Key Challenges in Online Social networks Analysis: A Survey

K Sailaja Kumar, D Evangelin Geetha, T V Suresh Kumar
Department of Computer Applications
M S Ramaiah Institute of Technology
Bangalore-54, India
sailajakumar.k@msrit.edu

Abstract

Online Social Networks (OSNs), an emerging multidisciplinary research field has become the important element of information society. OSNs provide a basis for maintaining social relationships, finding users with similar interests, and locating content and knowledge that has been contributed or endorsed by other users. The key aspect of many of the OSNs is that they are rich in data, and provide unprecedented challenges and opportunities from the perspective of knowledge discovery and data mining. This survey paper reviews the current state-of-art on the selected key challenges in OSNs such as data gathering techniques, heterogeneity, scalability and missing data, which helps the research community and also suggests that significant further research is required in this area.

Keywords: heterogeneity, scalability and missing data

1. Introduction

The Internet has spawned different types of information sharing systems, including the Web. Unlike the Web, which is largely organized around content, OSNs are organized around users. Large quantities of information are shared through OSNs, making them attractive sources of data for social network research. The key aspect of many of the OSNs is that they are rich in data, and provide unprecedented challenges and opportunities from the perspective of knowledge discovery and data mining. Social network data consist of a set of actors and a collection of social relations between the actors. Two kinds of data can be analyzed in the context of social networks [1].

•**Linkage-based and Structural Analysis**: In linkage-based and Structural analysis, an analysis of the linkage behavior of the network can be constructed in order to determine important nodes, communities, links, and evolving regions of the network. Such an analysis provides a good overview of the global evolution behavior of the underlying network.

•**Adding Content-based Analysis:** Many social networks contain a tremendous amount of content which can be leveraged in order to improve the quality of the analysis. It has been observed that combining content-based analysis with linkage-based analysis provides more effective results in a wide variety of applications.

The rapid growth of the social networks observed several key challenges such as data gathering techniques, heterogeneity, scalability, missing data etc.

The amount and kinds of data generated by social network usage are too rich to be captured by only one of these methods. The data may be collected from OSNs, (i) from the social network websites; (ii) from surveys, by asking participants about their behavior; (iii) through deployed applications, by directly monitoring users as they share content online. Hence, we believe that a single data collection method is insufficient to capture all aspects of users' experience.

The heterogeneity of data in OSNs is characterized by huge data sets and varied data types, both semistructured and unstructured (videos, images, audio, click-streams, weblogs, text, and e-mail). In essence heterogeneous data is from any number of sources, largely unknown and unlimited, and in many varying formats.

Managing and processing on a network consisting of hundreds of millions of edges on a single machine [5], distributing status updates to millions of users [4, 6] and managing and distributing user generated content (UGC) to millions of users spread geographically [4, 2]. The growth and popularity of this is unprecedented and pose unique challenges in terms of scaling, management and maintenance [8].

The increasing volume of generated data in OSNs, and the growing concerns of users will exacerbate the problem of missing data over time. Prediction of missing information is an important part of data analysis in social networks [1, 2, 3].

This paper aims to present the current state-of-art on the key challenges such as heterogeneity, scalability, missing data in OSNs. The rest of the paper is organized as follows: Section 2 presents the state-of-the art in the selected challenges. In Section 3 the provisional research challenges are analyzed. Finally, in Section 4 we present the conclusions.

2. State-of-the art in the selected challenges

OSNs have recently become more and more important element of information society [1, 6]. A social network is the set of the actors (a single person is the node of the network) and ties, called also relationships, which link the nodes [1, 4]. The evolution of the social network depends on the mutual experience, knowledge, relative interpersonal interests, and trust of human beings [3, 11]. The measurements can be collected to investigate the number and the quality of the relationships within the network.

Data can be collected from OSNs using various methodologies. Collecting data from different sources enhances data analysis, and provides results than could not be obtained through only one method. This involves mixing measured data from OSNs and deployed applications, and self reported data from questionnaires, interviews, and experience sampling. Nevertheless, applying this methodology to a larger scale and in an ethical fashion is still an outstanding challenge that needs to be addressed [1].

