International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1709-1715

http://www.ripublication.com

Sequential Rule Mining, Methods and Techniques: A Review

Amanjeet Kour
Department of Computer Science
Rungta college of Engineering and Technology
Bhilai (C.G), India.

Abstract

In the present era of software technologies the amount of data is increasing with every hour. Knowledge extraction or more accurately data mining from such a large data is a great challenge in itself . Although we have various techniques of knowledge discovery still in the field of data mining, sequential rule mining is considered as one of the most vital technique. There are numerous algorithms yet being discovered for sequential rule mining. our study in the field of data mining focuses on techniques from frequent pattern mining to sequential pattern mining followed by sequential rule mining.

INTRODUCTION

With the increase in technology in this era of information, the amount of data to be stored and maintained is also increasing. almost every field generate large amount of data such as financial transactions, marketing data, scientific data, genomics, from satellites and so on. These large datasets contains both useful and waste data. Finding useful relevant information form such a vast data is a challenging task. Here comes the concept of data mining. The data can be from simple numeric data to text, pictures, video or much more complex data. We require powerful means to extract useful information from such a vast data which is called data mining or knowledge discovery.This data mining is done through various steps – Data cleaning, Data integration,Data selection,Data Transformation,Data mining, pattern evalution and Knowledge representation.

Pattern mining is an important subfield of data mining. Generally, frequent pattern were mined mostly updating the technique as frequent pattern mining. Frequent pattern mining is the extraction of frequent itemsets from a list of item sets or subsequences that come into sight is a data set all of whose frequency is greater than a threshold value. This threshold value is specified by the user. Apriori algorithm –A pattern mining algorithm is a well known pattern mining algorithm. When it comes to sequential datasets another subfield of data mining comes up i.e sequential data mining. Sequential data mining is another specialized data mining task for sequential data which extracts useful sequential patterns from a large set of sequences. If the technique is used to find frequent patterns the it is more accurately termed as frequent sequential pattern mining. The prerequisite for sequential pattern mining is a sequential data base and a parameter termed as *minsup* – minimum support threshold. *Minsup* is a minimum threshold value which specifies minimum frequency of a particular pattern to occur in the sequential dataset to be considered as of interest. Algorthms yet introduced for this technique are PrefixSpan, Spade, SPAM, GSP and still a lot of work is being done in this part of mining. It was seen that frequent sequential pattern mining could be misleading so came up the concept of sequential rule mining which also took into account the probability that the pattern will be followed. A sequential rule is of the form X -> Y such that X and Y are sets of items which means that if an item from set of items of X occurs then some item from set of items of Y will also occur.

Big data is nothing but such large and complicated that data that can no more be processed using traditional processing or mining techniques. Big data was very firstly defined using three V's, "Volume,Velocity,Variety" later on it changed to "Veracity, Variability, Visualization, and value" and further it expanded to six V's as follows[27] :

Volume : referring to size from terabytes to petabytes

Variety: data can be structured, unstructured, text, audio, video and many more.

Velocity: In real time data is getting generated rapidly, its analysis is also required to be very fast.

Veracity: requires correct form of data.

Variability: data of same form but different semantics.

Visualization: data should be easy to process to bring out intelligence.

There were a number of techniques applied to big data for analysis and knowledge extraction most commonly used is mapreduce of hadoop framework as it is an efficient parallel algorithm for processing large data. The Mapreduce divides the large processing problems into numerous small nodes where they can be parallel processed with in fraction of time. It also hides the parallelization and data

distribution problem from the user so that the user. Mapreduce comprises of two highly important functions i.e *map* and *reduce.* Map takes as imput a key/value pairs and produces a set of intermediatory key/value pairs then the mapreduce library groups up the values in association with intermediate key *"I"* which is passed to reduce function with a set of values for that key[9]. The reduce merges and produces the smaller set of value may be just zero or one output value[9].

LITERATURE REVIEW

Data mining

As it is known that data mining is an admired means for handling gigantic amount of data so, Ms. Aruna J. Chamatkar, Dr. P.K. Butey[15] in their research article described about challenges and application areas of data mining. Some of them are mining sequence data and time series data, mining complex knowledge from complex data, mining in a network setting, distributed data mining, data mining for biological problem and so on[15]. They also gave the application areas as health care and health care system, manufacturing engineering, makrte basket analysis and much more[15].

