IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Associate Rule Mining for Social Network Data using Map Reduce

Miss Shruti S. Gadgil [1], Prof. L. M. R. J. Lobo [2]

Student M.E, Dept of Computer Science and Engineering, Walchand Institute of Technology, Solapur, India [1]

Associate Professor, Dept of Computer Science and Engineering, Walchand Institute of Technology, Solapur, India [2]

Abstract: Social Network is a network of individuals connected by interpersonal relationships. A social network data refers to the data generated from people socializing on this social media. This user generated data helps to examine several assets of the socializing community when analyzed and mined. Data mining is the process of analyzing data from different perspectives for finding unknown insights. Association Rule mining, one of the significant task of data mining helps in the discovery of associations, correlations in social networks. This paper presents an approach of extracting associations between the contents of Social Network Data using Apriori Algorithm based on MapReduce framework which is a programming model working in parallel form to find frequent itemsets for the Social Network data and then makes use of Genetic Algorithm for generating association rules from the frequent itemsets. Here Genetic Algorithm is used as an optimization technique to generate Association Rules.

Keywords: Data mining, Genetic Algorithm, Apriori Algorithm, Association rules, MapReduce.

I. INTRODUCTION

Few years ago people used to communicate verbally or non verbally. In non verbal communication they used to draft letter or write articles in news paper, magazines, journals, etc. There were no much means of non verbal communication. The advent of internet made the people to receive the information globally in various aspects. In due course of time the use of internet was not only restricted for receiving the knowledge but also stepped in two way communication like sharing the ideas, views, opinions through social media on internet. Social networking has thus escalated around the world with noteworthy speed. As a result the data on these Social Networking sites started bulging. This huge amount of data was necessary to analyze and then necessarily channelize to end users. This has led to develop powerful means for analysis and interpretation of such data together with the extraction of interesting knowledge that could help in decision-making process.

This has given rise to the use of the concept of Data Mining. Data mining is the process of analyzing data from different perspectives for finding unknown insights. Data mining deals with the process of discovering patterns in large data sets. The goal of data mining is to excerpt the information from a data set and remodel it into a structure for future use.

Many of the Social Networking sites such as Facebook, Twitter have millions of active users in present. And these sites have stored millions of data related to user communication, posts, images etc. For example Twitter is one of the trendy social networking site. User communication on Twitter takes place in form of text messages of shorter length called as tweets which are being created and shared at a unique rate. These millions of tweets are being gathered by Twitter every day. These gathered tweets contain information and is likely to contain lots of redundancy [1]. From this example it can be seen that it's very cumbersome to maintain such a huge data. To avoid such complexity and to make system more reliable, we need to optimize the collected information and develop certain ways to analyze the gathered information so that the result of the analysis can be used in the decision making process.

In this paper, we present a developed system that represents an Association mining model for Social Network data, using MapReduce framework and Genetic algorithm to get optimized association rules. By experimentation it is found that the efficiency of the developed algorithm is 39% more than MORA-GA Algorithm and the accuracy of the obtained rules is increased by 25% as compared to optimized association rule mining using genetic algorithm [2].

Association Rule Mining

Association rule mining is one of the significant task of data mining. Association rule mining is a method wherein relations between the various attributes in a transaction or a database are found out. Association rule is generally of the form X->Y where X and Y are itemsets. Here X is called as antecedent and Y is called as consequent. For example, the following rule {bread}->{butter} would indicate that if a customer buys bread , then he is also likely to buy butter. Rules have associated support and associated confidence. Support is a measure of what fraction of the transaction satisfies both antecedent and consequent of the rule. Confidence measures how often consequent are true when antecedent is true. Here, bread

DOI 10.17148/IJARCCE.2016.51029

and milk is support and butter is confidence. Such information can be used for decision making purpose.

The MapReduce Framework

MapReduce is a programming model reintroduced by Google in 2004 to support distributed computing on large datasets on clusters of computers. In the MapReduce framework the input problem is broken into map and reduces phases that are processed by these two in parallel manner. The map function takes a set of key/value pair to generate a set of intermediate key/value pairs. The output key/value pairs of the map phase are sorted by their key and each reducer gets a particular key and a list of all the values belonging to that key. The reduce function then merges all intermediate values associated with the same intermediate key.
The user specified map and reduce functions are of the following type[1]:
Map(k1,v1)→list(k2,v2)
Reduce(k2,list(v2))→ list(v2)

Genetic Algorithm

A Genetic Algorithm (GA) is a heuristics search technique used in computing to find exact or approximate solutions to optimization and search problems. The basic principles of Genetic Algorithms were first laid down by John Holland at University of Michigan in the United States in the year 1970[3]. Genetic Algorithm uses genetic operators such selection, crossover and mutation. Genetic Algorithm runs to generate solutions for successive generations [1]. The functions of these operators are as follows:

