Using Mining Predict Relationships on the Social Media Network: Facebook (FB)

Dr. Mamta Madan
Professor
Vivekananda Institute of Professional Studies
GGSIP University,India

Meenu Chopra
Assistant Professor
Vivekananda Institute of Professional Studies
GGSIP University,India

***Abstract*—The objective of this paper is to study on the most famous social networking site Facebook and other online social media networks (OSMNs) based on the notion of relationship or friendship. This paper discussed the methodology which can used to conduct the analysis of the social network Facebook (FB) and also define the framework of the Web Mining platform. Lastly, various technological challenges were explored which were lying under the task of extracting information from FB and discuss in detail the about *crawling agent* functionality.**

***Keywords*—Online Social Media Networks (OSMNs); Facebook (FB); Data Mining; Crawling Process; Protocol**

I. INTRODUCTION

The web mining architecture called as crawler agent, that allow us to pull out the various different specimens of the popularly known, SNS (social networking site) Facebook and to study the network topology anatomy of the above social network graph. To be more concise, the two main techniques of OSMN (online social media network) are, the first one based on the idea of visual extraction (called as uniform sampling based on rejection policy without bias) and the second one based on sampling procedure (called as Breadth-first Search or Traversal having bias).

II. BACKGROUND AND RELATED WORK

The process of mining and analyzing data from OSMNs has attracted many researchers from the world wide [1] [2] [3]. Our focus is to discuss the techniques which are used to crawl huge and complex social networks and extract the data from them. Then this collected data is mapped with the graph data structures with the aim of understanding their structural traits. Kleinberg [4], laid the foundation for all efforts, by indicating that the geographical properties of social graphs may be the trustworthy indicators of user's behaviors. The spectrum of targeted research queries arising from the analysis of OSMNs is unlimited. But for our research paper is focusing on the three important themes which are as follows:

A. OSMNs Dataset

The task of extracting relevant data from Web mining Platforms by means of OSMNs web extraction techniques. Since OSMNs Datasets resides in back-end servers and are not available publicly, so they are accessible only through Web interface. The research done on the friendship graph of the FB by Gjoka et al. [5] using many visiting algorithm for example (Random Walk or BFS) with the aim to produce a uniform sample of the FB graph. Our focus in this paper is to creep the little part of the social network graph like FB and to figure out the structural characteristics of the crawled data. In [6], researchers crawled data from the complex SMNs like Live Journal, Flickr and Orkut.

B. Uniform Node Detection (UND)

The task of acquiring the extent of uniformity of two nodes or users in SM graphs. Finding users of common properties and also to calculate their uniformity is by means of Jaccard coefficient similarity metrics on the sets of their neighbors [7].But the disadvantage of this coefficient is firstly, not taking global information into consideration, secondly, it showed the similarity between nodes even if nothing real similarity exists between them because of the fact that nodes having high number of acquaintances would have high probability of sharing. In [8], authors suggested uniformity between two users increases, if one user exchanges acquaintances with another who have less number of acquaintances. Many other methods have explored in this like *Regular Equivalence* (two nodes are uniform or similar if they have uniform acquaintances too), in [9] authors, used the approaches Katz coefficient, Simrank [10], provides a method on iterative fix point, where in [11], researchers, have given the nodes uniformity as optimization problem and in [12], they worked upon directed graphs and exploited an iterative approach.

The other approaches for the node similarity in social media network analysis are *Formal Concept Analysis* (it depends upon the formal relationship between nodes and then calculate the nodes similarity which is hard to compute because it rely on the concept of number of common friend between the nodes) and *Singular value Decomposition (SVD)*[13] which used a technique from Linear Algebra and able to compute the uniformity degree of two nodes even if number of friendship relationship they share is less or close to zero.

C. Effective User Detection (EUD)

The process of discovering users having potential of charging others users to participate discussions/events/activities in their network. Few algorithms being designed for blog analysis such as HITS algorithm[14], Random Walk technique to search for initiators, HP Labs researchers [15], used Twitter to analyze behavior of the users in a network, in [16] authors found the concept of initiator i.e. user who starts the conversation in the network and last but not least in [17], authors recommended a model which

represent blogosphere as a graph and consist of nodes and edges where former represent the bloggers and later represents the blogger cites.

III. EXPLORING THE GRAPH STRUCTURE OF FB

As of March, 2014 (the data is collected) Facebook1 has 802 million (Daily), 1.28 billion (monthly) active users, 609 million (daily) and 1.01 billion (monthly) mobile active users. Approximately 81.2% of our daily active users are outside the U.S. and Canada. Our interest in exploiting the characteristics and the properties of this social network on a wide-scale. To achieve this goal, first is to collect the data from this online platform and then perform the analysis on it.

