International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
International Manuscript ID : 22296166V7I3201704
Volume 7 Issue 3 September 2017

Improved Apriori Algorithm for Association Rule Mining using Cost Optimization and Effective Recommendations

Sukriti Mittal
M.Tech. Research Scholar
Department of Computer Science and Engineering
R. P. Inderaprastha Institute of Technology
Karnal, Haryana, India

Er. Ritika Mehra
Assistant Professor
M.Tech. Research Scholar
Department of Computer Science and Engineering
R. P. Inderaprastha Institute of Technology
Karnal, Haryana, India

ABSTRACT

Recommendation Engines and Rule Mining are closely related in terms of fetching the dataset for predictive analytics. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout. Programmers use association rules to build programs capable of machine learning. Machine learning is a

type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed. In the existing / classical Apriori Algorithm, only the frequent datasets are evaluated rather than the associated cost factor. Till now, the cost of each association rule is not evaluated by any researcher. In our proposed work, the cost with each combination is considered so that the global cost optimal results can be achieved. The simulation is done in web based scenarios for existing and proposed Apriori algorithm. The dataset of shopping mart is taken and final results are achieved in form of best fit cost optimized results.

Keywords: *Data Mining, Machine Learning, Predictive Analytics, Rule Mining*

INTRODUCTION

Data mining refers to the analysis of the large quantities of data that are stored in computers. Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. Data mining requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis.

APPROACHES FOR ASSOCIATION RULES

There are many Data Mining algorithms to mine frequent patterns for finding association rules. The two widely used algorithms are FP-Growth and Apriori.

Frequent pattern growth a very popular association rule mining algorithm for discovering itemsets in a database. The algorithm follows two step approaches for finding interesting rules. The Step1 of the algorithm builds a tree known as FP tree and in step2 frequent items are extracted from this FP tree. FP Growth algorithm is a 2-pass algorithm over database.

Apriori Algorithm is a decisive algorithm for mining frequent itemsets for Boolean association rules. It uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1) itemsets.

Eclat (alt. ECLAT, stands for Equivalence Class Transformation) is a depth-first search algorithm using set intersection. It is a naturally elegant algorithm suitable for both sequential as well as parallel execution with locality enhancing properties. It was first introduced by Zaki, Parthasarathy, Li and Ogihara in a series of papers written in 1997.

Mohammed Javeed Zaki, Srinivasan Parthasarathy, M. Ogihara, Wei Li: New Algorithms for Fast Discovery of Association Rules. KDD 1997.

AprioriDP utilizes Dynamic Programming in Frequent itemset mining. The working principle is to eliminate the candidate generation like FP-tree, but it stores support count in specialized data structure instead of tree.

CBPNARM is the newly developed algorithm which is developed in 2013 to mine association

rules on the basis of context. It uses context variable on the basis of which the support of an itemset is changed on the basis of which the rules are finally populated to the rule set.

FIN, PrePost and PPV are three algorithms based on node sets. They use nodes in a coding FP-tree to represent itemsets, and employ a depth-first search strategy to discovery frequent itemsets using "intersection" of node sets.

GUHA is a general method for exploratory data analysis that has theoretical foundations in observational calculi.

OPUS is an efficient algorithm for rule discovery that, in contrast to most alternatives, does not require either monotone or anti-monotone constraints such as minimum support. Initially used to find rules for a fixed consequent it has subsequently been extended to find rules with any item as a consequent. OPUS search is the core technology in the popular Magnum Opus association discovery system.

Multi-Relation Association Rules (MRAR) is a new class of association rules which in contrast to primitive, simple and even multi-relational association rules (that are usually extracted from multi-relational databases), each rule item consists of one entity but several relations.

Context Based Association Rules is a form of association rule. Context Based Association Rules claims more accuracy in association rule mining by considering a hidden variable named context variable which changes the final set of association rules depending upon the value of context variables. For example the baskets orientation in market basket analysis reflects an odd pattern in the early days of month.