- Selection: This operator decides which individuals to select for reproduction and which one to preserve.
- Crossover: Crossover produces new elements in the population called as off springs by combining parts of two elements called as parents currently in the population.
- Mutation: Modifies one or more gene values of the individual in order to find for better solutions

II. LITERATURE REVIEW

Rakesh Agrawal, Tomasz Imielinski, Arun Swami [4] focused on mining association rules between Sets of Items in Large Databases. They have worked upon customer transactions. They have divided the problem of generating association rules into two sub-problems. The first sub-problem finds out frequent itemsets using a threshold support value and the second sub-problem deals with the generation of association rules form these frequent itemsets found in the previous step. Their proposed algorithm makes multiple passes over the database to find frequent itemsets. In each pass candidate itemsets along with its support are generated from the tuples in the database. At the end of a pass, the support for a candidate itemset is compared with minSupport to determine if it is a large itemset.

Rakesh Agrawal, Ramakrishnan Srikant[5] developed two algorithms for finding association rules between items in a large database of sales transactions. These algorithms are named by them as Apriori and AprioriTid.Their proposed algorithms find large itemsets making multiple passes over the data.They proposed a concept of seed set for generating new large itemsets,called candidate itemsets, and count the actual support for these at the end of the pass until no new large itemsets are found.
J. Han, J. Pei, and Y. Yin [6] worked on mining frequent patterns without candidate generations, and developed an efficient FP-tree-based mining method called as FP-growth for mining the frequent patterns based on the concept of fragment growth. They tackled the problem in following three aspects. First a data structure, called frequent pattern tree or FP-tree is constructed, where only frequent length 1 items will have nodes in the tree. Second, they developed FP-tree-based pattern fragment growth mining method, which starts from frequent length-1 pattern, examines only its conditional pattern base and then constructs its FP-tree, and performs mining re-cursively with such a tree. Third, the search technique employed in mining is a partitioning-based, divide-and-conquer method rather than bottom-up approach.
S. Cong, J. Han, J. Hoeflinger, and D. Padua [7] focused on development of a sampling-based framework for parallel data mining. They presented a strategy for mining frequent itemsets from terabyte-scale data sets on cluster systems. The algorithm includes the holistic notion of architecture conscious data mining, taking into account the capabilities of the processor, the memory hierarchy and the available network interconnects. The solution proposed contains one of the fastest known sequential algorithms (FPGrowth), and extends it to work in a parallel setting, utilizing all available resources efficiently.
Jongwook Woo and Yuhang Xu [8] worked on a market basket analysis algorithm with Map/Reduce for Cloud Computing. They presented Market Basket Analysis algorithms with Map/Reduce, which proposes the algorithm with (key, value) pair and execute the code on Map/Reduce platform. Their proposed methodology sorts the transactions in the alphabetical order before generating key/value pair in order to avoid redundancy. Their proposed algorithm adopts joining function to produce paired items.
Zahra Farzanyar, Nick Cercone [9] proposed “Efficient Mining of Frequent itemsets in Social Network Data based on MapReduce Framework”. In their paper, an Improved MapReduce based Apriori algorithm is proposed for efficient mining of frequent itemsets.Their proposed work aims to reduce the number of the partial frequent itemsets produced during the Map and Reduce phase hereby improving the processing time.
D. Kerana Hanirex and K.P. Kaliyamurthie [10] focused on mining frequent itemsets using genetic algorithm. This paper proposes the use of Genetic Algorithm (GA) to improve the efficiency of finding frequent itemsets. In their methodology an initial population is created consisting of randomly generated transactions. Their

DOI 10.17148/IJARCCE.2016.51029

proposed genetic algorithm based method for finding frequent itemsets repeatedly transforms the population by executing the steps of fitness evaluation, selection, recombination and replacement.

Reza Sheibani and Amir Ebrahimzadeh [11] have proposed a novel mining algorithm named Improved Cluster Based Association Rules (ICBAR) which can explore efficiently large itemsets.Their presented method prunes considerable amounts of data by comparing with the partial cluster tables and also reduces the number of large candidate itemset.

III.METHODOLOGY

As methodology, Enhanced Apriori MapReduce using Genetic Algorithm [EAMRGA] is implemented. The algorithm is broadly divided into two stages.

