IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.
www.ijiset.com

ISSN 2348 – 7968

Apriori- A Big Data Analysis in Education

Asif Ansari[#1], Ajit Parab[#2]

[#1]PG Scholar, AlamuriRatnamala Institute of Engineering & Technology, Mumbai, India
[#2]H.O.D, BabasahebGawde Institute of Technology, Mumbai, India
[#1]ansariasif23@gmail.com

Abstract

Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions , or having no timestamps (DNA sequencing). Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Apriori applied in the field of education refers to as educational data mining. It uses various computational approaches to deal with the educational questions.

Keywords: Apriori, Apache Hadoop, MapReduce, EDM.

INTRODUCTION

Data mining is the efficient discovery of previously unknown patterns in large datasets. It has attracted a lot of attention from both research and commercial communities for finding interesting information hidden in large datasets. One of the most important areas of data mining is association rule mining; its task is to find all subsets of items which frequently occur and the relationship between them by using two main steps: finding frequent item sets (core step) and generating association rules.

Educational Data Mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses different computational approaches to analyzeeducational data in order to study educational questions

DOMAINS WHERE APRIORI IS USED

Application of the Apriori algorithm for adverse drug reaction detection. The objective is to use the Apriori association analysis algorithm for the detection of adverse drug reactions (ADR) in health care data. The Apriori algorithm is used to perform association analysis on the characteristics of patients, the drugs they are taking, their primary diagnosis, co-morbid conditions, and the ADRs or adverse events (AE) they experience. This analysis produces association rules that indicate what combinations of medications and patient characteristics lead to ADRs.

Application of Apriori Algorithm in Oracle Bone Inscription Explication

Oracle Bone Inscription (OBI) is one of the oldest writing in the world, but of all 6000 words found till now there are only about 1500 words that can be explicated explicitly. So explication for OBI is a key and open problem in this field. Exploring the correlation between the OBI words by Association Rules algorithm can aid in the research of explication for OBI. Firstly the OBI data extracted from the OBI corpus are pre-processed; with these processed data as input for Apriori algorithm we get the frequent itemset. And combined by the interestingness measurement the strong association rules between OBI words are produced. Experimental results on the OBI corpus demonstrate that this proposed method is feasible and effective in finding semantic correlation for OBI.

The algorithm can be used in the field of education with the various mining association rules. In this paper we take a review of the application of Apriori in the field of education.

ISSN 2348 – 7968

APRIORI IN EDUCATION

The use of computers in learning and teaching has advanced significantly over the last decades through different e-learning systems. Nevertheless, the need for improvement has always been present. Students learn, explore content, and by doing so, they leave trail of log information. There is an important research question: what can we do with this data? These parameters could easily be recorded in e-learning system, analysed.

Apriori algorithm is the first and best-known for association rules mining. It is an influential algorithm for mining frequent itemsets for Boolean association rules. This algorithm uses a level -wise search, where k-itemsets are used to explore (k+1)-itemset, to mine frequent itemsets from transactional database.

ASSOCIAION RULES IN EDM

Association rule mining has been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset), including, the following tasks: building recommender agents for on -line learning activities or shortcuts, automatically guiding the learner's activities and intelligently generate and recommend learning materials , identifying attributes characterizing patterns of performance disparity between various groups of students , discovering interesting relationships from a student's usage information in order to provide feedback to thecourse author , finding out the relationships between each pattern of a learner's behaviour , finding student mistakes often occurring together , guiding the search for the best fitting transfer model of student learning , optimizing the content of an e learning portal by determining the content of most interest to the user , extractinguseful patterns to help educators and web masters evaluating and interpreting on -line course activities , and personalizing e-learning based on aggregate usage profiles and a domain ontology

LIMITATION OF APRIORI ALGORITHM EDM

In spite of being simple and clear, Apriori algorithm has some limitation. It is costly to handle a huge number of candidate sets. For example, if there are 104frequent 1-item sets, the Apriori algorithm will need to generate more than107length-2 candidate'sand accumulate and test their occurrence frequencies. This is the inherent cost of candidate generation, no matter what implementation technique is applied. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. We can see from the above description of the Apriori algorithm that in general there exist two major shortcomings in Apriori algorithm [7]. Firstly, it needs to scan database repeatedly and secondly, it needs to generate large number of candidate item set. Apart from this Apriorialgorithm also assumes that the weight of each item and transaction in database is same which is not true always.

LITERATURE SURVEY

As basic Apriori Algorithm can't be used for education data mining due previously mentioned shortcoming, various authors have suggested variant in Apriori algorithm to address the mentioned shortcomings.

 Mixed Weighted Association Rules Mining Algorithm: Xin-huaZhu,Ya-qiong Deng, Qing-ling Zeng [8] in their paper have taken grades of computer cultural foundation course as example. The weighted association rules algorithm have been used to analyse grades of college-wide examination course in this paper, giving the model and mining method of mixed weighted association rules on grade database. Compared to directly apply the

ISSN 2348 – 7968

Apriori algorithm, more valuable correlations have been obtained between the chapters, chapters and scores, colleges and chapters at the same threshold values. Thus, it is more helpful to guide teachers for their teaching and for the improvement of teaching quality.

 Clustering with Apriori algorithm: Zhiyu Zhang [9] in his paper used Clustering algorithm to first categories the students and courses and then with the help Apriori algorithm various hidden information is extracted from the large amount of education data.