Contrast set learning is a form of associative learning. Contrast set learners use rules that differ meaningfully in their distribution across subsets.

Weighted class learning is another form of associative learning in which weight may be assigned to classes to give focus to a particular issue of concern for the consumer of the data mining results.

High-order pattern discovery facilitate the capture of high-order (polythetic) patterns or event associations that are intrinsic to complex real-world data.

K-optimal pattern discovery provides an alternative to the standard approach to association rule learning that requires that each pattern appear frequently in the data.

Approximate Frequent Itemset mining is a relaxed version of Frequent Itemset mining that allows some of the items in some of the rows to be 0.

Generalized Association Rules hierarchical taxonomy (concept hierarchy)

Quantitative Association Rules categorical and quantitative data

Interval Data Association Rules e.g. partition the age into 5-year-increment ranged

Sequential pattern mining discovers subsequences that are common to more than minsup sequences in a sequence database, where minsup is set by the user. A sequence is an ordered list of transactions.

Sequential Rules discovering relationships between items while considering the time ordering. It is generally applied on a sequence database. For example, a sequential rule found in database of sequences of customer transactions can be that customers who bought a computer and CD-Roms, later bought a webcam, with a given confidence and support.

Warmr is shipped as part of the ACE data mining suite. It allows association rule learning for first order relational rules.

RECOMMENDER SYSTEM

The advancement in technology in the field of Electronic commerce or e-commerce has enabled businesses to open up their products and services to a massive client base. As the competition between businesses becomes increasingly fierce, consumers are faced with a myriad of choices and hence information overload. Although this might seem to be nothing but beneficial to the consumer, the sheer wealth of information related to the various choices can be overwhelming.

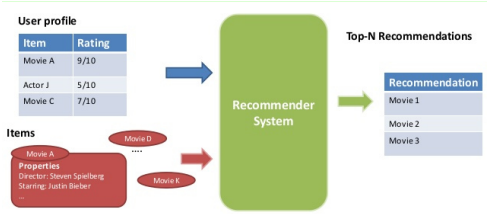

Figure 1: Recommender System

- Collaborative Filtering
- Content-based Filtering
- Knowledge-based Filtering
- Hybrid Systems
- User Model

REVIEW OF LITERATURE

In this chapter, various research papers include introduction to data mining, association rule mining and recommender systems.

Kumar et al. implements three phases of Web usage mining namely preprocessing, pattern discovery, and pattern analysis. Apriori algorithm is used to generate an association rule that associates the usage pattern of the clients for a particular website. The output of the system was in terms of memory usage and speed of producing association rules.

Martinez-Romo et al. have analyzed different information retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement for a broken link. To test the sources, they have also defined an evaluation methodology which does not require the user judgments, what increases the objectivity of the results.

Fayyad et al. have focused on web log file format, its type and location. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file. Data preprocessing is an important step to filter and organize appropriate information before using to web mining algorithm.

Kleinberg categorized web mining into three areas of interest based on which part of the Web to mine; Web Content mining, Web Structure mining, and Web Usage Mining. In Web mining, data collected at the server-side, client-side, proxy servers or a consolidated Web/business database. Hedberg provided data sources that can be used to construct several data abstractions, namely users, page-views, click-streams and server sessions.

Tang et al. have used re-ranking method and generalized Association Rules to extract access patterns of the Web sites pattern usage.

Lekhi (2015) - The Outlier detection is very active area of research in data mining where outlier is a mismatched data in dataset with respect to the other

available data. In existing approaches the outlier detection done only on numeric dataset. For outlier detection if we use clustering method then they mainly focus on those elements as outliers which are lying outside the clusters but it may possible that some of the unknown elements with any possible reasons became the part of the cluster so we have to concentrate on that also.

Zengyou He et al proposed Squeezer algorithm, a clustering algorithm for categorical data. It takes n tuples as input and produces clusters as output. Initially, the first tuple is read and cluster structure is constructed.