A. The Structure of the Online Social Network

The network layout of FB is simple. Every node is connected to each other by a relation called friendship. The social network graph is called as unimodal because it doesn't follow any hierarchy whereas friendship is called as bilateral reason being the relationship confirms among them. This FB graph is represented by G= (V, E): where V->End Users: E-> Edges (relationship). The graph is having two features, firstly, unweighted (Because within the network all the relationships have same value) and secondly undirected graph. In [18] adopted this kind of model for FB social network which has no loops simple unweighted undirected graph. In contrast to FB, the configuration or structure of other online social networks is more complex. For e.g. Nobii [19], YouTube and Flickr [20]. Twitter represents a multiplex directed network reason being it represent different types of relationships among users like "mention", "reply to" ,"following" etc.

This paper tries to explore the two things Firstly, Network Structural Information Retrieval Process of the FB network, secondly, FB data extraction process.

B. How to Retrieve the Structural Information of the FB

Various options are available to extract the information about the structure of FB, like one of option is acquire the data directly from the social networking company, which is not viable solution. Another option is acquire the data, directly from the platform itself, which is needed to reconstruct the model of the network; actually, we could take the representative sample of the social network, which further predicts its structure. Using various web mining techniques, this solution is viable, but the drawback of this option is that, large computational overhead of a large and complex Web Mining task. Moreover, network is not static; it is evolving, so its structure keeps on changing every time, because of this dynamism property of the network, the resultant sample would be a snapshot of the structure of the graph only at the time of data collection process.

There are many different data sampling algorithms that can be used for above mention task, but for our paper we zero down to only two approaches discuss in Table 1, firstly, "Breadth-First-Sreach (BFS) (Biased Approach)" and secondly, "Uniform (Un-Biased Approach)". Following are the characteristics of the above mention sampling algorithms.

TABLE I. TYPES OF APPROACHES FOR FETCHING STRUCTURAL INFORMATION EXTRACTION

Attributes	**BFS Algorithm**	**Uniform Sampling Algorithm**
1. Definition	Uninformed Traversal	Rejection-based Sampling
2. Advantages	• Easy to implement • Efficient • Optimal solution for un-weighted graphs [25,26,63,28,277, 27]	• Easily estimate the probability of a user by statistically[1,6] • To fetch the desired dimension of a sample, we randomly generate no. of User-Ids.
3. Hypothesis	Produces Biased Data towards high degree nodes [24]	Unbiased and Comparable Sample
4. Description	• User-Id's maintained in FIFO queue. • Time constraint is Adopted	• Parallelize the process of extraction. • User-Ids were stored in different queues.

C. How to Extract the Facebook Data

Once data collected could be used for comparing and analyzing their properties, behavior and quality. The quality parameters on which the collected data samples can be evaluated are: i) Significance with respect to statistical or mathematical models, ii) The quality of agreeing with results with other similar research studies. Because of the privacy and protection of data in FB, Twitter, etc., companies running these social networking services do not shared their data about users [21, 22]. We can access the information through graphical user interface with some technical glitches for example, using an asynchronous script; the friend-list can be crawled. Some of other online services like "Graph API (Application Programming Interface) [2]" etc., provided by FB developers team in 2010 and in by the end of 2011, using the Web data Mining techniques, we can able to access the structure of FB.

IV. THE SAMPLING FRAMEWORK OF FB

Figure 1 depicts the architecture of Web data mining process, which is composed of the following components.

1) A web-server executing Agents for Mining,

2) A Java based platform independent application, which executes the code of the agent,

3) An Apache interface, which controls and manages the flow of information through online network.

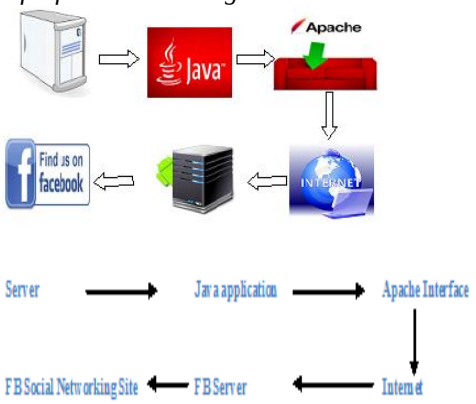

Fig. 1. Topology of the data Mining Platform

While executing, the mining agents inspect the FB server(s) to fetch the list of web pages of the friends connected to the specific requested users, reforming the structure of friendship among them. Finally, the data which has been collected would be stored on the web server and thereafter, goes post-processing task and delivered in an XML-format [23] for further processing.