In the first stage Apriori Algorithm which is based on MapReduce Framework is applied on the processed Social Network Data. This phase of the EAMRGA implements Apriori based MapReduce Algorithm in three steps. The first step of this phase calculates single item support for the processed data. The second step is a pruning step. The Map task of this step is given a minimum support value which acts as a threshold to prune the infrequent item sets from the step 1 while the reduce task of this step emit the result of the Map phase. In the third step the Map task diffuse the transactions while the Reduce task calculates the support of the frequent itemsets hereby resulting into frequent itemsets with their support.

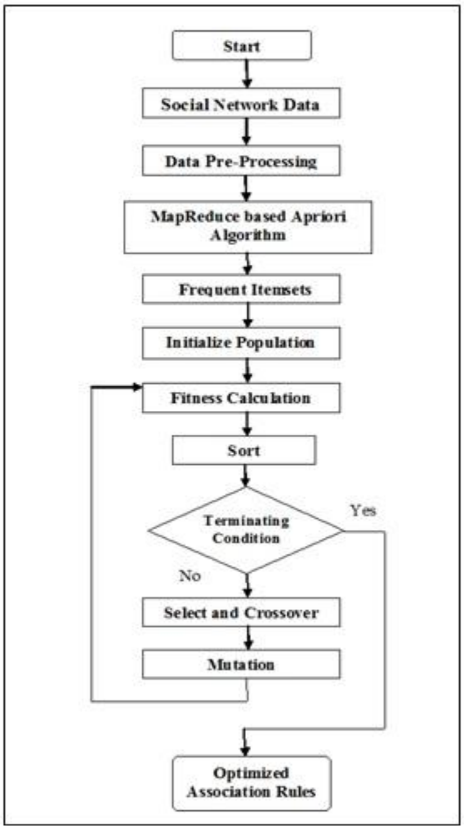

Fig. 1. System Flow

The Second stage of the EAMRGA implements Genetic Algorithm (GA).The frequent itemsets found by the MapReduce based Apriori phase of the EAMRGA are used as input by the GA phase of the EAMRGA.

The flow of the system is as shown in Fig.1. Here based on previously found frequent itemsets, an initial population is created and its fitness is calculated.

This process continues until the terminating condition is met. If the terminating condition is not meet then the GA performs selection and crossover, mutation on the population. Again the individual fitness of new population is evaluated and the least-fit population is replaced by the new one. This process continues until a terminating condition is met resulting into optimized Association Rules. Thus we are left with optimized Association Rules.

IV. EXPERIMENTAL SETUP AND RESULTS

The system was developed using Hadoop 2.7.1 running on stand-alone mode on a single machine running Ubuntu 14.04 operating system. Both of the algorithms MapReduce based Apriori Algorithm and Genetic Algorithm are implemented using JAVA with Jdk version 1.8.

To evaluate the performance of the system empirical studies were conducted on BMS-POS dataset .In this experiment we compare our Algorithm with IMP Apriori Algorithm in the primary factor execution time.

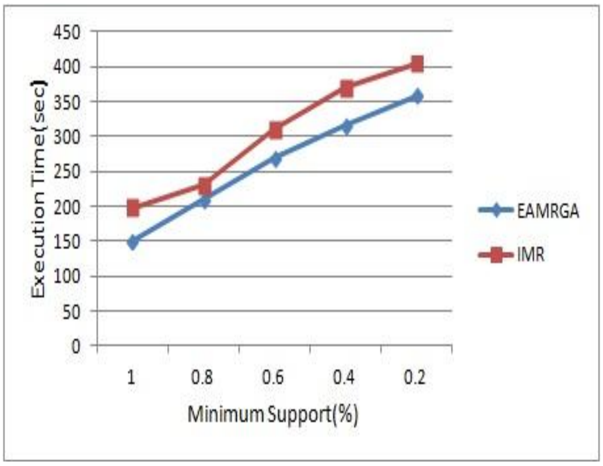

Fig. 2. Result Analysis

Fig. 2 provides runtime efficiency comparison between IMR Apriori Algorithm and the proposed system. It can be seen from Fig. 2 that the algorithm performs 36% up well with respect to execution time.

Another experiment was conducted to investigate the efficiency of the proposed method. The testing was performed using the real time dataset and comparison was done between our proposed algorithm and MORA-GA Algorithm. The EAMRGA produces more optimizes rules with higher confidence value as compared to MORA-GA. It can be seen from the charts in Fig. 3 and Fig. 4 that our proposed algorithm is 39% more efficient than MORA-GA Algorithm.