André Baresel et al proposed Evolutionary Structural Testing. It uses Evolutionary Algorithms (EA) to search for specific test data that provide high structural coverage of the software under test. A necessary characteristic of evolutionary structural testing is that the fitness function is constructed on the basis of the software under test.

Zengyou He et al proposed FindCBLOF Algorithm for detecting outliers. This algorithm computes the value of CBLOF for each record which determines the degree of record's deviation. This algorithm is efficient for handling large datasets . Zengyou He et al proposed NabSqueezer algorithm, an improved Squeezer algorithm. NabSqueezer algorithm gives more weight to uncommon attribute value matches for finding similarity in similarity computation of Squeezer algorithm. In this algorithm weight of each attribute is precalculated using More Similar Attribute Value Set (MSFVS) method.

RESULTS AND DISCUSSION

The Association Rule Mining Apriori Algorithm has few drawbacks such as the iterations involved reduce the minimum support until it finds the required number of rules with the given minimum confidence. There is need to propose and implement the work based on cost factor optimization.

Figure 2: Implementation Screenshot

Phase - 1 : Products Table

Table 1 – Product Table of User Purchase

Product	Price
Milk	100
Milk	100
Yogurt	80
Yogurt	80
Jam	40
Butter	30
Bread	20
Bread	20

Phase - 2 : Products Occurrences Count

Table 2 – Product Occurrences

Product	Occurrences	Price
milk	2	100
yogurt	2	80
jam	1	40
butter	1	30
bread	2	20

Phase - 3 : Sorting

Table 3 – Sorted Product Occurrences

Product	Occurrences	Price
milk	2	100
yogurt	2	80

bread	2	20
jam	1	40
butter	1	30

Phase - 4 : New Arrivals Products Table

Table 4 – New Arrival of Products

Product	Price
french jam	120
sugar	90
american tea	150
tea	90
jam	190

Phase - 5 : Recommendations (Classical Approach)

New Arrival Products - Array (=> french jam => sugar => american tea => tea => jam)

Table 5 – Recommendations in Classical Approach

Earlier Similar Puchased Item	Price	Recommended Puchase Item	Price
jam	40	Jam	190

Execution Time -> 1.0321700572968 MicroSeconds

Phase - 1 : Products Table

Table 6 – Product Table

Product	Price
milk	100
milk	100
yogurt	80
yogurt	80
jam	40
butter	30
bread	20
bread	20

Phase - 2 : Products Occurrences Count

Table 7 – Product occurrences

Product	Occurrences	Price
milk	2	100
yogurt	2	80
jam	1	40
butter	1	30
bread	2	20

Phase - 3 : Sorting

Table 8 – Sorted Product occurrences

Product	Occurrences	Price
milk	2	100
yogurt	2	80
bread	2	20
jam	1	40
butter	1	30

Phase - 4 : New Arrivals Products Table

Table 9 – New Arrival of Products

Product	Price
french jam	120
sugar	90
american tea	150
tea	90
jam	190

Phase - 5 : Recommendations (Proposed Approach)

Table 10 – Recommendations in Proposed Approach

Product	Price
sugar	90
tea	90

Execution Time -> 0.042825937271118 MicroSeconds

Comparative Analysis

Table 11 – Comparison of Execution Time

Classical Approach	Proposed Approach
1.0620608329773	0.29401683807373
1.0590600967407	0.22601199150085
1.0650610923767	0.059003114700317
1.0620610713959	0.057003021240234
1.0740621089935	0.032001972198486
1.0820620059967	0.038002014160156
1.040060043335	0.026001930236816
1.0260589122772	0.029001951217651
1.0260591506958	0.077003955841064
1.0412991046906	0.031000852584839
1.0403831005096	0.033002138137817
1.0382359027863	0.037002086639404
1.0321700572968	0.034002065658569

