RESEARCH ARTICLE OPEN ACCESS

Role and Analysis of Big Data Technology in Social Network

Anitha R[1], Udhayakumar U[2]
1Department of Computer Science, Shanmuga Industries Arts and Science College, Tiruvannamalai
2Assistant Professor, Department of Computer Science, Shanmuga Industries Arts and Science College, Tiruvannamalai

ABSTRACT

The increasing use of social networks, such as Facebook, Twitter and Weibo has produced and is producing huge volume of data. Business firms and other organizations are interested in discovering new business insight to increase business performance. By using advanced analytics, enterprises can analyze big data to learn about relationships underlying social networks that characterize the social behavior of individuals and groups.An approach of this Social Network Analysis is to finding the importance of each individual subscriber of tele-communication networks. As these networks can be very large , the methods used to study them must scale linearly when the network size increases. Thus, an integral part of the study is to determine which social network analysis algorithms that have this scalability. Moreover, comparisons of software solutions are performed to find product suitable for these specific tasks.

Using data describing the relationships, we are able to identify social leaders who influence the behavior of others in the network and on the other hand to determine which people are most affected by other network participants. This study focuses on modeling the knowledge diffusion in social networks. We will present a new evolving model of a directed, scale-free network. We will test the effectiveness of our model by a simulation using data of a real world social network. For that reason, a complete process flow for finding influential subscribers in a telecommunication network has been developed. The flow analysis, machine learning is employed to uncover what behavior is associated with influence and pinpointing subscribers behaving accordingly.

KEYWORDS

Social Network Analysis, Telecommunication networks, Hadoop, Machine learning.

INTRODUCTION

As the amount of mobile data is growing fast. This implies not only increasing demands on telecommunication equipment, but also a possibility of analyzing and deducting more information from network traffic. For a telecommunication operator, this provides means of getting more information of specific subscribers. Data Challenges today are often categorized as “Big” because they deal with one or more of the following. “Big” volume, velocity or variety. While the challenges of analyzing such “big data” are most often discussed, growing volume, velocity and variety of data are produced in social media. The increasing use of social networks, such as Facebook, Twitter and Weibo has produced and is producing huge volume of data. Twitter posts more than 500 million tweets everyday. Weibo is reported to have over 766 million active users per day in 2014. Business firms and other organizations are interested in discovering new business insight to increase business performance.

The Big Data produced by social networks can be analyzed by current computer technologies Map reduce, Hadoop and NoSQL techniques have supported distributed data storage, parallel data retrieval and processing. Many analytical methods and algorithms are designed for business analytics, such as K-means clustering, Association rules, Linear/ Logistic regression and Time series. Many software companies have developed their BDA products. For example, IBM has a series of software to support BDA including Infosphere Puredata, Cognos and SPSS modeler. The applications of this are many, such as segmentation for

marketing purposes or detection of churners, people about to switching operator. Thus the analysis and information extraction is of great value.

OTHER STUDIES

The main idea is to divide a problem into parts and let several machine processes it at the same time, separately. Google is, as of now the pioneer of this approach with their software solutions MapReduce and Pregal. Additionally, solving the problem in parallel using MapReduce or Hadoop implies storing different parts of the data on separate hard drives. Many models for social networks were established based on undirected networks. In this study, we will construct an evolving model of a new directed, Scale-free network on the basis of the BA model. Our propose model adopts the mechanism of preferential attachment during network evolution, which is considered one of the key factors in the formation of scale-free networks. We will test the effectiveness of our model by a simulation using data of a real-world Chinese social network.

By using advanced analytics, enterprises can analyze big data to learn about relationships underlying social networks that characterize the social behavior of individuals and groups. Using data describing the relationships, we are able to identify social leaders who influence the behavior of others in the network, and on the other hand, to determine which people are most affected by other network participants. We can also use diffusion analysis to identify the individuals most affected by the group leaders and target the marketing to them.

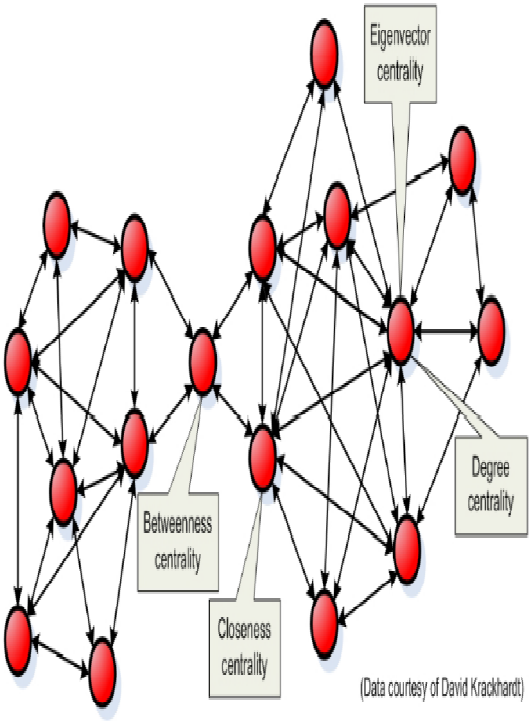
RELATED WORK

Clever algorithms, taking into account the issues of parallelization, are a solution to this. Using the technology of the Hadoop framework and making use of knowledge form the field of machine learning, a prototype for detecting subscribers of particular interest is developed. This uses Call Detail Records (CDR) generated by the operator, for charging purposes as input. A training set is provided to create a decision model for classifying the subscribers. A social network is a set composed of nodes and links between each nodes. The nodes also represent social actors, and the links are relationships. People or organizations as well as their social relations are connected by a social network. A social network structure is the actual or potential relational schema existing between actors. A social network structure can not only help us understand the kinds of gathering ways among a set of actors in a special space but also help us understand the significance of one's behavior.