Collection of large and complex, structured and unstructured data sets were termed as Big Data. Data mining for big data again became a challenge . Bharti Thakur, Manish Mann gave an overview for types of big data and future challenges regarding it[23].They divided data mining into six activities –Classification, Estimation, Prediction, Association rule, clustering and description[23]. They gave three major challenges of big data. Those are: (i)privacy security and trust.(ii)Data management and sharing. (iii)Technology and analytical system[23].Where as Pravin Anil Tak, Dr. S. V. Gumaste, Prof. S. A. Kahate, in their paper expressed that much work is required to overcome the challenges of big data such as heterogeneity, scalability, speed, accuracy, trust, Provenance, Privacy and inter-activeness [25].

Applications of dat mining

Alagukumar. S, Lawrance. R applied data mining in the field of bioinformatics by proposing an approach for analysis of microarray data which extracts frequent patterns and also extracts relations among microarray genes which helped in cancer detection and treatment[3].

Nidhi Bhatla, Kiran Jyoti employed different data mining techniques for heart disease prediction where they showed that neural network with 15 attribute gave 100% accuracy while decision tree with 15 attribute gave 99.62% accuracy and in combination with genetic algorithm decision tree with 6 attribute gave 99.2% accuracy[5].

M.A.Nishara Banu, B Gomathy have presented an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses for the efficient prediction of heart attack[17].

Have mined medical data sets on the basis of classification whether the patient is normal or abnormal an further detecting the heart disease using map reduce[26].

Mapreduce

Jens Dittrich, Jorge-Arnulfo Quian´e-Ruiz, gave a tutorial to familiaries the readers with mapreduce and have also discussed the idea to use Hadoop mapreduce for big data analysis[8].Kyuseok Shim in their tutorial introduced map reduce frame work based on Hadoop to make developers be aware of State of the art of mapreduce algorithms[23].Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, Bongki Moon made a survey report over map reduce and discussed its pro and cons such as it is simple but still provides good scalaility and fault-tolerence for massive data processing, input output cost of mapreduce is still to be addressed for successful implementation[7]. JeffreyDean and Sanjay Ghemawat implemented Mapreduce programming model at google for many different uses and gave several reasons of success such as firstly considered as easy to use model even for programmers with no knowledge of distributed systems, large number of problems were easily addressed by mapreduce technique, implementation of mareduce made efficient use of thousands of machines for large computational problems that google came across[9].

Later on, Ming-Yen Lin, Pei-Yu Lee, Sue-Chen Hsueh proposed three algorithms namely SPC [Single Pass Counting], FPC [Fixed Passes Combined-counting], DPC [Dynamic passes Counting] to examine the implementation of apriori algorithm over map reduce frame work and it was concluded that DPC outperforms both FPC and SPC[6].

Pattern mining using mapreduce

Shivanagouda B.Patil, Manjula G,[19] in their paper wrote about frequent pattern mining of big data as because the apriori algorithm was inefficient for big data due to its technique of multiple scans and expressed requirement of some parallel working algorithm. They came up with a parallel algorithm on the basis of map reduce which was capable of mining frequent pattern from big transactional data in which data was first divided into many subsets and were paralleley processed, finally the frequent pattern were selected.

Zahra Farzanyar, Nick Cercone proposed IMRApriori for efficient frequent item set mining on mapreduce framework for data sets being collected from social networking where they alsocompared IMRApriori and MRApriori and showed that IMRApriori

was more efficient[4].

Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong designed a parallel algorithm for mining frequent pattern over big transactional data based on an extended mapreduce framework[20].in this technique data is to be split into multiple files and the pattern from each file is extracted using bitmap computation using single scan of the data. The method was declared efficient and strong in scalabiity.

Guru Prasad M S, Nagesh H R and Swathi Prabhu, identified the factors affecting the performance of frequent item set mining algorithm through hadoop mpreduce and found approaches for optimizing the performance where the technique proved to be better in terms of execution time ans disk space utilization[10].

G.Vaishali, V.Kalaivani have mined medical data sets on the basis of classification whether the patient is normal or abnormal an further detecting the heart disease using map reduce[26].

