MapReduce to Find Association Rules Representing Social Network Data

Shruti S. Gadgil
Student M.E
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur, India

L.M.R.J. Lobo
Associate Professor
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur, India

ABSTRACT
Social Network is a network of social involvements and personal relationships. Social Networks involve information sharing between people at all times which results in producing large amount of data produced in this social network environment which can be extremely useful. As social networks are increased, its storage also increases. By observation, it has been discovered that most of social sites have redundant, noisy data. To get such optimized information, Social network analysis focuses on mining out the pattern of user's interaction. For such mining the paper proposes to implement Mining of association rules which helps in the discovery of associations, correlations, statistically relevant patterns, causality, emerging patterns, and other data mining tasks in social networks. Most of the traditional frequent item set mining algorithms is ineffective due to either enormous resource requirements or large communications overhead. Cloud computing has shown that processing very large datasets over clusters can be done by providing the right programming model. As a programming model working in parallel form, Map-Reduce, one of techniques for cloud computing, has emerged in the mining of datasets scaling from terabyte or larger on clusters of computers. The present paper focuses on making use of a proposed algorithm for association rule mining employing the MapReduce frame of reference which deals with Hadoop, a parallel store and computing platform. This will help to improve efficiency and accuracy of the given system.

General Terms
Data mining, Genetic Algorithm

Keywords
Association rules, MapReduce.

1. INTRODUCTION
Data mining deals with the process of discovering patterns in large data sets. The goal of data mining is to extract the information from a data set and transform it into a structure for future use.

Using some of data mining techniques, the proposed methodology aims to develop mining rules to optimize the data.

1.1 Overview
Nowadays, millions of people use the internet to express their ideas and share information. Most use Mail service for sharing and storing data, Social sites, blogs for connecting with people and sharing information. Social network analysis [SNA] is the calculating and aligning of relationships and flows between people, groups, users and other connected information or knowledge entities. Social networking is spread around the world with remarkable speed. Most of the Social Networking sites such as Facebook have more than 845 million active users in February 2012. And it has stored millions of data related to communication, posts, blogs etc. Another example Twitter is one of the very popular blogging sites. Small-text messages called as tweets are being created and shared at a unique rate. Twitter collects millions of tweets every day, which contain lots of noise and redundancy & non structural tweets. From the above two examples, it can be concluded that it's very difficult to maintain such a huge data. To avoid such complexity and to make system more reliable, it is imperative to optimize the collected information. This paper, proposes to develop an Association mining model for optimizing the data, using MapReduce framework and Genetic algorithm to get optimized association rules.

1.2 Association Rule
Mining of Association rule is a popular method for finding interesting relations between variables in large databases. For example, the rule {bread, butter}->{milk} would indicate that if a customer buys bread and butter together, they are likely to also buy milk. Here, bread and butter is support and milk is confidence. Such information can be used for decision making purpose. It plays an important role in many Data Mining tasks that try to find interesting patterns from databases, such as mutual relationship, ordering, episodes, classifiers and clusters.

2. LITERATURE REVIEW
Rakesh Agrawal, Tomasz Imielinski,Arun Swami[1] focused on mining association rules between Sets of Items in Large Databases. They brought into use association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example ,the rule {onions, potatoes}->{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used for decision making purpose.

Rakesh Agrawal, Ramakrishnan Srikant[2] developed the Apriori algorithm which is one of most well-known methods for mining frequent itemsets in a transactional database. The algorithm works within a multiple-pass generation-and-test framework, comprising the joining and the pruning phases to reduce the number of candidates before scanning the database for support counting.

J. Han, J. Pei, and Y. Yin[3] worked on mining frequent patterns without candidate generations, a novel frequent pattern tree(FP-tree) structure was developed , which is an extended prefix-tree structure for storing complex, critical facts about frequently occurring patterns and developed an

efficient FP-tree-based mining method FP-growth for mining the complete set of frequent patterns by fragment growth.