PROPOSED SYSTEM

First, the general field of social network analysis is visited and the relations to telecommunication networks are discussed. Following that, two possible techniques for performing social network analysis effectively, parallelization and machine learning are studied. Moreover, it introduces a prototypical solution of the whole process of identifying subscribers of particular interest, from the input of CDR's to a list of significant subscribers as output.

The summary of the results obtained when testing the algorithms related to this study. This is divided into parts comprising the pre-study results, the results of test of hadoop in clustered mode and the experiments related to the prototype. Additionally, some result of the distribution of social network analysis metrics in a real telecommunication network is accounted for.



PARALLELIZATION TECHNIQUE

An idea gaining foothold among computer scientists is parallelization. This is basically doing parts of a process separately and in parallel. MapReduce is a framework of parallelizing a computation process on a cluster of computers. A similar open-source project is Hadoop. These systems are frameworks which deal with the parallelization, fault tolerance and scheduling of machine communication for large tasks.

In a virtual community, a social structure is some mode of stable relationships and is always expressed as a network formed by a series of nodes (actors) and links that represent relationships among nodes. There are three basic substances of social network structures: the actor, relation and network which are introduces below.

i)Actor:

Nodes in network are actors. Everything such as a user, a book, or a movie can be an actor in the Virtual community. When doing research on a virtual study community based on a network environment, we can focus on some particulars actors who have a very important influence on the majority of others on one side, and pay attention to the entire network on the other side.

ii)Relation:

A relation is a link of nodes. It is one of the core parts in basic research on virtual learning communities. Nodes (actors) share, transfer and acquire information through direct or potential links (relations). There are three features of relation: content, direction and strength.

Direction could be divided into directed and undirected. In some websites, A can yellow B without B's acceptation. Then, we can say their relation has a clear direction. However in other websites, A and B can be friends only if they follow each other. Therefore, their relation has no direction.

There are weak as well as strong ties between individuals. In light of the trust in a society, the clan trusts exists in strong connections, and the foundation of weak connections is impersonal trust in society. Strong connections can become weak or dissolve due to various reasons. In contrast, weak connections may also become strong with an increase in confidence. In a social network, we distinguish relations strength by the frequency of their interaction.

iii) Network:

A network appears to be a set of relations, describing its relational scheme or connection mode. There are two typically types of network: ego- centered network and whole network.

When one analyzes an ego-centered network, one generally pays attention to special actors, named "Internet stars", and the networks they (actors) establish and relations (links) with their neighbors.The whole network method aims to study relations of all the actors within a certain range. Therefore, a large amount of relational data is of huge importance. This leads to the problem of big calculation. To solve this problem, it is necessary to get help from computer assistance computation.

Sociogram as well as Social Matrix are powerful tools as we seek to describe networks. The former makes the network more intuitive than the latter. However, when there is a large amount of actors, it is difficult to analyze the relational structure within Sociogram. Thus, social matrix seems like the best option.

A very important part of the MapReduce framework is the blocking mechanism. This is the way MapReduce deals with machine failures. If the task is not completed it will be re-scheduled for a different machine. If the task is completed however, it will be treated differently depending on if it is a map task or a reduce task.

MACHINE LEARNING

Machine learning is a part of the field of artificial intelligence. Two main groups can be defined white box and black box methods.

White box: White box methods are translucent in that they reveal the process by which the output is created from the input

Black box: Black box methods are do not inform the user of how an output is produced.

We use white box methods of decision trees and logistic regression, as well as the black box methods of neural networks.

The information era has witnessed an unprecedentedly high speed of data creation and knowledge diffusion. Many models for social networks were established based on undirected networks. In this study, we will construct an evolving model of a new directed, scale-free network. Our proposed model adopts the mechanism of preferential attachment during network evolution, which is considered one of the key factors in the formation of scale-free networks. We will test the effectiveness of our model by a simulation using data of a real- world Chinese social network.

IMPLEMENTATION

There is a challenge for the network operator to know which algorithms to use and how to evaluate the resulting measures from all different algorithms in order to pinpoint the most influential subscribers or other interesting segments of subscribers. In order to deal with this difficulties machine learning algorithm is implemented. The general idea is to generate a social network graph from the information in a CDR-file.

This algorithm is implemented is implemented to created a model for handling each subscriber. For instance, this can be a classification model where each subscrible is classified with respect to an attribute. To be able to create the model, a training set must be provided. The training set is constituted of a subset of the network in which for each subscriber, the pre defined attribute is known exactly.