DOI 10.17148/IJARCCE.2016.51029

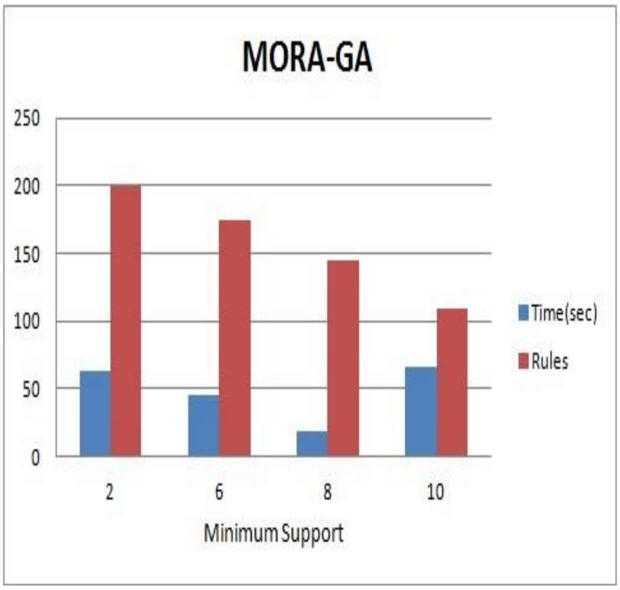

Fig. 3. MORA-GA Result

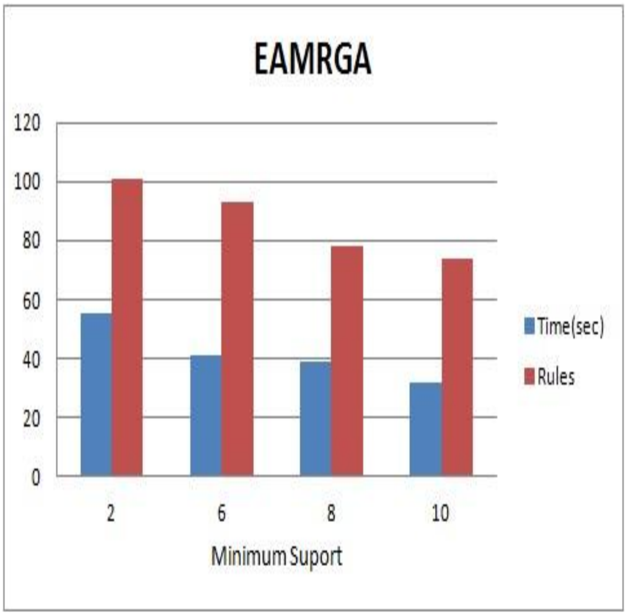

Fig. 4. EAMRGA Result

V. CONCLUSION AND FUTURE SCOPE

The Social Network sites are becoming increasingly popular now a day. Social sites have tremendous data. So, mining of data is very useful. A system is developed to exhibit a parallel programming model. The model aims at finding association rules from such a data-rich environment by using efficient algorithm based on MapReduce framework. Genetic algorithm will be used for optimizing the item sets and finding optimized and relevant association rules. As a future work to this idea proposed would require handling data in range of terabytes or more and reduction in processing time which could be handled by a parallel or hierarchical approach encountering more extensive use of the features provided by Hadoop.

REFERENCES

[1] Gadgil S.,Lobo L.M.R.J,“MapReduce to Find Association Rules Representing Social Network Data”,IJCA RTDM 2016 Proceedings.

[2] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K.,” Optimized association rule mining using genetic algorithm”, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04

[3] http://lancet.mit.edu/mbwall/presentations/IntroToGAs

[4] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.

[5] Agrawal and R. Srikant:”Fast Algorithms for Mining Association Rules in Large Databases”. In: Proceedings of the Twentieth International Conference on Very Large Databases.

[6] J. Han, J. Pei, and Y. Yin.” Mining frequent patterns without candidate generations”. In Proceedings of the International Conference on Management of Data, 2000.

[7] S. Cong, J. Han, J. Hoeflinger, and D. Padua. “A sampling-based framework for parallel data mining”. New York, NY, USA, 2005. ACM.

[8] Jongwook Woo and Yuhang Xu, “Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing”, Las Vegas, July 18-21, 2011.

[9] Zahra Farzanyar, Nick Cercone “Efficient Mining of Frequent itemsets in Social Network Data based on MapReduce Framework”, 2013 IEEE International Conference on Advances in Social Networks Analysis and Mining.

[10] D. Kerana Hanirex and K.P. Kaliyamurthie,” Mining Frequent Itemsets Using Genetic Algorithm”, Middle-East Journal of Scientific Research 19 (6): 807-810, 2014 ISSN 1990-9233 © IDOSI Publications, 2014.

[11] Reza Sheibani and Amir Ebrahimzadeh,“ICBAR: An Efficient Mining of Association Rules in Huge Databases”, International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012.

DOI 10.17148/IJARCCE.2016.51029