International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

AN EFFICIENT APPROACH FOR FREQUENT ITEMSET MINING IN BIG DATA

R. SANGEETHA
M.Phil Research Scholar
Department of Computer Science

Mr. S.AMARESAN MCA., M.Phil., M.Tech
Associate Professor
Department of Computer Science
PRIST UNIVERSITY

PONNAIYAH RAMAJAYAM INSTITUTE OF SCIENCE & TECHNOLOGY [PRIST]
THANJAVUR – 613403 – TAMIL NADU

Abstract- The concepts of Data Mining are one of the most important of item set. This two decades many research works have been done in Frequent Item set Mining. And it becomes a very difficult task while they are applied to Big Data. Too Many efficient pattern design in data mining algorithms have been discovered in last two decades.

In terms of memory as well as execution speed, tree based Pattern growth algorithm considered as most efficient other Frequent Item set Mining (FIM) methods. Another important thing is considered in Frequent Itemset Mining is often generates a very large number of item sets, and reduces not only an efficiency but also the effectiveness of mining. So Constraint based FIM have been proved to effective in reducing the search space in the FIM task and its improves the efficiency. Above drawbacks an efficient algorithm called as Modified FP Growth has been proposed to mine Frequent Item sets of big data. In this mining algorithm, Map Reduce concept as used to find Frequent Item sets from Big Data set. In each and every Data Node as

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

Frequent 1-itemsets are generated using a new tree structure called support count tree.

This support count tree can easily be embedded into any of the existing algorithms aimed at FIM. With help of this tree Frequent 1-Itemsets are found with efficiently and quickly and in-turn speeds up the generation of entire database for frequent Itemset. In additions, to still increase more efficiency of Map Reduce task a cache has included in Map phase to maintain support count tree for calculating the Frequent-1 item sets of each mapper. This reduces the calculating total time of Frequent-1 item sets since it bypasses the sort and the combine task of each Mapper in the original Map Reduce tasks. This turn reduces the total time of execution generating Frequent Item sets of the entire database.

Keywords: Data Mining, Frequent Itemset, Constraints, Hadoop, Map Reduce, support count, Frequent 1-itemsets, patterns, cache

Introduction:

Patterns becomes a very difficult task when they are applied to Big Data. Data storage has increased exponentially in the world over the past few years. Data coming from different sources such as web logs, machine logs, human-generated data, etc. are being stored by companies. This phenomenon is known as "Big Data" and nowadays it is trending everywhere. With the incredibly fast growth of data, comes the need to analyze the huge amount of data.

Due to lack of adequate tools and programs, data remains unused and underutilized because many important knowledge which is useful to the mankind remains hidden. Recent improvements in the field of parallel programming have provided good tools to tackle this problem.

Some of the real-life applications to which different Data Mining techniques can be applied are (i) forming groups people based on their interests of the people or grouping similar constraints based on their properties (clustering), (ii) categorizing new insurers based on records of similar old claimants (classification) and (iii) detecting unusual credit transactions.

Frequent Pattern Mi

Besides clustering, classification and anomaly detection, frequent pattern mining and association rule mining are also important because the latter two analyses

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

valuable data (e.g., shopper market basket data) and help shop owners/managers by finding interesting or frequent itemsets that reveal customer Purchase behavior.

Big Data

The data can be a structured, semi-structured and unstructured data which can be mined for information. Here the definition uses size or volume of data as the only criterion. Another interesting definition of Big Data is that it is a technology designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.

This definition is based on the 3Vs model coined by Doug Laney in 2001. He did not use the term Big Data but predicted that data management will get more and more important and difficult. He then identified the 3Vs which is shown in Figure 1.2 (data volume, data velocity and data variety) as the biggest challenges of data management. Data volume is the size of the data, data velocity is the speed at which data arrives and data variety is the data extracted from different sources which can be unstructured or semi-structured.

