REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.2331(UIF)
VOLUME - 7 | ISSUE - 5 | FEBRUARY - 2018

SOCIAL NETWORK EXTRACTION USING WEB MINING TECHNIQUES

Duppati Sanjay Kumar
Dept of Computer Science & Informatics, Kakatiya University Warangal.

ABSTRACT

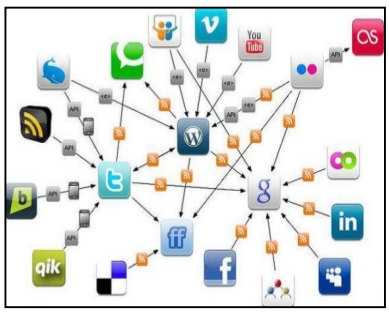
In the current scenario, at the crossroad of computational linguistics and data retrieval opinions and emotions are more valuable than the subject of the document. Linguistic resources are used to retrieve sentiments and also to classify it. Over the internet, not only the large volume of unstructured data is available but also the large amount of text is also generating day by day in the form of blogs, emails, tweets and feedbacks e.t.c. Text analysis is much more mature than unstructured data. WWW is large source for data mining. Email is a valuable and pervasive mean of communication in the information society, is one of the primary ways that people use to communicate and access their widespread social networks. A social network is a structured representation of the social actors (nodes) and their interconnections. Social network is an aggregation of social groups (communities) that share common interests and therefore include different relationships such as positions. Online social networking sites like Facebook, twitter, etc. contain huge number of user profiles containing semi-structured personal data and account for majority of Internet traffic.

KEYWORDS: *Social Net Work, Web Mining, Social Network Extraction*.

INTRODUCTION

A social network is a structured representation of the social actors (nodes) and their interconnections (ties). These networks can be represented as a graph G = (V, E). The set V denotes people joined in pairs by edges in E denoting acquaintances. Social networks form social groups that share common interests. These groups have emerged on the Web also and the demand for forming an on demand social network is immense. Members of virtual communities profit from being linked to other people sharing common interests despite their geographically dispersed affiliations. Social networks can be constructed for business entities like a company or firm, for educational entities like a school or University, or for any other set of entities.

Extraction and visualization of social relations can benefit many end users. It finds application in areas like crime and terrorism prevention, organizational network analysis, customer interactions, recommendation systems and communities. To visualize these networks, an understanding of the structure and the metrics associated with their graphs is required.



SOCIAL NETWORK ANALYSIS

The focus of Social Network Analysis (SNA) is relationships, their patterns, implications, etc. Using it, one can study these patterns in a structural manner. SNA can be used to identify important social actors, central nodes, highly or sparsely connected communities and interactions among actors and communities in the underlying network. SNA has been used to study social interaction in a wide range of domains, e.g. collaboration networks, directors of companies, inter-organizational relations, etc. Social networks have got a lot of attention from the research community long before the advent of the Web. Between 1950

and 1980, when Vannevar Bush's proposed hypertext medium 'Memex' was gaining acceptance, Social Sciences also contributed a lot in measuring and analyzing social networks. There are numerous examples of social networks 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, connections, and transmitting infections between people; hyper linking between Web pages; and citations between papers.

Web Mining

Data on the Web is huge, diversified, scattered and largely unstructured. Search engines have become the most widely used tool for searching required information on the Web. One often gets irrelevant responses because the words which form the query can have various meanings and contexts, e.g. while searching for a word 'java' on Google2 search engine it returns quite a number of links, around 20 million, to web pages ranging from java programming language to geographical locations. Although the responses which one gets from a query to the search engine are indexed based on the relevance of the page retrieved, it depends upon his expertise to find the relevant information from such a large corpus of web pages. Same thing happens for queries containing person names. It is interesting that person names form a major part of the queries submitted to a search engine. In such situations one need to look for other tools and techniques to obtain the required information in an efficient manner.