S. Cong, J. Han, J. Hoeflinger, and D. Padua [4] focused on developing a sampling-based framework for parallel data mining. This paper makes use of pattern growth algorithm to solve the problems sequentially related to frequent itemset mining and sequential pattern mining. They presented a framework for parallel mining of frequent itemsets as well as sequential patterns making use of divide and conquer strategy for the pattern growth. They proposed a sampling technique known as selective sampling to address the problem related to load balancing.

W. Fang, K. K. Lau, C. K. Lam, Y. Yang, B. He, Q. Luo, P. V. Sander [5], proposed a novel parallel data mining system making use of new generation graphics processing units(GPUS).The system that relied on multi-threaded SIMD (Single Instruction, Multiple-Data) architecture given by GPUs. The GPUMiner designed by them consisted of the following three components: (1) a CPU-based storage and buffer manager, (2) a GPU-CPU co-processing parallel mining module, and (3) a GPU-based mining visualization module. The GPUMiner is implemented based on the k-means clustering and Apriori frequent pattern mining algorithm.

L. Liu, E. Li, Y. Zhang, and Z. Tang [6] reduced the complexity by implementing the optimization method for multi core processor in their paper, two techniques were developed to deal with utilization of multi-core system by the current FP-tree based algorithms: a cache-conscious FP-array (frequent pattern array) and a lock-free dataset tiling parallelization mechanism to address this problem.

E. Ozkural, B. Ucar, and C. Aykana.[7] developed parallel frequent itemset mining with selective item replication. They proposed a transaction database distribution scheme which splits the frequent item set mining task in a top-down fashion. Their method works on a graph where vertices correspond to frequent items and edges correspond to frequent itemsets of size two. The study indicated partitioning this graph by a vertex separator is sufficient to decide a distribution of the items resulting in the independent mining of sub databases This strategy is used in the design of algorithms: NoClique replicates the work induced by the separator and NoClique2 computes the work conjointly.

Jongwook Woo and Yuhang Xu [8] worked on a market basket analysis algorithm with Map/Reduce framework. 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. The algorithm makes use of joining function to produce paired items.

Zahra Farzanyar, Nick Cercone [9] developed 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.

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.

3. PROPOSED METHODOLOGY

3.1 MapReduce

A MapReduce framework job splits the input data-set into independent chunks that are processed by the map task in a parallel fashion. The function of the mapper is to map the job into key and value. The framework sorts the outputs of the map tasks, which are then given as input to the reduce tasks. The reduce function integrates all the intermediate values associated with the same intermediate key. Then frequent itemsets are generated.

The user specified map and reduce functions are of the following type:

Map(k1,v1)→list(k2,v2)

Reduce(k2,list(v2))→ list(v2)

3.2 Genetic Algorithm

Genetic Algorithm uses genetic operators such selection, crossover and mutation. Genetic Algorithm runs to generate solutions for successive generations. The functions of these operators are as follows:

- Selection: Use a fitness function to evaluate the current solution.
- Crossover: Crossover produces new elements for the population by combining parts of two elements currently in the population.
- Mutation: Modifies the new solutions in order to find for better solutions.

The fitness function is implemented in order to calculate the fitness of the individuals and to decide which the best candidates are in the following generations.

3.3 Proposed Architecture

The system architecture of the proposed work is as shown in Figure.1.

The proposed system works as follows:

Input: Untagged tweets:

Output: optimized association rules.

Step 1: Getting tagged tweets and generating Itemsets:

- **Tweet Collector**

The tweet collector sends request for the tweets and in response sends the tweets to the DB Module to be stored in the database.

- **DB Module**

The DB Module performs the task of storing the tweets collected by the tweet collector responding to the request of retrieving the tweets for processing, storing the tagged tweets as processed by the tweet tagger.

- **Tweet Tagger**

The tweet tagger takes untagged tweets from the database and taking tagger training data as an input to the classification model classifies the untagged tweets and stores the tagged tweets into the database through the DB Module.

