Journal of Advanced Computing and Communication Technologies ***(ISSN:* 2347 - 2804*)***
Volume No.4 Issue No.2, April 2016

A parallel algorithm for reading the different variables in social networks using data mining techniques.

By

Anand.R[1] Pushpalatha.M[2] Dr Rajshekhar M Patil[3]
[1]Assistant Professor, Department of CSE, BMS Institute of Technology,Avalahalli
[2]IV Semester PG Student, Department of CSE, BMS Institute of Technology,Avalahalli
[3]Professor, Department of ISE/CSE, HKBK college of Engineering,
Bengaluru-560064, Karnataka, India.
Anandor@bmsit.in, pushpalatha197@gmail.com, pusmr1@gmail.com.

ABSTRACT

A social networks is a social structures made up of a set of nodes(such as individuals or organizations),set of different variables, and other social interactions between nodes. The social network provides a set of methods for analyzing the structures of whole social networks as well as different variables observed in structures. A social network analysis is a important factor for mapping and measuring of relationships and data flows between in organizations or groups. The links shows the relationships and flows between the nodes in social networks. The process of reading the different variables and information retrieving in social networks using data mining techniques such as preprocessing, data analysis and data interpretation for analyzing the information of different variables concurrently in social networks. A data mining techniques for reading a different variables in social networks and different algorithm approaches for reading variables concurrently.

Keywords:

Social Networks, Social Network Analysis and Data Mining Techniques, Preprocessing, Data Analysis, Data Interpretation.

1. INTRODUCTION

Social network is used to describe web-based services that allow individuals to create a profile within a specific network domain such that they can communicatively connect with other users within the network The occurrence of social networks has been one of the most exciting events in this decade. Many different popular social networks such as *Twitter*, *LinkedIn*, and *Face book* have become increasingly popular. In addition, a number of social multimedia networks such as *Flickr* have also seen an increasing level of popularity in recent years. Many such social networks are extremely rich in different content, variables and they typically contain a tremendous amount of *content* and *linkage* data which can be leveraged for analysis. [4] Data mining techniques is most commonly used to search a pattern and rules of data sets in social networks such as amount of data is flows to different network. Data mining techniques use of various statistical, machine learning and graphical methods in network and separate the knowledge in to a form which is very much useful for many real world applications. Social network analysis has become a very popular field of research as it is useful for networks many applications.

Data mining is a popular tool that can help to find patterns and relationships within our social data. Data mining discovers hidden information from large data in networks. Social network analysis has drawn much attention in graph data management research in network field[6].To outcome meaningful data mining results, we must understand our social data. There are several factors which has made the study of social networks gain enormous importance by networks factors. Few such factors include the availability of huge amount of social network data, the representation of social network data as graphs, nodes and so on.

Data mining techniques of social networks can be done using the graph mining methods such as classification/topologies, prediction, efficiency, pattern detection, measurement and metrics, modeling, evolution structure, data processing, and communities[8] .To reading the information represented in graphs we need to define metrics that describe the global structure of graphs, find the community structure of the network, and define metrics that describe the patterns of local interaction in the graphs, develop efficient algorithms for mining data on networks, and understand the model of generation of graphs. Social network and its analysis is an important factor and it is widely spread among many young researchers. Social networks research gained from psychology, sociology, statistics and graph theory. Based on graph theoretical concepts a social networks interprets the social relationships of individuals as nodes and their relationships as the lines connecting them shown in fig1.

Figure 1. Social network analysis.

2 COMMON CONCEPT IN SOCIAL NETWORK ANALYSIS

• Centrality: This measure gives a rough indication of the social power of a node based on how well they "connect" to the network. "Betweenness", "Closeness", and "Degree" are considered to fall under the measures of centrality.

.• Betweenness: It is defined as the extent to which a node lies between other nodes in the network. Here, the connectivity of the node's neighbors is taken into account in order to provide a higher value for nodes which bridge clusters. This metrics reflects the number of people who are connecting indirectly through direct links.

• Closeness: This refers to the degree with which an individual is nearer to all others in a network either directly or indirectly. Further, it reflects the ability to access information through the "grapevine" of network members. In this way, the closeness is considered to be the inverse of the sum of the shortest distance (sometimes called as geodesic distance) between each individual and all other available in the network.

