Critical Analysis of Social Networks with Web Data Mining

[1]Sanjeev Dhawan, [2]Kulvinder Singh, [3]Vandana Khanchi
[1,2]Faculty of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra-136119, Haryana, India.
[3]M.Tech. (Software Engineering) Research Scholar, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra-136119, Haryana, India.
E-mail (s): [1]rsdhawan@rediffmail.com, [2]kshanda@rediffmail.com, [3]vandanakhanchi777@gmail.com

Abstract: Analysis of social networks concern with the communication among users by using them as nodes of a network (graph) and their relations which are considered as network edges. Study of such type of structures place on the intersection of different fields of research: graph theory, sociology, and data mining. This paper presents study about social networks using Web mining techniques. The structure of this paper is framed into five different sections. Section-I presents the background and introduction classification related to the social network, section-II describes the general Web mining process, taxonomy and techniques for social networks analysis, section-III reviews the related literature work about Web mining and social networks, and finally section-IV discusses the research direction that can be incorporated for data mining for social networks analysis. In this paper the efforts are made for critically analyzing the users' behavior on social networks.

Keywords: Data mining, social network, social network analysis, Web data mining.

1. Introduction

The scale of data on computers is growing at exponential rate in form of databases and files. Users need information out of these databases and files. At present, the use of Internet is increasing at rapid rate specifically associated to e-business and e-commerce applications. Data Mining is one of the kinds to support such kind of demand. Data Mining is considered as finding latent information in database. There are several challenging difficulties in data, Web and text mining research. The mining data may be structured or non-structured. Mining is of three kinds: data mining, Web mining, and text mining. Data mining concerns with structured data organized in a database while text mining concern with unstructured data and on the other side Web mining data deals the combination of unstructured and structured data. Web mining uses data mining and / or text mining approaches.

Table 1: Web mining v/s data mining

Comparison	**Web mining**	**Data mining**
Structure	It is the information from semi structured, structured and un-structured from Web pages. It elicits information from wide database.	Data mining take the information from explicit structure. It does not get the information from wide database compares to Web mining database.
Accessibility	Data access is publicly. Here not latent the data that is access in Web database along with get permission to Web log master.	It access data privately and only authorizes user access in database.
Scalability	Here the search processing is not a big, 10 million job in Web server database.	In data mining the search processing is large, 1 million jobs in database.
Storage of data	Web data is stored in Web server database and server logs.	In this data warehouse store the data.
Data	It works on online data.	It works with offline data.

1.1 Classification of Web Mining

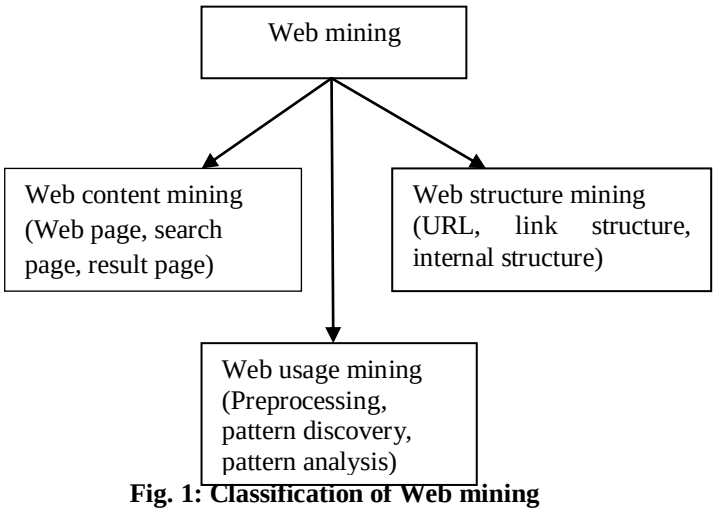

Fig. 1: Classification of Web mining

- Web Content: The data actually present in the pages that conveys information to the Web users. The Web page contains multimedia data e.g. text, HTML, audio, video, images, etc. It mainly comprises: (a) Mine the data/information/ content of documents/pages, and (b) Retrieval, filtering, clustering of search results, summarization, classification/categorization, etc.
- Web Structure: The organization of the Web pages linked through hyperlinks i.e. many HTML tags used to link one page to another and one Web site to other Web site. It basically includes: (a) Study the link structure of sites and pages and sites, and (b) Authorities and hubs, detection of communities, and page ranking (Google).
- Web Usage: The data that express the usage of Web collected on proxy server, Web servers and client browser with IP address, date, time etc. It generally contains (a) Analyze surfing behavior/patterns, usage data, and (b) Site marketing and restructuring.