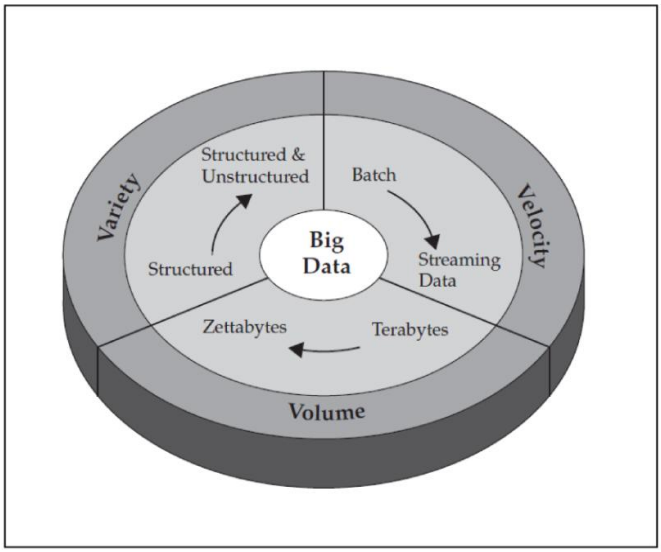

Figure 1.2 Big Data characterization

Data Volume

The size of the data is growing at a rapid speed nowadays. A text file is a few kilobytes; a sound file is a few megabytes while a full-length movie is a few gigabytes. These data comes from many sources. In the olden days all data are generated within the company itself by the employees. But the data are now generated by employees, partners and customers. In addition to this, machines also generate data for a company which has many

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

branches. For example, hundreds of millions of smart phones send a variety of information to the network infrastructure. Handling such huge volume of data is obviously the most widely recognized challenge. Processing huge volume of data is not an issue if the data is loaded in bulk and there is enough processing time. But handling small amount of data gets problematic if the data is unstructured, arriving at high velocity and also need to be processed within seconds. From this it is very clear that it is not the volume which decides whether the data is big but also other data characteristics like velocity and variety must also be considered.

Data Velocity

The speed at which data arrives is known as velocity. There are two folds of data velocity. First, the rate at which new data are flowing in and existing data is getting updated called the „acquisition rate challenge". Second, the time acceptable to analyze the data and process while the data is on move, called „timeliness challenge". The first problem is data acquisition rate. The challenges involved in this is of how to receive, filter, manage and store the continuously arriving data. Traditional relational database systems are not suitable for this task as they process many overhead in the form of locking, logging, buffer pool management and formulating threaded operations. One way to handle this problem is to analyze the flow of data for anomalies and discard unnecessary data by filtering and only store important data. This filtering of data stream without missing any important data not only requires an intelligent tool, but also consumes time and resources. Also it is not always possible to filter data. Another necessary task is to automatically extract and store metadata together with the streaming data. This is useful to track how data is stored and measured.

The second problem is regarding reaction to incoming data streams. In many situations real time analysis becomes necessary, otherwise the information gets useless. As mentioned earlier, it is not only sufficient to analyze the data and extract information in real time but also necessary to react on it and apply. The speed of handling the whole case becomes the decisive issue. For

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

mining of data streams not only speed is important but also the time interval of the actual processing of data. This is very important with respect to data streams with characteristic feature of evolving over time.

Data Variety

Variety refers to the diversity of data sources. Data comes from different sources in various types such as web data, social media, machine generated logs, human-generated data, biometrics, transactional data, etc. This not only implies an increased amount of data sources but also structural differences among those data sources. Furthermore, the structure or schema of different data sources is not necessarily compatible and also the semantics of data can be inconsistent. Therefore, managing and integrating of multi-structured data from a wide variety of sources poses many challenges. First comes the storage and management of this data in a database like systems. Relational database management systems may not be suitable for all types and formats of data. The next challenge is related to the semi and fully unstructured character of data. In the context of integrating different data sources, different data, be it structured unstructured or semi-structured needs to be transformed to some structure or schema that can be used to relate different data sources.

