviews. We also proposed a general framework that can be used for building of social network extraction systems. But several challenges like data sampling, combination of different types of web mining techniques, information representation on the Web are still to be addressed properly. It is necessary to design efficient tools and techniques for data extraction from the Web because it becomes difficult for end users to find useful data because of a number of issues. Information representation is one of them.

Social network profiles change all the time, not just in structure but in content as well as such there is a need for more work for their extraction from semi-structured pages like online social networking profiles. There is severe problem of lack of benchmarks for extracting relations from the blogosphere and very little work has been done for social network extraction from multimedia data. There is need to investigate relation extraction methods driven by a given schema describing the relations of interest in a particular time window. These methods are especially useful for extracting relations from a specific domain, such as stock market or politics. Automatic extraction of attributes is the way forward because it allows us to process much larger volumes of data which can be extracted over a period of time without user intervention.

Automatically obtained networks, e.g., Web-mined social networks, provide a good view of prominent persons, but they do not properly record relationships of novices, students, and other "normal" people. There is need to devise methods to address this problem and study of the interaction between the users and the system.

Multi-relational and multi-layered social network have got very little attention by the research community which can be attributed to their complexity. Such networks are more difficult to be analyzed than simple single relational or one-layered social networks and as such no established methods have been developed for their extraction.

5. REFERENCES

[1] http://www.ebizmba.com/articles/social-networking-websites. [last accessed: Dec,02, 2013]

[2] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B. "Measurement and Analysis of Online Social Networks.", In Proceedings of the 5th ACM/USENIX Internet Measurement Conference-IMC'07, San Diego, CA, October 2007, pp 29-42.

[3] Wasserman, S. and Faust, K. "Social Network Analysis: Methods and Applications." Cambridge University Press, New York, 1994.

[4] Chakrabarti, S. "Mining the Web: Discovering Knowledge from Hypertext Data." Morgan Kaufmann Publishers, USA, 2003.

[5] Tang, J., Zhang, D., and Yao, L. "Social Network Extraction of Academic Researchers." In Proceedings of International Conference on Data Mining-ICDM'07, Nebraska, USA, October 2007, pp 292-301.

[6] Tomobe, H., Matsuo, Y. and Hasida, K. "Social Network Extraction of Conference Participants." In Proceedings of 12th International Conference on World Wide Web-WWW'03, Budapest, Hungary, May 2003.

[7] Kosala, and Blockeel, "Web mining research: A survey." SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, 2, 2000, pp 1-15.

[8] Troyano, R., Lopez, G. and Gasca, M. "Competitive Intelligence Based on Social Networks for Decision Making." International Journal of Software Engineering and its Applications, 4(4), 2010, pp 93-104.

[9] Nowson, S., and Oberlander, J. "Identifying More Bloggers." In Proceedings of International AAAI Conference on Weblogs and Social Media, Colorado, USA, 2007.

[10] Ting, I-Hsien. "Web Mining Techniques for On-line Social Networks Analysis." In Proceedings of International. Conference on Service Systems and Service Management, Melbourne, Australia, 2008, pp 696-700.

[11] Kautz, H., Selman, B., and Shah, M. “The Hidden Web.” American Association for Artificial Intelligence magazine, 18(2), 1997, pp 27–35.

[12] Arif, T., Ali, R. and Asger, M. “Scientific Co-authorship Social Networks: A Case Study of Computer Science Scenario in India.” International Journal of Computer Applications, 52(12), USA, pp 38-45, 2012.

[13] Matsuo, Y., Mori, J., and Hamasaki, M. “POLYPHONET: An advanced social network extraction system from the web.” In Proceedings of the 15th Intl. Conference on World Wide Web-WWW'06, Edinburgh, Scotland, May 2006, pp 397-406.

[14] Mika, P. “Flink: Semantic web technology for the extraction and analysis of social networks.” Journal of Web Semantics, 3(2), 2005, pp 211-223.

[15] Jin, Y.Z., Matsuo, Y., and Ishizuka, M. “Extracting Social Networks among Various Entities on the Web.” In Proceedings of the 4th European Semantic Web Conference, Innsbruck, Austria, June 2007, pp 251-266.

[16] Manning, C. D. and Schutze, H. “Foundations of statistical natural language processing.” The MIT Press, London, 2002.

[17] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. “Arnetminer: Extraction and Mining of an Academic Social Network.” In Proceedings of 17th International World Wide Web Conference-WWW’08, Beijing, China, April 2008, pp 990-998.

[18] Lafferty, J., McCallum, A. and Pereira, F. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” In Proceedings of 18th International Conference on Machine Learning, MA, USA, 2001, pp 282-289.

[19] Pouliquen, B. and Atkinson, M. “Extracting and Learning Social Networks out of Multilingual News.” In Proceedings of the Social Networks and Application Tools Workshop (SocNet-08), Slovakia, 2008, pp 13-16.

[20] Oka , M. and Matsuo, Y. “Weighting Relations in Social Networks Using the Web.” In Proceedings of 23rd Annual Conference of the Japanese Society for Artificial Intelligence, Takamatsu, Japan, 2009, pp 1-2.