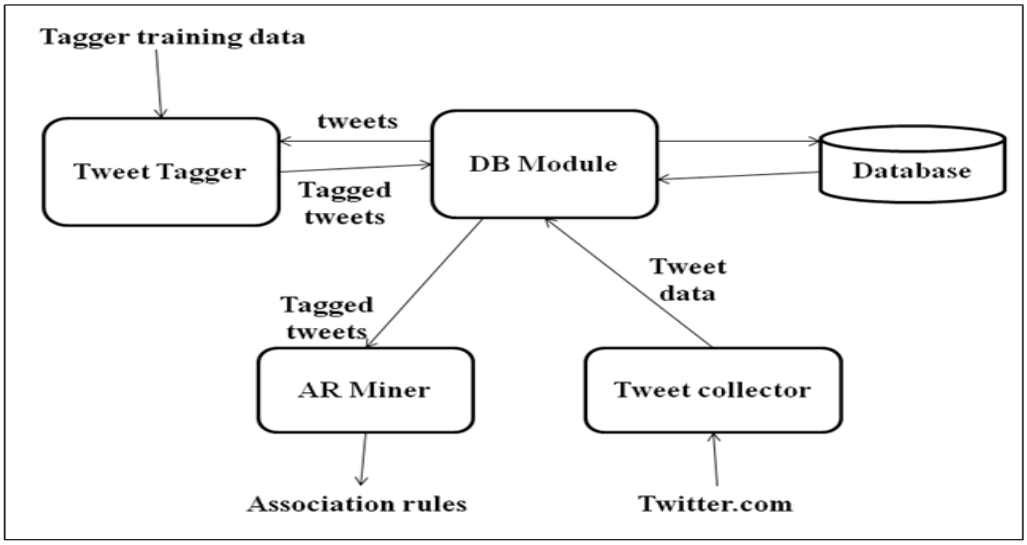

Fig 1: Proposed System Architecture

Step 2: Apply Map reduce algorithm

- **AR Miner**

The AR Miner takes the tagged tweets as an input and implements the Apriori algorithm based on Map Reduce framework to generate association rules. These association rules are then optimized using Genetic Algorithm resulting into optimized association rules for the Social Network Data. The flow of the proposed Apriori Algorithm based on MapReduce Framework is as follows:

Step 1: Calculate single item support

Step 2: Remove infrequent items from the

Transactions

Map task: remove infrequent items using

Minimum support.

Reduce task: emit transactions.

Step 3: Map task: Diffuse transactions.
Reduce task: Calculate support.

Step 3: Apply Genetic algorithm

Using Genetic Algorithm the fitness value for the generated itemsets are calculated and a new population is produced based on crossover and mutation. Again the individual fitness of new population is evaluated and the least-fit population is replaced by the new one. This process continues until terminating criteria is met. Thus, an optimized rule is developed.

4. EXPERIMENTAL SETUP

4.1 Hardware Resources

1. RAM:Minimum 4GB
2. Processor:Minimum 2.01 GHZ
3. HDD(storage):500GB or more

4.2 Software Resources

1. Language:Java
2. JDK:1.8
3. Database:My SQL
4. Apache Hadoop 1.0/2.0
5. Operating System: Linux

5. CONCLUSION

The Social Network sites are more popular now a days. Social sites have tremendous data. So, mining of data is very useful. The data produced in such environment grows dynamically at highest range in terabytes or more. As a result basic algorithms are inefficient to handle such large data. Thus, it is proposed to develop a parallel programming model. The proposed 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 it would be possible to go for a multimedia approach which would make things more realistic. It would also require reduction in processing time which could be handled by a parallel or hierarchical approach encountering more extensive use of the features provided by Hadoop.

6. REFERENCES

[1] B 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.

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

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

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

[5] W. Fang, K. K. Lau, C. K. Lam, Y. Yang, B. He, Q. Luo, P. V. Sander, "Parallel data mining on graphics processors", The Hong Kong University of Science & Technology, 2008.

[6] L. Liu, E. Li, Y. Zhang, and Z. Tang. "Optimization of

frequent itemset mining on multiple-core processor". In Proceedings of the 33rd international conference on Very large data bases, VLDB '07, pages 1275–1285. VLDB Endowment, 2007.

[7] E. Ozkural, B. Ucar, and C. Aykanat." Parallel frequent item set mining with selective item replication". Parallel and Distributed Systems, IEEE Transactions on, oct. 2011.

[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.