• Degree: It is the count of the number of ties to other actors in the network.

• Clustering coefficient: This provides the likelihood that two associates of a node are associates with themselves. A higher clustering coefficient indicates a greater "cliquishness".

• Centralization: It is calculated as the ratio between the numbers of links for each node divided by maximum possible sum of differences. While a centralized network will have many of its links dispersed around one or a few nodes, the decentralized network is one in which there is little variation between the number of links each node possesses.

• Density: It is the degree that measures the respondent's ties to know one another. The density may be sparse or dense network depends upon the proportion of ties in a network relative to the total number of possibilities.

3. EVOLUTION OF SOCIAL NETWORKS

Social networks are now so well popular, that there are now a core 'top 4' social networks which are most popular which doesn't change from year-to-year. But, as we'll see in this scenario, the most popular social media sites vary a lot by level of usage in different countries and with graphics. So understanding these differences in popularity of different social networks is really important when targeting specific applications. When comparing the popular social networks it's best to review them by active account usage, not just the number of user accounts. We'll also see in this summary that some social networks are growing more rapidly than others while some are now in decline.

Social Media has grown massively, but it's growth is now starting to plateau. Interestingly the over 65s segment are now driving growth, as other age groups have plateau completely and use is hardly growing it all. Among the 50-63 age cohort, use hasn't increased since 2012.

3.1 The most popular social networks worldwide in 2015?

Here is the latest Global Web Index summary in January 2015 (the most recently published) showing social network account ownership and active usage. It's useful to have both since it's the active social media use statistic which really shows the potential of a platform. Although Face book is no longer growing at the rate it was based on the previous chart, it's clearly the number one.

The ongoing importance of the Google social platforms YouTube and Google+ may be a surprise since Google+ is no longer actively promoted, but they are integrated into their same account sign-in.

3.2 Social network popularity by applications

This is a great visualization of the popularity of social networks based on the interviews in the different applications. If you pick out your application it's probably way behind the data of users in which these four core social networks are most popular. The rise of social media over the past decade has presented brands with the opportunity

to interact with its clients in ways that have been application are used. The possibility to communicate with millions of potential customers at relatively low cost would have been use to over internet. Growth of social networks so well increasing the usability of users year by year like facebook ,twitter so on and increasing the users activity in social networks. The growth of social networks patents of different applications running on social activity with different users on networks to provides the communication to each other so well increasing the social networks as shown in figure4.1describes growth of social networks and corresponding to number of users.

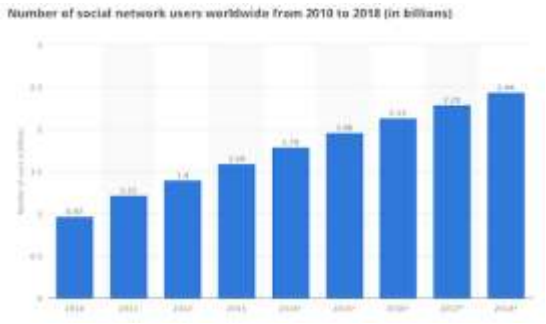

Figure 3.2.1 Growth of social networks and different applications.

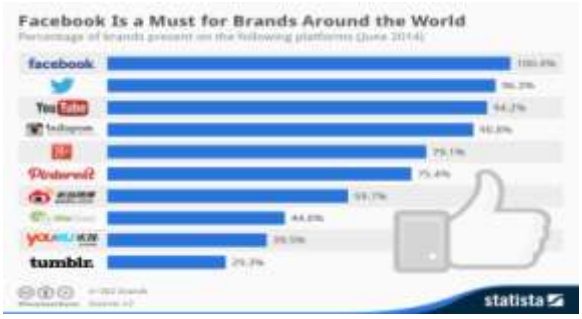

Figure3.2.2 Table shows the social networks verses year of using the social media site.