[21] Tyler, J. R., Wilkinson, D. M., Huberman, B. A. “Email as spectroscopy: automated discovery of community structure within organizations.” In Proceedings of International Conference on Communities & Technologies, Amsterdam, 2003, pp 81-96.

[22] Culotta, A., Bekkerman, R., and McCallum, A. “Extracting social networks and contact information from e-mail and the web.” In Proceedings of Conference on Email and Anti-Spam, CA, USA, 2004.

[23] Van Alstyne, M., and Zhang, J. “EmailNet: A system for automatically mining social networks from organizational email communication.” In Proceedings of 2003 North American Association for Computational Social and Organizational Science, 2003.

[24] Bird, C., Gourley, A., Devanbu, P., Gertz, M., and Swaminathan, A. “Mining Email Social Networks.”, In Proceeding of MSR 2006, Shanghai, China, 2006, pp 137-143.

[25] Wilkinson, D. and Huberman, B.A. “A Method for Finding Communities of Related Genes.” In Proceedings of the National Academy of Sciences, USA, 2003, pp 5241-5248.

[26] Mutton, P. “Inferring and Visualizing Social Networks on Internet Relay Chat.” In 10th IEEE Symposium on Information Visualization, Austin, TX, USA, 2004, pp 35–43.

[27] Resig, J., Dawara, S., Homan, C., and Teredesai, A. “Extracting social networks from instant messaging populations.” In Proceedings of Workshop on Link Analysis and Group Detection (LinkKDD2004), USA, 2004, pp 22-25.

[28] http://www.jibble.org/pircbot.php

[29] Fruchterman, T.M.J. and Reingold, E.M. “Graph Drawing by Force-Directed Placement.” Software Practice and Experience, 21(11), 1991, pp 1129-1164.

[30] Mesquita, F., Merhav, Y. and Barbosa, D. “Extracting Information Networks from the Blogosphere: State-of-the-Art and Challenges.” In Proceedings 4th International AAAI Conference on Weblogs and Social Media--Data Challenge, Washington, 2010.

[31] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. and Etzioni, O. “Open information extraction from the Web.” In Proceedings of International Joint Conference on Artificial Intelligence, Hydrebad, India, 2007, pp

D. Zhou, G. and Tan, C. L. “Discovering relations between named entities from a large raw corpus using tree similarity-based clustering.” In Proceedings of The International Joint Conference on Natural Language Processing, Korea, 2005, pp 378–389.

[33] Tang, J., Wang, T. and Wang, J. “Measuring the influence of social networks on information diffusion on blogosphere.” In Proceedings of the 8th International Conference on Machine Learning and Cybernetics, Baoding, July 2009, pp 3492-3498.

[34] Alim, S., Abdulrahman, R., Neagu, D. and Ridley, M. “Online social network profile data extraction for vulnerability analysis.” International Journal of Internet Technology and Secured Transactions, 3(2), 2011, pp 194–209.

[35] Park, J. and Barbosa, D. “Adaptive record extraction from web pages.” In Proceedings of the 16th International Conference of the World Wide Web-WWW’07, Alberta, Canada, 2007, pp 1335–1336.

[36] Salvatore Catanese, S., Pasquale De Meo, P., Ferrara, E. and Fiumara, G. “Analyzing the Facebook Friendship Graph.” In Proceedings of the 1st International Workshop on Mining the Future Internet-MIFI'10, Berlin, Germany, 2010, pp 14-19.

[37] Song, M., Lee, T., and Kim, J. “Extraction and Visualization of Implicit Social Relations on Social Networking Services.” In Proceedings of the 24th AAAI Conference on Artificial Intelligence-AAAI’10, Atlanta, Georgia, July 2010, pp 1425-1430.

[38] Liu, B. “Web Data Mining-Exploring Hyperlinks, Contents and Usage Data.” Springer, 2006.

[39] Agrawal, R. and Srikant, R. "Fast Algorithms for Mining Association Rules.” In Proceedings of the 20th Very Large Databases Conference, Santiago, Chile, 1994, pp 487-499.

[40] Luo, D. and Huang, H. “Link Prediction of Multimedia Social Network via Unsupervised Face Recognition.” In Proceedings of MM’09, Beijing, China, October 2009, pp 805-808.

[41] Ting, I., Wu, H., and Chang, P. “Analyzing Multi-Source Social Data for Extracting and Mining Social Networks.”, In Proceedings of 12th International Conference on Computational Science and Engineering, Vancouver, Canada, 2009, pp 815-820.

[42] Wang, K., Ting, I., Wu, H., and Chang, P. “A Dynamic and Task-Oriented Social Network Extraction System Based on Analyzing Personal Social Data.” In Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Denmark, 2010, pp 464-469.

[45]

[43] Ansari, A. and Jalali, M. “A system for social network extraction of web complex structures.” International Journal of Computer Science and Information Security, 9(8), 2011, pp 67-75.

[44] Kazienko, P., Musial, K., Kukla, E., Kajdanowicz, T., and Brodka, P. “Multidimensional Social Network: Model and Analysis.” In Proceedings of International Conference on Computer and Computational Intelligence, Bangkok, Thailand, 2011, pp 378-387.