DISCUSSION AND EVALUATION

The result was done on a single machine, thus a natural extension is to try the solution on a computer cluster. Initially, parts of the cluster are tested and following that the complete cluster. The analysis is presented in the same order. After that, a discussion of the statistical relation between the social network analysis measures obtained from the real tele-communications network are explored. Additionally, the results related to the prototype are looked at.

i)Modelling and visualization of networks:

Visual representation of social networks is important to understand the network data and convey the result of the analysis. Many of the analytic software have modules for network visualization. Exploration of the data is done through displaying nodes and ties in various layouts and attributing colors, size and other advanced properties to nodes. The NSA has been performing social networking analysis on call detail records (CDR's) also known as metadata.

EXPERIMENTAL RESULTS

One of the greatest challenges of analysis graphs by applying graphs by applying parallel computing is to make the algorithms work in parallel. The fact that many analytic tools need access to the whole graph causes problems as information regarding the graph structure will be distributed on the different nodes of the computer

cluster. For that reason, many graph analysis algorithms involves sending important information between the nodes and in the end gathering all relevant and necessary data in one place.

Specifically, the map phase generally consists of a vertex sending some sort of information to its neighbors. In the reduce phase, the information from every neighbor is gathered by a vertex and processed into the output of the iteration.

Socio- centric (whole) network analysis:

- Emerged in Sociology
- Involves quantification of interaction among a socially well defined group of people
- Focus on indentifying global structural patterns
- Egocentric (Personal) network analysis
- Emerged in anthropology and psychology
- Involves quantification of instructions between an individual (called ego) and all other persons (called alters) related (directly or indirectly) to ego
- Make generalizations of features found in personal networks
- Difficult to collect data, so till now studies have been rare

PERFORMANCE ANALYSIS

Machine learning algorithms is more likely to and relations to the training set of classification if a larger number of attributes is used . Either, more attributes could be given in the input examples in the training set of one machine learing algorithm, or the attributes could be split to from inputs to several models. In the later case, the different models would ultimately be combined to yield an improved result.

Knowledge based network analysis:

- Emerged in computer science
- Involves quantification of interaction between individuals, groups and other entities
- Knowledge discovery based on entities associated with actors in the social network

SUMMARY

Testing that framework could be a relevant study as well. Another possible way of speeding up the calculation of the social network analysis metrics considered in this study is to implement a clever partitioning of the input data. This partitioning should ensure that vertices that are closely related ends up in the same map task. This could reduce the amount of data that needs to be send between computers as well as improve the performance of any combiner in use. An investigation of this can be found in for web graphs, but if an effective and computational partition function could be defined, this could be tried in the case of telecommunication network as well.

Applications:

- Viral marketing
- Social Analytics
- Expert Finding
- Image analysis
- Fraud detection

CONCLUSION

A complete solution for analyzing telecommunication networks and deducting subscribers of interest have been developed. Starting from Call Detail Records generated for charging purposes, a social network is created. This network is analyzed to and an array of a total of nine different values for each subscriber, indicating that particular subscribers influence within the network. A supervised learning method is ultimately used to determine a model for deduction of relevant subscribers.

SUGGESTIONS

Furthermore, unsupervised learning might be tested. For example, k mean clustering to see if there are any obvious groups within the network. If there is, then perhaps one could be identified as the influential users. For instance, a supervised learning part could be introduced asking the operator to name a group of desirable

influential users and the group finding most of them can be regard as the influential group.

FUTURE WORK ON SOCIAL NETWORK ANALYSIS

Lastly, when it comes to machine learning algorithms should be tested using a larger training set. The one used in the test of this study was rather small, less than 1% of the complete graph. This could possibly improve the accuracy of the algorithms. To be able to generate training sets of larger sizes, a different model for classifying the subscribers should be developed.

REFERENCES

[1] Yongmin Choi, Hyun Wook Ji, Jae-yoon Park and Hyun-chul Kim, A 3W Network Strategy for Mobile Data Traffic loading. Communications Magazine, 118-123, 2011.

[2] Bob Emmerson, M2M: The Internet of 50 billion devices. WinWin by Huawei, 19-22, 2010.

[3] Christine Kiss, Andreas Scholz and Martin Bichler, Evaluating Centrality Measures in Large Call Graphs. Proceedings of the 8th IEEE International Conference on E-Commerce Technology, 2006.

[4] Muhammad Usman Khan and Shoab Ahmed Khan, Social Networks Identifcation and Analysis Using Call Detail Records. Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, 192-196, 2009.

[5] Angela Bohn, Norbert Walchhofer, Patrick Mair and Kurt Hornik, Social Network Analysis of Weighted Telecommunications Graphs. ePub Institutional Repository, 2009.

[6] Jefrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. Proceedings of OSDI '04: 6th Symposium on Operating System Design and Implemention, 2004.

[7] Apache Hadoop, retrieved 15 November 2011, http://hadoop.apache.org/.

[8] Neo4j: NOSQL For the Enterprise, retrieved 15 November 2011, http://neo4j.org/.