Jinggui Liao, , yuelong Zhao, Saiqin Long proposed MRPrePost is a parallel algorithm for mining big data on hadoop platform which was an improvement over PrePost where MRPrePost was declared to be better than PrePost in terms of stability and scalability[22].

Fabio Fumarola and Donato Malerba in their paper proposed a parallel algorithm for Approximate Frequent Itemset mining using Mapreduce named as MrAdam, a novel parallel distributed algorithm which helps in making reasonable decision in the absence of a perfect answer[21].

LI Bing, Keith C.C. Chan gave a parallelized algorithm on the basis of mapreduce to mine opinion and analyse sentiments from the ambiguous, unstructured big data collected from social media such as twitter where the speed of execution accelerates with increase in size of data[16].

Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu introduced i^2MapReduce, A mapreduce framework for incremental big data processing which combines fine-grain incremental engine, a egenral purpose iterative model and a set of effective techniques for incremental iterative computation[18]. This effectively reduces the runtime.

Rule mining

Association rule mining or sequential rule mining is another important aspect of data mining after pattern mining. Philippe Fournier-Viger, Roger Nkambou, Vincent Shin-Mu Tseng presented RuleGrowth, An Algorithm for mining sequential rules based on pattern growth approach which discover more efficient and scalable sequential rules[1]. Philippe Fournier-Viger with others in [2] their work proposed algorithm names CMRules for mining sequential rules that are common to several sequences as

similar rules can be of same phenomenon[2].

Genetic algorithm

As it is known that association rule miningis an important technique so a strong rule generation technique is found by Umesh Kumar Patel by combining Apriori and FP growth algorithm further genetic algorithm was implemented to optimize the rule generation[11].

K.Y. Fung, C.K. Kwong , K.W.M. Siu, K.M. Yu proposed a two stage approach for rule generation where thefirst stage is simple chromosome design and second in refinement of rule [12].

B. Minaei-Bidgoli,R.Barmaki,M. Nasiri, proposed numerical association rule mining approach which was based on genetic algorithm where three measure were used to mine efficient rule : confidence, comprehensibility and interestingness [13]. The technique was declared to be useful and helpful.

V. Purushothama Raju,G.P. Saradhi Varma implied genetic algorithm technique for mining closed sequential pattern on the basis of fitness function and pruning method which was named as G-CSPM [14].

CONCLUSION

Here, presented is a small review in the field of data mining starting from frequent pattern mining, sequential pattern mining and till rule mining. Here various challenges and applications of data mining and a number of techniques yet being researched are discussed. Similarly with upcoming era of information concept of bigdata arrived bringing up hadoop mapreduce for distributed parallel processing. We also found that genetic algorithm helps in optimizing the results. At the end we conclude that even if great research,s have been done in the field of data mining yet alot is to be done as the complication such as size and dimension of data in increasing day by day. So the field of data mining always have a scope of research. It can also be proposed from the study that combining of the technique of hadoop mapreduce and genetic algorithm would be fruitfull as they would help in over coming each others limitation.

REFERENCES

[1] Philippe Fournier-Viger, Roger Nkambou, Vincent Shin-Mu Tseng, "*RuleGrowth: Mining Sequential Rules Common to Several Sequences by Pattern-Growth*", SAC'11, March 21-25, 2011, TaiChung, Taiwan.

[2] Philippe Fournier-Viger, Usef Faghihi, Roger Nkambou, Engelbert Mephu

Nguifo, "*CMRules:Mining sequential rules common to several sequences* ", Knowledge-Based Systems 25 (2012) 63–76.

[3] Alagukumar. S, Lawrance. R, "*Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals* ", International Journal of Computer Applications (0975 – 8887).

[4] Zahra Farzanyar, Nick Cercone, "*Efficient Mining of Frequent itemsets in Social Network Data based on MapReduce Framework*", 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[5] Nidhi Bhatla, Kiran Jyoti, "*An Analysis of Heart Disease Prediction using Different Data Mining Techniques* ", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October – 2012 ,ISSN: 2278-0181.