Number of users	Year	Applications	Percentage
0-.5	2007	tumblr	29.30%
.5-1.0	2008	ypuku.com	39.50%
1.0-1.5	2009	Wechart	44.00%
1.5-2.0	2010	welbo.com	59.70%
2.0-2.5	2011	printerest	75.40%
2.5-3.0	2012	gooleplus	79.30%
3.0-3.5	2013	instragram	90.80%
3.5-4.0	2014	youtube	94.20%
4.0-4.5	2015	twitter	96.30%

4 DATA MINING TECHNIQUES FOR SOCIAL NETWORKS

There are three methods processing of large data in different set of data in social networks such as preprocessing, data analysis and data interpretation as describe below.

a). Data preprocessing methods:

The preprocessing of the data in social networks is highly susceptible to noise, missing values and inconsistency .the quality of data will affects on network. In order to improve the quality of the data and consequently of the mining results data is preprocessed so as to improve the efficiency and ease of the mining process. Data preprocessing is a most critical steps in data mining process which deal with preparation and transformation of initial datasets .Data preprocessing methods are divided into following categories.

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction.

i).Data cleaning: data is to analysis by data mining techniques is incomplete ,noise and inconsitance. Data cleaning routines work to clean the data filling by missing values, smoothing noise data identify or removing outliers and resolving inconsistencies.

ii). Data integration: combines data from multiple sources into a coherent store. Schema integration the integrates metadata from different sources.Entity identification problem identify real world entities from multiple data sources.

Detecting and resolving data value conflicts for the same real world entity, attribute values from different sources are different possible reasons: different representations, different scales.

iii).Data Transformation: „ Smoothing: remove noise from data, aggregation summarization, data cube construction, generalization concept hierarchy climbing.

Normalization: scaled to fall within a small, specified range , min-max normalization, z-score normalization and normalization by decimal scaling .

iv).Data Reduction Strategies Warehouse may store terabytes of data. Complex data analysis/mining may take a very long time to run on the complete data set. Data reduction obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.

b) Data Interpretation and Analysis Technique.

The analysis of data via statistical measures and/or narrative themes should provide answers to your assessment questions. Interpreting the analyzed data from the appropriate perspective allows for determination of the significance and implications of the assessment.

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical or structural models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

c) Data Interpretation and Analysis Process.

- Consider the data from various perspectives. Whatever your project may be or whatever data you have collected from your business it's always best to ask what that data means for various actors or participants.
- Think beyond the data but do not stray too far from the data. Be mindful that you are not making too much of your data or too little. Make the link between the data and your interpretations clear. Base your interpretations in your research.
- Make visible the assumptions and beliefs, or mental models, that influence your interpretation. We each carry images, assumptions, and stories in our minds about ourselves, others, the organizations we work in, etc. As a composite, they represent our view of our world. Because these models are generally unarticulated, i.e., below our level of our awareness, if left unexamined, these assumptions and beliefs can lead to incorrect interpretations. Reflect on your own thinking and reasoning. Individually and/or collectively list your assumptions about the inquiry focus.
- Take care not to disregard outlying data or data that seems to be the exception.

5 A PARALLEL IN DATA MINING FOR SOCIAL NETWORKS

Parallel data mining (PDM) deals with tightly-coupled systems including shared-memory systems (SMP), distributed-memory machines (DMM), or clusters of SMP workstations (CLUMPS) with a fast interconnect. Distributed data mining (DDM), on the other hand, deals with loosely-coupled systems such as a cluster over a slow social network. It also includes geographically distributed sites over a wide-area network like the Internet. The main differences between PDM to DDM are best understood if view DDM as a gradual transition from tightly-coupled, fine-grained parallel machines to loosely-coupled medium grained LAN of workstations, and finally very coarse-grained WANs. There is in fact a significant overlap between the two areas, especially at the medium grained level where is it hard to draw a line between them. In another view, we can think of PDM as an essential component of a DDM architecture. An individual site in

DDM can be a supercomputer, a cluster of SMPs, or a single workstation. In other words, each site supports PDM locally. Multiple PDM sites constitute DDM, much like the current trend in meta- or super-computing. Thus the main difference between PDM and DDM is that of scale, communication costs, and data distribution. While, in PDM, SMPs can share the entire database and construct a global mined model, DMMs generally partition the database, but still generate global patterns/models. On the other hand, in DDM, it is typically not feasible to share or communicate data at all; local models are built at each site, and are then merged/combined via various methods. PDM is the ideal choice in organizations with centralized data-stores, while DDM is essential in cases where there are multiple distributed datasets. In fact, a successful large-scale data mining effort requires a hybrid PDM/DDM approach, where parallel techniques are used to optimize the local mining at a site, and where distributed techniques are then used to construct global or consensus patterns/models, while minimizing the amount of data and results communicated.