2.1. Heterogeneity

The data on the web (and also in businesses and governments) is heterogeneous, unstructured, and often incomplete. Globally, the generation of data can be considered as a loosely coupled bottom-up process [2]. Many existing systems, e.g., Araneus [3], Ariadne [12], Information Manifold [9], Lore [12], Tsimmis [7], and others, have attempted to integrate heterogeneous web sources under a common interface.

2.2. Scalability

Scalability is often-overlooked property in OSNs. To address this issue [1] presented a novel approach to scale up OSN called One Hop Replication (OHR). This system combines partitioning and replication in a middleware to transparently scale up a centralized OSN design, and therefore, avoid the OSN application to undergo the costly transition to a fully distributed system to meet its scalability needs.

In case of peer-to-peer (P2P) network OSNs needs to trim the active connections per node, thereby enabling the system to scale and remain practical. As mentioned by [15] the following are the requirements are to be satisfied in P2P network OSNs

Network scalability: A P2P OSN system must link millions of users without any noticeable loss of performance

Tailored topology: OSN users may have hundreds or thousands of friends. A P2P OSN system must therefore tailor the number of connections for each node.

Connectivity: Updates from one user must reach all her friends. A P2P OSN system must ensure with high probability that all user friends remain connected.

Cost-effective update propagation: Users profiles are often updated and most of these updates represent small messages. A P2P OSN must therefore allow (i) the propagation of updates through a small number of edges; and (ii) the aggregation of different small messages into bigger composite messages, amortizing encapsulation costs.

Data Availability: Users may join or leave the network at any time, but their profiles must remain available from other online nodes. A P2P OSN must thus provide mechanisms to ensure high availability and topic-connectivity, even for unpopular users and under high churn rates. Topic-connectivity measures the fraction of users that can be reached by all their friends in the P2P overlay.

The challenge faced by the OSNs designers is: either to follow best practice and build a scalable OSN in order to accommodate a potential success that might never come, with the associated high costs in terms of time and resources, or to follow common practice, starting with a small centralized system with a short time to market and a low impact on the resources, but take the risk of death-by success if the OSN takes off.

2.3. Missing Data

The analysis of social networks is even more aggravated by missing values, because the complexity of network surveys is more likely to give scope for missing data, and the analysis and mapping of the structure of the network is especially sensitive to missing data [3, 16, 17, 18].

The studies in the literature, mostly by statisticians, include reconstruction of the unknown connections in a social network [4, 5], analyzing non-ignorable non-responses in a survey sampling [6, 7] and many others. Related to [5] there are the works that study the effects of missing data on measured properties of social networks [9] and the study of biases when obtaining a graph of the Internet based on measurements [10, 1]. Another related line of work is on sampling in large networks [12, 13, 14], where given a large network we would like to find some procedure to sample a small set of nodes such that important structural properties of the network are preserved.

Missing values are frequently found in data collected in empirical research. The easiest option is to simply ignore the missing data and only analyze the observed responses. However, this practice results in (serious) loss of information and a decrease in statistical power, and, more important, may lead to serious bias [19, 20]. Other missing data treatments include weighting procedures, model-based procedures (often likelihood-based), and imputation. Much is already known about the effects of missingness on (statistical) data analysis and the effectiveness of the various treatment procedures [19, 20]. However, the effects of missing data on the structural properties of social networks, and especially the treatment of missing network data are scarcely studied.

3. Research Challenges

There are several issues which researchers can cope with:

3.1. Heterogeneity

The data on the web (and also in businesses and governments) is heterogeneous, unstructured, and often incomplete. Researchers in the database community have called collection of heterogeneous data a data space, and have formulated a long-term research agenda to provide technologies for managing such data spaces.

3.2. Scalability

OSNs face serious scalability challenges due to their rapid growth and popularity. Scaling up is in general a non-trivial endeavour and, in the case of OSNs; the problem is particularly acute due to the rapid growth that they can potentially experience. Postponing scalability is dangerous, especially for OSNs system that can experience an extreme growth.