[6] Ming-Yen Lin, Pei-Yu Lee, Sue-Chen Hsueh, "*Apriori-based Frequent Itemset Mining Algorithms on MapReduce*", ICUIMC'12,February20–22, 2012,Kuala Lumpur, Malaysia.

[7] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, Bongki Moon, "*Parallel Data Processing with MapReduce: A Survey*", SIGMOD Record, December 2011 (Vol. 40, No. 4).

[8] Jens Dittrich, Jorge-Arnulfo Quian´e-Ruiz, "*Efficient Big Data Processing in Hadoop MapReduce*", August 27th - 31st 2012, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 12

[9] JeffreyDean and Sanjay Ghemawat, " *MapReduce: Simplified Data Processing on Large Clusters*".

[10] Guru Prasad M S, Nagesh H R and Swathi Prabhu," *High Performance Computation of Big Data: Performance Optimization Approach towards a Parallel Frequent Item Set Mining Algorithm for Transaction Data based on Hadoop MapReduce Framework*", I.J. Intelligent Systems and Applications, 2017, 1, 75-84.

[11] Umesh Kumar Patel, "*Optimization of Association Rule Mining Using Genetic Algorithm*", International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 5, Issue 6 June 2016.

[12] K.Y. Fung, C.K. Kwong , K.W.M. Siu, K.M. Yu, "*A Multi-objective Genetic algorithm approach to rule mining for affective product design*", Expert Systems with Applications 39 (2012) 7411–7419.

[13] B. Minaei-Bidgoli,R.Barmaki,M. Nasiri, "*Mining Numerical association rules via multi-objective genetic algorithm"*, Information sciences 233(2013) 15-24.

[14] V. Purushothama Raju,G.P. Saradhi Varma, "*Mining Closed Sequential Patterns Using Genetic Algorithm*", 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (lCACCCT).

[15] Ms. Aruna J. Chamatkar, Dr. P.K. Butey, *"Importance of Data Mining with Different Types of Data Applications and Challenging Areas*", Ms. A J. Chamatkar et al Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 5(Version 3), May 2014, pp.38-41.

[16] LI Bing, Keith C.C. Chan," *A Paralleled Big Data Algorithm with MapReduce Framework for Mining Twitter Data* ", 2014 IEEE Fourth International Conference on Big Data and Cloud Computing.

[17] M.A.Nishara Banu, B Gomathy," *Disease Predicting System Using Data Mining Techniques*", International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45

[18] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu, Member, IEEE, "*i2MapReduce: Incremental MapReduce for Mining Evolving Big Data*". IEEE Transactions on Knowledge And Data Engineering, VOL. 27, NO. 7, JULY 2015.

[19] Shivanagouda B.Patil, Manjula G," *Frequent Pattern Mining in Big Transactional Data Using Map Reduce Distributed Computing Framework*", International Conference on Computer Science, Electronics & Electrical Engineering-2015.

[20] Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong, "*Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce*", 2013 IEEE International Conference on Granular Computing (GrC).

[21] Fabio Fumarola and Donato Malerba," *A Parallel Algorithm for Approximate Frequent Itemset Mining using MapReduce*", 978-1-4799-5313-4/14/$31.00 ©2014 IEEE.

[22] Jinggui Liao, yuelong Zhao, Saiqin Long, "*MRPrePost-A parallel algorithm adapted for mining big data*", 2014 IEEE Workshop on Electronics, Computer and Applications.

[23] Bharti Thakur, Manish Mann," *Data Mining for Big Data: A Review*", Volume 4, Issue 5, May 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[24] Kyuseok Shim, "*MapReduce Algorithms for Big Data Analysis*", August 27th - 31st 2012, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 12.

[25] Pravin Anil Tak, Dr. S. V. Gumaste, Prof. S. A. Kahate, "*The Challenging View of Big Data Mining"*, Volume 5, Issue 5, May 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[26] G.Vaishali, V.Kalaivani, "*Big data analysis for heart disease detection system using Map reduce technique* ", 2016 international conference on computing technologies and intelligent data Engineering (ICCTIDE'16).

[27] Soumya Shukla, Vaishnavi Kukade, Sofiya Mujawar, "*Big Data :Concept, Handling and challenges : An overview",* International Journal of computer application(0975=8887)volume 114-No.11,March 2015.