A. *Facebook Crawling Process*

Figure 2 shows the architecture of FB Crawler, it is a cross-platform java based agent which actually crawl the GUI of the Facebook (front platform) and also the crucial part of the web data mining process. The given figure 2 below depicts the logic of the java agent, irrespective of the sampling algorithm executed. For the crawling agent execution, which is first preparative step in data mining process, includes two things, firstly chosen sampling algorithm and secondly, setting up some of technical parameters like maximum execution time, existing criteria etc. Therefore, the crawling process can initiate or start from the previous back-step. During the process of execution the java based crawling agent visits the friend-list web page of the requested user, obeying the rules of the selected sampling algorithm directives, for searching the social network or graph. To save I/O operations, all the data about newly discovered nodes and relationships among them are saved in a compact format. Termination of the process of crawling takes place when termination condition met.

Figure 2 shows the flowchart, which depicts the process of HTTP requests flow of the crawler with proper authentication and mining steps. First step, in the data mining process, is the front-end platform uses the Apache HTTP Request Library[3] to have a communication with the FB server(s). Second step, after establishing a secure connection (i.e. an authentication phase) and obtaining "cookies" for logging into the FB platform, finally getting the HTML web pages of the friend-list of the user through HTTP requests. This process is describes in Table 2.

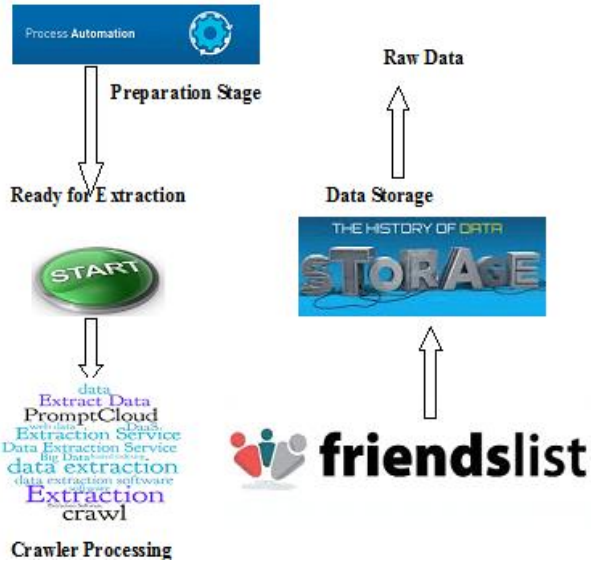

Fig. 2. The Flow diagram of the Data Mining task

The web-crawler has two executing modes:

a) HTTP Request-Based Execution: This mode is faster on large-scale of extraction.

b) Extraction Based on Visual Component: In this crawler embeds a Firefox Browser via XPCOM[4] and XULRunner[5]. The advantage of using this mode is its ability to perform asynchronous requests, for e.g. AJAX scripts but disadvantage of slower execution, time-consuming in rendering the web page.

At last, the paper discuss about the technical constraint imposed by the FB, which has been noticed during data mining task, is the limit of the generated friend-list web pages (which is not above 400 friends), through or via HTTP requests. To decrease network traffic, this limitation is put on, and if friends exceeds by 400, then asynchronous scripts fills the web page, this will led to a non-reproducible crawler or agent based on HTTP requests. This problem can be rectify by using different mining approach, for example use of visual crawler which is less cost effective and not viable for large-scale data mining tasks.

TABLE II. THE MINING AND AUTHENTICATION STEP OF THE CRAWLER VIA HTTP REQUESTS

Action taken		Protocol/ Method	URI
1.	Access the FB page	HTTP/GET	www.facebook.com/
2.	Authentication/Login	HTTPS/POST	Login.facebook.com/login.php
		HTTP/GET	/home.php
3.	Visit Friend-List	HTTP/GET	/friend-list/ajax/friends.php?id=#&filter=afp

V. CONCLUSION

The analysis as well as extraction of data from OSMN is a challenging task. This paper had discussed and explored, the different sampling algorithms that have been implemented to search or examine the social network graph Facebook that consist of countless friend-friend relationships. Out of the two sampling techniques, the visiting technique, BFS is known to deliver biasness in the scenario of incomplete traversal. Lastly, this paper described the random FB crawler agent, which could be used to generate samples of anonymous types. Analysis of these samples, SNA (social network analysis) using *graph theory (nodes and relations), diameter metrics, degree distribution and coefficient of clustering distribution* is the part of future discussion.