Web mining is an interesting area which can help provide user getting the required information one hand and managing the quantity of data on the other. In case of social networks, Web mining techniques have been used to extract and analyse social networks of academic researchers, students, conference participants, artists, firms, etc. from diverse sources like webpages, blogs, movie databases, digital libraries, etc. Web mining techniques can be categorised into three types as: web content mining; web usage mining; and web structure mining. In the context of social network extraction and analysis, web content mining can be used to categorise or classify webpages, blogs, etc. based on the similarity between their contents; web usage mining can be used to infer relationships between various entities from different types of log files e.g. e-mail communication logs, instant messaging logs, etc. by analysing their usage patterns; and web structure mining can be used to understand the ties between webpages by way of hyperlinking, digital citations by way of citing each other, etc.

Aims & Objectives

1. To understand the Web Mining and Social net working.
2. Developing a Social Network Extraction system that will be using multiple data sources, multi-relational in nature, based on spectral and temporal considerations where possible.
3. To know the Instant Messaging/ Chat-based Social Network Extraction.

Methodology

Extraction of the most recent publications from the Web, search engines and even those which the search engines are not able to crawl/index or which don□t find any mention in the citation digital library, will be extracted from whatever source they are present in. Blogposts, wiki edits, etc. can also be included in a researcher's publication list.

Ambiguity in Author Names

Person names are fundamental to our civilization. Every one of us is recognized by a certain name which depends upon a number of factors like place of origin, ethnicity, religion, family, etc. Despite that, person names are not unique. It is not that difficult to find people who share same or similar names. Names are not unique like certain numbers like Social Security Number (SSN), Permanent Account Number (PAN), etc. which are unique. We cannot have same SSN or PAN allocated to two different individuals. The problem gets aggravated further because same person may be referred to by different name variations in different places and contexts, e.g. a researcher may be listed by his full name on his organization's website or his own homepage but he may be referred to by the initial of his first name and full last name in a digital library such as DBLP. When we search Google for a name 'Rashid Ali', we get much more confusing results as compared

to the results which we got for the word 'java'. The search engine returned 19 million links to webpages which included those referring to a singer from Hyderabad, India, a former Prime Minister of Iraq, a founder of London based Architecture and Design Studio, and a Pakistani cricketer. Surprisingly, none of the first fifty links pertained to the Rashid Ali we were looking for. It becomes obvious that the task of searching the information for a person on the Web becomes quite challenging because of this name ambiguity problem.
Social media has changed the way user driven content and information is produced, transmitted and consumed. In an online social media setup, content is generated when someone performs a status update, posts a blog, comments on a post or status, tweets or retweets, etc. This content acts as a link between these social actors i.e. producers and consumers of information. This link or relationship is the most important piece of information for social network analysis. Social Network Analysis has a lot of potential for businesses and other consumers of information. Tracking and analyzing the flow of information on online social media provides a means for companies to gain feedback on their products or services. This information helps them to improve their products and services, market those products in a better way, and maintain competitive advantage. Participants on these networks have the advantage of making much more informed decisions by using the wisdom of crowds on these networks. This can be attributed to the availability of vast and diverse amount of information on these networks.

Instant Messaging/ Chat-based Social Network Extraction

Instant messaging (IM) or Internet Relay Chat has been a popular form of real time computer-based communications service. Relationship identification and extraction is a central problem in the analysis of such large-scale social networks as there is no clear measure of relationship strength. In 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. Relationship identification from status logs is a difficult task notwithstanding its apparent simplicity. The difficulty can be alleviated by obtaining acquaintances (e.g. buddies in AOL) list for each user but unfortunately, such lists are not public. List owners need to be contacted to obtain these lists which seem impractical. The solution lies in constantly tracking the status (online, busy, away, offline etc.) of each user relative to the IM service. It is possible to track status and transition (user state transitions) from electronically published IM data. In, these status logs are used to measure the degree of relationship between any two AOL IM users.