Need for the study

FPM has proved to be one of the promising fields in carrying out the research work because of its wide use in all Data Mining tasks such as clustering, classification, prediction and association analysis. Mining frequent itemsets enables humans to take better decisions in a wide range of applications including market basket analysis, traffic signals analysis and in Bioinformatics identify frequently co-occurring protein domains in a set of proteins. Many researches have proposed many algorithms to generate FIM, but the execution time and storage space plays a key difference in all these algorithms. Pruning unimportant patterns becomes another important research area in FIM.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

Now it is an era of Big Data. There are some applications where frequent patterns have to be extracted from Big Dataset. One such example is Web Log Mining, which helps us to identify frequent web pages visited by the user. By using this information one can improve their advertising process. To handle Big Dataset parallel mining becomes necessary for which MapReduce concept can be used.

Problem Statement

Generating all frequent itemsets is typically very large which some applications do not require. The subset that is really needed by these applications usually contains only a small number of itemsets. Thus more time is spent in considering all unwanted frequent itemsets to extract frequent itemsets. In addition to this, memory is also wasted in storing all unimportant frequent itemsets. So constraints can be introduced to remove these unimportant itemsets. All the existing algorithms hold good only when the dataset is small. So there is a need to propose an efficient algorithm to find frequent item sets from Big Dataset using constraints. In almost all FPM algorithms, Frequent 1-itemsets are generated to find the support count (occurrences) of each item in the entire database. This task is itself a tedious task in generating Frequent itemsets when considering the hugeness of modern databases available.

Conclusion

As the world now is in its information era, massive volume of data is accumulated every day. A real universal challenge is to extract knowledge from extremely large datasets.

In this research work, the main focus is on the problem of Mining Frequent Itemsets from Big Data efficiently and effectively. Due to the fact that a large number of patterns or rules are often found in FIM, Constrained Frequent Itemset algorithm has been proposed and it is implemented in Big Data using MapReduce task in Hadoop. Most of the FPM algorithms spend half the time in generating Frequent 1-itemsets. A simple and easy to implement support count tree algorithm has been proposed which has reduced the time of

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 7, Issue 1, No 1, 2019. ISSN: 2348-6600
Reference ID: IJCS-350 PAGE NO: 2352-2358

generating frequent 1-itemsets. This algorithm can easily embedded into any of the existing algorithms aimed at association rule mining in order to extract Frequent 1-itemsets and their corresponding counts. To still more reduce the execution time of extracting Frequent Itemsets from Big Data using MapReduce, a modified MapReduce has been proposed. In this a cache has been included in the map phase to maintain support count tree for calculating the frequent-1 itemset of each mapper. This reduces the total time of calculating Frequent-1 itemsets since it bypasses the shuffle, sort and the combine task of each Mapper in the original MapReduce tasks. Another feature of finding cumulative frequent itemsets from multiple files is also added.

Thus, efficient algorithms for mining Frequent itemsets from Big Data has been proposed and proved to be an efficient than the already existing algorithms. Hadoop with its map-reduce programming paradigms, overall architecture, ecosystem, fault- tolerance techniques and distributed processing, Hadoop offers a complete infrastructure to handle Big Data. Users can fully utilize the useful information of Big-Data by adopting Hadoop infrastructure for data processing.

References

1. Agrawal, R. and Srikanth, R. "Fast algorithm for mining associationrules", International conference on Very large databases, 1994.

2. Bhatia, Priyaneet and Gupta, S."Correlated Appraisal of Big Data, Hadoop and MapReduce", Advances in Computer Science: an International Journal, Vol. 4,2015.

3. Bonchi, F., Giannotti, F., Mazzanti, A. and Pedreschi, D. " ExAMiner: Optimized level wise FPM with monotone constraints", In Proceedings of International Conference on Data Mining, 2003.

4. Choubey, A., Ravindra, P. and Rana, J. L. "Graph based new approach for FPM ", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 4, No. 1, 2012.

5. Dong, Jie and Han, M. "BitTableFI: An efficient mining frequent itemsets algorithm", Knowledge based Systems, Elsevier, 2006.