5.1 The parallelism of this algorithm can be exploited at three levels of granularity in social networks.

Coarse-grained: the computation starting from each source vertex can be considered as a task; the algorithm needs n tasks to compute partial values, which can proceed in parallel. If p processors are used, p copies of data structures are required. In a real world, this space usage for a large scale graph easily exceeds the available physical memory on conventional parallel computers.

Medium-grained: the BFS explores all neighbors of each vertex. One exploration of a vertex can be considered as one task, thus, all tasks could proceed totally in parallel if there was no shared neighbor between any two vertices. Otherwise, memory access conflicts occur and a synchronization mechanism is required to exploit this granularity of parallelism. Fine-grained: the task of exploring the neighbors of a vertex itself can also be parallelized. The amount of available parallelism depends on the degree of a vertex.

6. WHY PARALLELIZE DATA MINING?

Data-mining applications fall into two groups based on their internet . In some applications, the goal is to find explanations for the most variable elements of the data set that is, to find and explain the outliers. In other applications, the goal is to understand the variations of the majority of the data set elements, with little interest in the outliers. Scientific data mining seems to be mostly of the first kind, whereas commercial applications seem to be of the second kind . In applications of the first kind, parallel computing seems to be essential. In applications of the second kind, the question is still open because it is not known how effective sampling from a large data set might be at answering broader questions. Parallel computing thus has considerable potential as a tool for data mining, but it is not yet completely clear whether it represents the future of data mining.

7. CONCLUSION.

Data mining has become more popular today with the increase in the amount of data generated every minute in social network, with this issues like increase in size, data distribution, unstructured data, cleaning and pre-processing and is an open challenge. Data mining techniques can be speeded up by proper of parallel approaches for reading a different variable in social networks . In parallel scenario, we can get better performance in terms of memory utilization and speedup of retrieving information of different variables of nodes in social networks if there is utilization of proper parallel algorithm . Lot of advancements in the field of data mining is observed in last decade for networks. We have addressed receiving the details of different variables in the field of parallel data mining. Parallel data mining has to go long way for benefitting scientists, academicians and industries.

7. REFERENCES.

[1]. Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 533-542). ACM, 2012.

[2]. Aggarwal, C. An introduction to social network data analytics. Springer US, 2011.

[3]. Xu G, Zhang Y, Li L. Web Mining and Social Networking Techniques and applications, 1st edition. Springer; 2011.

[4]. Chelmis, C., Prasanna. VK.: Social networking analysis: A state of the art and the effect of semantics. Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) . IEEE, 2011

[5]. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (pp. 192-199). IEEE, 2011

[6]. Chen, Y., Lee, K.: User-centred sentiment analysis on customer product review. World Applied Sciences Journal 12 (special issue on computer applications & knowledge management) 32 – 38, 2011. ACM, New York, NY USA, 2011.

[7].Zeng, Li, et al. "Distributed data mining: a survey." Information Technology and Management 13.4 (2011): 403-409.

[8]Andrade, Diego, et al. "Task-parallel versus data-parallel librarybased programming in multicore systems." Parallel, Distributed and Network-based Processing, 2009 17th Euromicro International Conference on. IEEE, 2009.

[9]. Chen, Z. S., Kalashnikov, D. V. and Mehrotra, S. Exploiting context analysis for combining multiple entity resolution systems. In Proceedings of the 2009 ACM International Conference on Management of Data (SIGMOD'09), 2009

[10]. Chou, W. Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., Hesse, B. W.: Social media use in the United States: implications for health communication. Journal of medical Internet research , 11 (4), 2009

[11]. Du H. Data Mining Techniques and Applications an Introduction, 1st Edition. Cengage Learning Edition; 2010.

[12].X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, ―Top 10 Algorithms in Data Mining,‖ Knowl Inf Syst (2008) 141-37.

[13]. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[14]. Han J, Kamber M. Data Mining: Concept and Techniques, 2nd Edition. Morgan Kauffmann; 2006.

[15].Burt, R S.: Brokerage and closure: An introduction to social capital. Oxford University Press, 2005.

[16]. P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Proceedings of Knowledge discovery in Data Conference, pages 9–15, 1998.

[17]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, VLDB, pages 487–499. Morgan Kaufmann, 1994.

.

.