2. Social Network

A social network is generally constructed and formed by continuously daily communication among people and that is why includes different associations, such as the betweennesss, position and closeness among groups or individuals [1]. To understand the social structure, social relationships, social network analysis and social behaviors, it is a very useful and important technique. Research about online social networks could be gone back to sociology, epidemiology and anthropology. For social networks analysis, the analysis objectives are mainly concentrated on Web resources, like structures, content and the user behaviors. Application of data mining techniques to the World Wide Web, called as Web mining, can be utilized for the analysis of social networks [2]. In Web mining, key analysis objectives are from the World Wide Web, in the form of Web content mining, structure mining and usage mining [3].

2.1 General Web Mining Process for Social Networks Analysis

A general process to use Web mining for social networks analysis steps are:

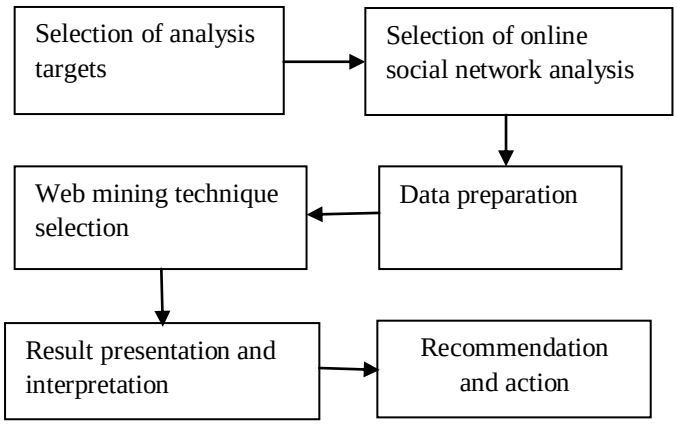

Fig. 2: Web mining process for social networks

(i) Selection of analysis targets: The first step is the choice of the analysis targets, such as e-mail, Web, telephone communications, etc. More than one target is to be selected.
(ii) Selection of social networks analysis: The social networks analysis methodology has been choose.
(iii) Data preparation: In this step, the relevant data will be gathered for analysis and thereafter the data is to be preprocessed and cleaned to store in database.
(iv) Web mining techniques selection: Selecting the Web mining techniques or their combination to be used and then performing operation with them.
(v) Result presentation and interpretation: The analysis results after Web mining are then presented and interpreted either automatically or manually or with visualization techniques.
(vi) Recommendation and action: This is an optional step, and the process may be ended after the analysis results have been produced.