Chandrani Singh, Dr.ArpitaGopal ,Santosh Mishra [10] in their paper deals with the extraction and analysis of faculty performance of management discipline from student feedback using clustering and association rule mining techniques. First the faculty members are categorised based on student feedback data and then Apriori algorithm has been implemented to undermine the hidden trends in Faculty performance and their behaviour.

 Matrix based Apriori algorithm: Hong Liu, Yuanyuan Xia [11] in their paper used matrix based Apriori algorithm to extract and analyse the Indicator-score in teaching evaluation data and found the information of Indicator-score that have high frequency, and then analyzed the strengths and weaknesses of the teaching, to provide recommendations of improving teaching quality of teachers. This method does not repeat the database scanning, thus reducing the I/O load.

 Improved Apriori algorithm based on Tid set: QiangYang,Yanhong Hu [12] intheir paper used improved Apriori algorithm to find the correlation rules of course which provided the directive significance information for the curriculum . This algorithm need to scan the original database only once when generating candidate item set, it compute support count of the other candidate item sets through stating the count of the corresponding Tid set, not scanning the database repeatedly, which saves the visiting time greatly.

Improved Apriori algorithm based on clipping technique: Jian Wang, ZhubinLu, WeihuaWu and Yuzhou Li [13] in their paper used improved Apriori algorithm of association rules to analyze the intrinsic link among various courses, dig out the precedence relationship and association of students' learning courses, reveal the teaching regularities and problems from large amount of data, as well as provide a strong basis for reasonable course -setting .

Improved Apriori algorithm uses clipping technique to remove all candidate itemset in Ck that doesn't belong to Lk-1. Improved Apriori algorithm based on logo list intersection: Lanfang Lou, Qingxian Pan, XiuqinQiu [14] in their paper proposed a novel association rules for data mining to improve Apriori algorithm. The new designed approach uses the intersection operation to generate frequent item sets. It is different from the existing algorithm as it scans the database only one time and then uses the database to mine association rules. The proposed technique has been implemented in a teaching evaluation system, to enhance the foundation in performance evaluation for staff in teaching issues.

 Improved Apriori Algorithm based on Modified Pruning process and flag bit: Deng Jiabin, Hu JuanLi, Chi Hehua, Wu Juebo[15] in their paper put forward a kind of intelligent evaluation method based on improved Apriori, which can beused to mine different levels of association rules and evaluate the teaching quality automatically. The improvement ideological of the frequent items:

ISSN 2348 – 7968

Between the Lk and Ck, introducing the Lk', when one item has been validated that it is not a frequent item set, it will be inserted into Lk', but not be deleted. In order to distinguish an item set whether it is frequent item sets or non-frequent item sets flag bit is introduced into the item sets. When it is thefrequent item set, we use 1, or else use 0. At the same time, the verification process and the pruning process also need to be modified: when verifying the candidate set Ck, each time, we select i tems from the item set Ck to verify.

However, each time, we select items from Lk' in the pruning process as the Pruning conditions and iteratively generate Lk +1'. Different techniques have implemented to improve the shortcoming of classic Apriori algorithm in education data mining. Although these improved algorithms can reduce the number of candidate itemsets or improve the mining efficiency by pruning methods, but still can't completely solve the problem of which candidate itemsets appear no longer. And, what's more, facing masses of education data for mining long pattern to adopt basic association rules mining is not the solution to problem as they will be producing a large number of candidate itemsets, using lots of memory space and CPU processing time. Apart from this setting the appropriate minimum support threshold is also an issue as it may lead to too many or too few rules.

CONCLUSION

Apriori algorithm with a main motive of reducing time and number of scans required to identify the frequent itemset and associationrules among education data using bottom up approach. Moreover the algorithm also replaces the user-defined minimum threshold with standard deviation based functional model as mentioning minimum support value in advance may lead to either too many or too few rules which can negatively impact the performance of entire model.

Apriori algorithm can be used with FP growth tree in the future scope for the data mining.

The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone.

REFERENCES

1- www.educationaldatamining.org
2- Cristobal Romero,"Educational Data Mining: A Review of the State of the Art", IEEE transaction on systems, MAN, and Cybernetics-part c: Application and Reviews, VOL. 40, NO. 6, November 2010
3- Agrawal, R., Imielinski, T. and Swami, A.N., Mining Association Rules between Sets of Items in Large Databases. In Proceedings of SIGMOD, 207-16, 1993.
4- Zheng, Z., R. Kohavi and Mason, L., Real world performance of association rules. Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2(2),86-98, 2001.
5- Agrawal R., SrikantR.,"Fast algorithms for mining association rules", In Proceedings 20th International Conference on VeryLarge Data Bases (VLDB' 94), pp. 487-499, 1994.
6- Deng Jiabin, Hu JuanLi, Chi Hehua, Wu Juebo, "An Apriori-based Approach for Teaching Evaluation", IEEE, 2010.
7- White, T., Hadoop: The Definitive Guide. O'Reilly Media, May 2009.
8- Lin, J. and Dyer, C., Data-Intensive Text Processing with MapReduce. To be published by Morgan and Claypool. Pre-press edition available at http://www.umiacs.umd.edu/~jimmylin/MapReduce-book-20100219.pdf, February 2010.