3.3. Missing Information

The Analysis of OSNs data is often suffered with non-responses and missing data. The most common way to deal with missing values is to replace them by some reasonable estimates using known or model-based cross-dependencies over the network in analysis. However, these methods do not typically consider networks that change with time, when another source of information is given by the temporal patterns arising from the network evolution. An important question when treating missing data is whether the data are systematically missing, and if so, whether missing data is related to the values of observed variables (properties or attributes) [4].

4. Conclusion

The availability and scale of data generated in Online Social Networks raises tremendous challenges and opportunities to gather structure and analyze the data. Collecting data from different sources which enhances the data analysis, and provides results, could not be obtained through only one method. Nevertheless, applying specific methodology to a larger scale and in an ethical fashion is still an outstanding challenge that needs to be addressed. OSNs face serious scalability challenges due to their rapid growth and popularity. Scaling up is in general a non-trivial endeavour and, in the case of OSNs; Postponing scalability is dangerous, especially for OSNs system that can experience an extreme growth for better predictive analysis. Overall, the rapid growth of the social networks themselves, the increasing volume of their generated data, and the growing concerns of users over privacy will likely to only exacerbate the problem of missing data over time. Missing data in networks is a longstanding but relatively poorly understood problem.

5. References

[1]. C C Aggarwal, “Social Network Data Analytics”, Springer Science Business Media, LLC 2011

[2]. J L Schafer and J W Graham, “Missing data: our view of the state of the art”, Psychological Methods, 7(2):147–77, 2002.

[3]. G Kossinets, “Effects of missing data in social networks”, Social Networks, Elsevier, 28:247–68, 2006.

[4]. M Huisman, “Imputation of missing network data: some simple procedures”, Journal of Social Structure, 10(1):1-29, 2009.

[5]. http://www.cmu.edu/joss/content/articles/volindex.html

[6]. P D Hoff, A E Raftery and M S Handcock,” Latent space approaches to social network analysis”, Journal of the American Statistical Association, 97(460): 1090–1098, 2002.

[7]. P D Hoff, “Multiplicative latent factor models for description and prediction of social networks”, Computational and Mathematical Organization Theory, 15(4):261–272, 2009.

[8]. T W Malone, K R Grant, F A Turbak, S A Brobst, and M D Cohen, “ Intelligent information-sharing systems”, Communications of the ACM, 30(5):390–402, 1987.

[9]. A Ansari, S Essegaier, and R Kohli, ”Internet recommendation systems”, Journal of Marketing Research, 37:363–375, 2000.

[10]. S Ha S, “Helping Online Customers Decide through Web Personalization”, IEEE Intelligent Systems, 17(6) 34-43, 2002.

[11]. P Kazienko, P Ko lodziejski, “WindOwls Adaptive System for the Integration of Recommendation Methods in E-commerce”, Proceeding of AWIC’05, Springer Verlag, pages 218-224, 2005.

[12]. L Terveen, W Hill, B Amento, D McDonald, and J Creter, “PHOAKS: A system for sharing recommendations, Communications of the ACM, 40(3):59-62, 1997.

[13]. R Hanneman and M Riddle, “Introduction to social network methods”, 2005.

[14]. J Hammer, H Garcia-Molina, J Cho, R Aranha, and A Crespo, “Extracting semistructured information from the Web”, Workshop on Management of Semistructured Data, PODS/SIGMOD’97, 1997.

[15]. A Olteanu and G Pierre, “Towards Robust and Scalable Peer-to-Peer Social Networks “, 5th Euro-Sys Workshop on Social Network Systems (SNS), Bern, Switzerland, April 2012.

[16]. R S Burt, “A note on missing network data in the general social survey. Social Networks, 9(1): 63-73, 1987.

[17]. A C Ghani, C A Donnelly, and G P Garnett, “ Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases”, Statistics in Medicine, 17: 2079-2097, 1998.

[18]. S P Borgatti and J L Molina, (2003), “Ethical and strategic issues in organizational social network analysis”, Journal of Applied Behavioral Science, 39: 337-349, 2003.

[19]. R A J Little, and D B Rubin, “Statistical Analysis with Missing Data”, New York: Wiley, 1987.

[20]. J L Schafer, and J W Graham, “Missing data: our view of the state of the art”, Psychological Methods, 7: 147-177, 2002.