3. Review of Literature

In this section, related literature about social network, the taxonomy and techniques of Web data mining are reviewed to demonstrate a broad view about these two topics. A social network is a social framework consists of individuals or group called nodes that are linked by one or more certain types of interdependency, like friendship, common interest, dislike, relationships of beliefs, knowledge. Meyappen et al. [4] stated that Web mining is the part of data mining that works with the fetching of interesting knowledge from the WWW. Internet is a base for E-Business and it provides a framework for the supplier to connect with the customer and serve them in a better way to appeal them to revisit their site. Satisfying one customer provides more customers to the particular vendor. Myra et al. [5] proposed that the ease and speed by which business transactions can be demonstrated over the Web has been a vital driving way in the rapid growth of electronic commerce. In addition, customer interactions, along with personalized content, online customer service, and online surveys, e-mail provide new medium of communication that were not previously exist or were very inefficient. The Web is revolutionizing the way businesses communicate with each other (B2B) and with each customer (B2C). It has introduced entirely new ways of performing commerce, involving auctions and reverse auctions, up-to-date content, micro-segmented offers, dynamic pricing. Web-based applications and Web services [6] are emerging at an exponential rate. This is generating a big amount of Web data having its own odd features. This in turn makes research in the field of Web data mining more difficult. Web data mining is a data mining application that deals with extraction of required or latent knowledge from the World Wide Web. Herrouz et al. [7] presented that today; the Web has become one of the most popular platforms for information retrieval and change. Since it is easier to present or publish documents, as the number of users, and so publishers, grows and as the number of documents increases, searching for information becomes into a very difficult and time-consuming operation. An online social network is a social structure consists of individuals (or groups) called nodes that are tied by one or more particular types of interdependency, like friendship, common interest, kinship, dislike, financial exchange, relationships of beliefs, knowledge or prestige [8]. Most widely, social network analysis represented social structure as a network that connecting members and channeling resources, influence the characteristics of connections instead on the characteristics of the individual members and consider communities as personal communities, which is, as networks of individual relationships that people foster, manage, and use in the work of their daily lives [9]. Popularity of academic social networking sites (SNSs) is increasing exponentially. Academic users at different levels of their career and from different disciplines are today becoming more interested in them. They are using those academic SNSs for several goals, and thus in different ways constituting some patterns of opinions. This paper aimed to describe usage patterns of an academic SNS (namelyAcademia.edu) through different academic users groups. To achieve this, users' profile data were gathered directly from Academia.edu Website [10]. Lee et al. [11] proposed that social networking sites (SNSs) have become novel phenomena in social interaction and communication patterns that have profound effect in the way people communicate and link with one another. The aim of this study was to test the enhanced cognitive-behavioral model of generalized problematic Internet use (GPIU) in the Facebook context. Professional networking has become a vital characteristic of several professionals' work [12] and various online social networking services (SNSs) provide help to the creation and maintenance of professional networks. This has also contributed to an increased potential for various professionals. Wagner et al. [13] said that with the emerging usage and popularity of online social media services, people now have accounts on several and diverse services like Twitter, YouTube, Facebook, and LinkedIn. In this paper [14], they systematically analyzed the problem of mining hidden communities in heterogeneous social networks. They proposed a new model for learning an optimal linear association of these relations which can best meet the user's preferences. From the obtained relation, better performance result can be achieved for community mining. Here community mining and social network analysis showed a major shift in methodology from single-network, user-independent analysis to multi-network, user-dependant, and query-based analysis. Kang et al. [15] said that graphs are of many types such as social networks, citation net-works, mobile call networks, biological networks, computer networks and the World Wide Web. By the lower cost of disk storage, the success of social

networking sites e.g. Facebook, Twitter and Google+ and Web 2.0 applications, and the high availability of data sources, graph data are being generated at a non paralleled rate. Now they are evaluated in terabytes and about petabytes, with over billions of nodes and edges. For example, Facebook every day loads 60 terabytes of new data, Yahoo had a 1.4 billion nodes Web graphs nodes in 2002, Microsoft had 1.15 billion URL-query pairs at 2009, and Google processes 20 petabytes per day. Mining and analyzing [16] Web data from online social networks can be difficult because of the huge amounts of data involved. Such activities are generally very expensive since they need a number of computational resources. Data analysis is going to be more accessible due to easier access to less expensive computational resources, with the cloud computing success recently. In this work they proposed use of cloud computing services such a possible solution for analysis of large amounts of data. Twitter was used as a source for a large data set that was a graph with 50 million nodes and 1.8 billion edges. Personal information mining (PIM) [17] can find out the hidden relationship and features of the target people which can be used for required post operation. A feature optimization method is proposed here to resolve the problem and perform efficient data mining. The method with the aim of data dimensionality deduction was based on combination of rough set theory with PCA approach.

4. Research Direction and Conclusions

As Web data is semi-structured /non-structured and non-homogeneous, there is a difficulty in discovery of required or unexpected knowledge information. This fact presented several challenging research problems. The challenges are in form of discovery of novel kinds of information, enhanced mining algorithms, incremental Web mining, finding relevant information, personalization of information, and learning about consumers or individual user. This paper studies the application of the techniques and concept of Web mining for social networks analysis, and reviews the related literature about Web mining and social networks. Social networks investigation carried out through the techniques of Web mining is an interesting field of research. However, there are many challenges in this research field to be resolve with improvement. For example, like finding communities in social networks structure, searching patterns in social networks and examining overlapping communities. We will shift our future research intension to handle the challenges discussed above, besides we will concentrate on how to utilize the Web mining techniques to some real on-line social networking Websites, such as on-line photo albums, comments and blogs.