Blogs-based Social Network Extraction Blogosphere is a virtual collection of social media sites which provide a platform for individuals to express their ideas, discuss various events, share their opinions, facts, events, etc. These discussions range from personal life to society, politics to religion, science and technology to superstition, etc. This environment stands out as a virtual network and is considered a rich source of social information. The dramatic increase in their size, diversity and popularity in recent years has made it a ripe field for automatic extraction of underlying social networks. Contextual similarity between two named entities in sentences of a blog is used as a measure of their being related in and entities sharing a certain predefined degree of similarity are clustered together by Hierarchical Agglomerative Clustering (HAC). Blogs are becoming an important and somewhat indispensable means of information dissemination. The hidden information in these social structures has a significant impact on the rate and extent of information flow. A few studies, like are exploring novel ways of measuring how the hidden social structure in these networks influences the flow of information in them. Information flow in a network is tracked by using strength of relationships. In appearance of any two words W1 and W2 repeatedly in same entries is considered as an indication of their belonging to the same topic. Experimental results obtained in indicate two aspects of these networks: (a) social structures play an important role in the diffusion of "interest" topics, and (b) the authority exercised by social networks in diffusion of information contained in them is somewhat directly related to the information characteristic.

Multi-Source Data-based Social Network Extraction

Most of the techniques discussed above focus on a single source and the issue of social network extraction from different sources on the Web has not been discussed well in literature. Combination of instant messages and e-mails for social network extraction has been proposed in. 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. Average time spent on each page by every user as recorded by a web server log file is the source of relationship data in. The log files, after pre-processing, go 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 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.

Online Social Networking Sites-based Social Network Extraction

Online social networking sites like Facebook, twitter, etc. contain huge number of user profiles containing semi-structured personal data and account for majority of Internet traffic. The study in extracts profile data from the deep web using the approach adopted in. Vector of tokens is obtained by parsing the HTML content in the data pre-processing phase. Breadth First Search is then used to traverse the specified profile webpage, in case if it has not been traversed earlier, and the extracted personal details from the traversed profile webpages 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. An online social network graph is generated from the populated repository for analysis. In, a web agent which mocks a real user, is used to obtain an undirected graph from explicit relationships between subscribed users of Facebook. Only those profiles which are publicly available are accessed. However, These latent social relationships are extracted from a social networking service by analyzing the users' activities using a modification of frequent set mining techniques. The hidden relationship between a set of users is extracted by determining the occurrence as well as frequency of common terms in their writing pattern. The frequency is used as a measure of the strength of the relationship. In, social networks are constructed from contents of photo albums on Facebook using unsupervised face recognition. It first determines the owner of the album from the frequency of the most occurring image (face) in the album. If two persons appear in the same photo, an edge is added between their albums.

Publication Data Extraction from the Web

The exponential growth of Web has made available a whole lot of data from varying fields on the Internet. This data when used in an intelligent way can help find answers to a number of interesting questions. However, one has to dig deep into these online sources to get the relevant information and the quality of data extracted depends largely upon the methodology used by the end user to search for the desired results. Search engines serve as the preferred means for people to look for relevant information on the Web. Commonly available search engines, such as Google, are very good at finding documents on the basis of certain keywords, but currently they are limited to simple relevance-ranking mechanism. The results obtained by keyword searching are vast and beyond the limit of comprehension of a normal human being. The demand for organized search has increased many folds owing to large number of irrelevant and misleading search results thrown up in traditional Web search. The search results should be highly expressive, e.g. if one searches for Prime Minister of a country, he should get a list of all the Prime Ministers of that country along with all other relevant details. Besides, the search results should present data extracted from various sources, such as Web pages, in an integrated fashion, and in a form that makes its analysis easy.