VI. WEB REFERENCES

1) *http://www.facebook.com/press/info.php?statistics*
2) *http://developers.facebook.com/docs/api*
3) *http://httpd.apache.org/apreq*
4) *https://developer.mozilla.org/en/xpcom*
5) *https://developer.mozilla.org/en/XULRunner*
6) *http://www.google.com/adplanner/static/top1000/*

REFERENCES

[1] Albert, R., Barabasi, A.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47-97 (2002).

[2] Garton, L., Haythornthwaite, C., Wellman, and B.: Studying online social networks. Journal of Computer-Mediated Communication 3(1) (1997).

[3] Ye, S., Lang, J., Wu, and F.: Crawling Online Social Graphs. In: Proc. of the 12th International Asia-Pacific Web Conference, pp. 236-242. IEEE (2010).

[4] Kleinberg, J.: The small-world phenomenon: an algorithm perspective. In: Proc. of the 32nd annual symposium on Theory of computing, pp. 163-170. ACM (2000).

[5] Gjoka, M., Kurant, M., Butts, C., Markopoulou, and A.: Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proc. of the 29th conference on Information communications, pp. 2498-2506. IEEE (2010)

[6] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, and B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007)

[7] Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques. Morgan Kaufman Pub (2011)

[8] Adamic, L., Adar, E.: Friends and neighbors on the web. Social networks 25(3), 211-230 (2003)

[9] Blondel, V., Gajardo, A., Heymans, M., Senellart, P., Van Dooren, P.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching.Siam Review pp. 647-666 (2004)

[10] Jeh, G., Widom, and J.: Simrank: a measure of structural-context similarity. In: Proc. Of the 8th SIGKDD international conference on Knowledge discovery and data mining, pp. 538-543. ACM (2002)

[11] Batagelj, V., Doreian, P., Ferligoj, A.: An optimization approach to regular equivalence .Social Networks 14(1-2), 121-135 (1992)

[12] Blondel, V., Gajardo, A., Heymans, M., Senellart, P., Van Dooren, P.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching.Siam Review pp. 647-666 (2004)

[13] Golub, G., Van Loan, C.: Matrix computations, vol. 3. Johns Hopkins University Press (1996)

[14] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604-632 (1999)

[15] Romero, D., Galuba, W., Asur, S., Huberman, B.: Influence and passivity in social media. In: Proc. of the 20th International Conference Companion on World Wide Web, pp. 113-114. ACM (2011)

[16] Mathioudakis, M., Koudas, N.: Efficient identification of starters and followers in social media. In: Proc. of the International Conference on Extending Database Technology, pp. 708-719. ACM (2009)

[17] Song, X., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proc. of the 16th Conference on Information and Knowledge Management, pp. 971-974. ACM (2007).

[18] Goldenberg, A., Zheng, A., Fienberg, S., Airoldi, E.: A survey of statistical network models. Foundations and Trends in Machine Learning 2(2), 129-233 (2010)

[19] Aiello, L.M., Barrat, A., Cattuto, C., Ruffo, G., Schifanella, R.: Link creation and profile alignment in the aNobii social network. In: Proc. of the 2nd International Conference on Social Computing, pp. 249-256 (2010)

[20] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007)

[21] Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Proc. of the Workshop on Privacy in the Electronic Society, pp. 71-80. ACM (2005).

[22] McCown, F., Nelson, M.: What happens when Facebook is gone? In: Proc. of the 9th Joint Conference on Digital Libraries, pp. 251-254. ACM (2009).

[23] Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.: GraphML Progress report structural layer proposal. In: Graph Drawing, pp. 109-112. Springer (2002).

[24] Kurant, M., Markopoulou, A., Thiran, P.: On the bias of breadth first search (bfs) and of other graph sampling techniques. In: Proc. of the 22nd International Teletrafic Congress, pp. 1-8 (2010).

[25] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G.: Analyzing the Facebook friendship graph. In: Proc. of the 1st International Workshop on Mining the Future Internet, vol. 685, pp. 14-19 (2010) 4, 52

[26] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Crawling Facebook for social network analysis purposes. In: Proc. of the International Conference on Web Intelligence, Mining and Semantics, pp. 52:1-52:8. ACM (2011).

[27] D'haeseleer, P.: How does gene expression clustering work? Nature Biotechnology 23(12),1499-1502 (2005).

[28] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007).

[29] Wilson, C., Boe, B., Sala, A., Puttaswamy, K., Zhao, B.: User interactions in social networks and their implications. In: Proc. of the 4th European Conference on Computer Systems, pp. 205-218. ACM (2009)