References

[1] Jin, Y. Z., Matsuo, Y., and Ishizuka, M., "Extracting Social Networks among Various Entities on the Web", In Proceedings of the Fourth European Semantic Web Conference, pp. 251-266, June 3-7, 2007, .Innsbruck, Austria, published by Springer Berlin Heidelberg.

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

[3] Cooley, R., Mobasher, B. and Srivastave, J., "Web Mining: Information and Pattern Discovery on the World Wide Web", In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, 1997, pp. 558-567, Newport Beach, CA, USA.

[4] Abdul Rahaman Wahab Sait and Dr.T.Meyappan, "Web mining – a catalyst for e-business", Advanced Computing: An International Journal (ACIJ), Vol.3, No.6, November 2012.

[5] Ron kohavi, Brij masand, Myra spiliopoulou, Jaideep srivastava , "Web mining", Data Mining and Knowledge Discovery, 6(1), pp. 5-8, ACM SIGKDD International Conference on Knowledge Discovery in Databases (KDD), 2002, Edmonton, Alberta, Canada.

[6] Ujwala Manoj Patil, J.B. Patil, "Web Data Mining Trends and Techniques", in Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI-2012), pp. 961-965, ACM New York, NY, USA, 2012.

[7] Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi, "Overview of Web Content Mining Tools", the International Journal of Engineering and Science (IJES), Volume 2, Issue 6, pp. 106-110, 2013.

[8] Wen-jun, S., and Hang-ming, Q., "A Social Network Analysis on Blogspheres", In Proceedings of the 15th IEEE International Conference on Management Science and Engineering, pp. 1769 - 1773, 2008, Long Beach, CA, USA.

[9] Borko, F., "Handbook of Social Network Technologies and Applications (1st ed.)", Springer publication, 2010.

[10] Omar Almousa, "Users' Classification and Usage-Pattern Identification in Academic Social Networks", 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pp. 1-6, 6-8 Dec. 2011, Amman.

[11] Zach W.Y. Lee, Christy M.K. Cheung, Dimple R. Thadani, "An Investigation into the Problematic Use of Facebook", in Proceedings of 45th Hawaii International Conference on System Sciences, pp. 1768-1776, January 04-January 07, 2012, Maui, Hawaii USA.

[12] Linda Elen Olsen, Frode Guribye, "The Adoption of Social Networking Services: a case study of LinkedIn's impact on professional networkers", in Proceedings of IEEE 2009 International Workshop on Social Informatics, pp. 3-9, 2009, Poland.

[13] Anshu Malhotra, Laum Totti, Wagner Meira Jr., Ponnurangam Kumaraguru, Virgilio Almeida, "Studying User Footprints in Different Online Social Networks", in Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1065-1070, 26-29 Aug. 2012, Istanbul.

[14] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan, and Jiawei Han, "Mining Hidden Community in Heterogeneous Social Networks", March 2005, Report No. UIUCDCS-R-2005-2538, UILU-ENG-2005-1731.

[15] U Kang and Christos Faloutsos, "Big Graph Mining: Algorithms and Discoveries", ACM SIGKDD Explorations, pp. 29-36, Volume 14, Issue 2, December 2012, ACM New York, NY, USA.

[16] Pieter Noordhuis, Michiel Heijkoop, Alexander Lazovik, "Mining Twitter in the Cloud: A Case Study", pp. 107-114, in Proceedings of 2010 IEEE 3rd International Conference on Cloud Computing, 5-10 July, 2010 IEEE Computer Society Washington, DC, USA, Miami, FL.

[17] Guilan Hu, Xiaochun Cai, "Research on Feature Optimization Method in Personal Information Mining", in Proceedings of 2009 WASE International Conference on Information Engineering, vol. 1, pp. 656-659, 10-11 July, 2009, Taiyuan, Shanxi.