E-mail Communication-based Social Network Extraction

Email is one of the primary ways that people use to communicate. Because of its inherent properties, it is considered as a promising area of work on communities and social networks. It is the number one online activity for most users, use e-mail communications to extract social networks from this highly structured information source. 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 make it a rich source for extraction of communication data. In some cases, only header data is used to extract a social network, whereas in some cases, contents in message body have also been used. Techniques like, compromise the privacy and confidentiality of the message, which may be of some concern. These concerns can be

addressed by accessing header information only. The information contained in the message body can be ignored to address privacy and confidentiality issues. However, ignoring information contained in message body significantly limits the potential of using email as source of information for analyzing social relationship. Mixed approaches like, obtain basic information from e-mail message headers and contact and other related information from the Web.

CONCLUSION

Social network extraction techniques based on the type of the information source they use. To the best of our knowledge, we were the first one to categorise them. Social relations play an important role in our life. In fact, we are defined in terms of our contacts and relations. Co-authorship is one of the most important relations for academics. Systematic analysis of this relation can help unravel hidden trends and interesting facts about individuals and institutions. There is huge amount of co-authorship data from which academic social networks can be extracted. The huge size and diversity of the data make it imperative to design automatic methods for extraction and analysis of these networks. Social networking services (SNSs) like Facebook, Orkut, LinkedIn and others have become very popular on the Web . An SNS that manages and stores social networks can become a base of information infrastructure in the future. The potential of Social Networking has been utilized to a good extent in the area of personal communication, evident from the growing popularity of these SNSs, but in the area of professional and academic interaction their potential has not been exploited as much. As a matter of fact, DBLP is just one of the various digital libraries and does not index all the publications. Therefore more efforts need to be made for extraction of publications data from multiple digital citation libraries. For this, issues like heterogeneity of digital libraries should also be taken into account. In ego-centric networks, the reason behind strong and weak collaborations may be revealed in future. In addition, the impact of removing the central node from an academic social network may also be studied in future. Due to rapid development of electronic communications, 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.

REFERENCES:

1. Accomazzi et al., 1997, Accomazzi, A., Eichhorn, G., Kurtz, M.J., Grant, C.S. and Murray, S.S. (1997) “The ADS article service data holdings and access methods.” In G. Hunt and H. Payne, editors, Astronomical Data Analysis Software and Systems VI, 125 of A.S.P. Conference Series, pp. 357-360.
2. Ansari and Jalali, 2011, Ansari, A. and Jalali, M. (2011) “A system for social network extraction of web complex structures.” International Journal of Computer Science and Information Security, 9(8), pp. 67-75.
3. Ali, 2010] Ali, R. (2010) “Performance evaluation of web search systems.” Ph.D. Thesis, Department of Computer Engineering, AMU Aligarh, India.
4. Bhattacharya and Getoor, 2007, I. and Getoor, L. (2007) “Collective entity resolution in relational data.” ACM Transactions on Knowledge Discovery from Data, 1(1).
5. Chakrabarti, 2003 Chakrabarti, S. (2003) “Mining the Web: Discovering knowledge from hypertext data.” Morgan Kaufmann Publishers, USA.
6. Cooley et al., 1997 Cooley, R. Mobasher, B. and Srivastave, J. (1997) "Web mining: Information and pattern discovery on the world wide web." In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, CA, USA, pp. 558-567
7. Mutton, 2004 Mutton, P. (2004) “Inferring and visualizing social networks on Internet relay chat.” In Proceedings of 10th IEEE Symposium on Information Visualization, Austin, TX, USA, pp. 35–43.
8. Tang et al., 2008, Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008) “Arnetminer: Extraction and mining of an academic social network.” In Proceedings of 17th International WWW Conference, Beijing, China, pp. 990-998.
9. Newman, 2010, Newman, M.E.J. (2010) “Networks: An Introduction.” Oxford University Press, United Kingdom.

10. Parimala, et al., 2011, M., Lopez, D. and Senthilkumar, N.C. (2011) "A survey on density based clustering algorithms for mining large spatial databases." International Journal of Advanced Science and Technology, 31, pp. 59-66.

Duppati Sanjay Kumar
Dept of Computer Science & Informatics, Kakatiya University Warangal.