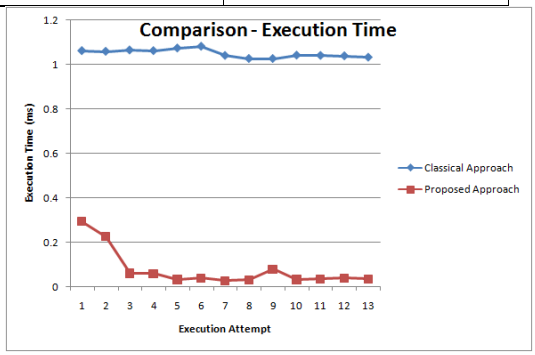

Figure 3: Comparison of Execution Time

Classical Approach	Proposed Approach
83	91
85	93
87	97
84	91
89	92
80	91
81	91
90	97
83	97
90	94
89	97

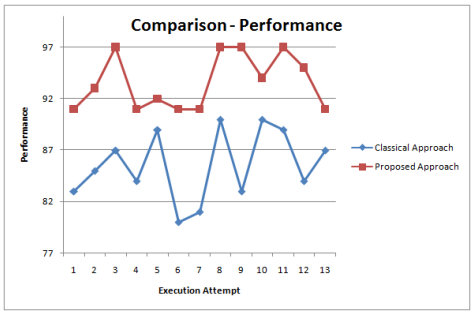

Figure 4: Comparison of Performance

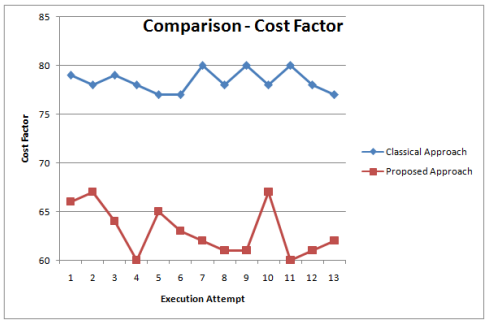

Figure 5: Comparison of Cost Factor

CONCLUSION

Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. A massive amount of research work is under process throughout the globe in assorted algorithms. In this research work, we have proposed and implemented a novel algorithm that makes use of the mathematical foundation and

evolutionary approach for the formation of clusters in efficient and effective manners in terms of execution time and associated results. A sample data set of shopping mart has been implemented and the algorithm performs in excellent manner on the desired aspects.

In addition and for further improvements, the nature inspired approaches and soft computing approaches can be used to achieve the global optimization. As deep Learning is one of the constituent of soft computing having core tasks associated with classification, recognition which is generally related with the artificial intelligence. Generally, these operations are performed using some metaheuristic approach in which the global optimization or simply effective results can be fetched from a huge search space of solutions.

The prominent soft computing approaches which can be used for further optimization include

- Fuzzy Logic
- Support Vector Machines
- Swarm Intelligence
- Metaheuristics
    - Ant Colony Optimization
    - Cuckoo Search
    - Bees Algorithm
    - Particle Swarm Optimization
    - Firefly Algorithm
    - Bat Algorithm
    - Simulated Annealing
    - Flower Pollination Algorithm
- Bayesian Network
- Evolutionary Approaches
- Nature Inspired Algorithms
- River Formation Dynamics

REFERENCES

[1] Dohare, M.P.S., Arya, P. and Bajpai A., 2012, "Novel Web Usage Mining for Web Mining Techniques" International Journal of Emerging Technology and Advanced Engineering, vol. 2, Issue 1, pp. 253-262.

[2] Gantner, Z., Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2011, October). MyMediaLite: a free recommender system library. In Proceedings of the fifth ACM conference on Recommender systems (pp. 305-308). ACM.

[3] Sharma, P. and Bhartiya, R., 2011, "An efficient Algorithm for Improved Web Usage Mining" International Journal of Computer Technology & Applications, Vol. 3, No.2, pp. 766-769.

[4] Kumar, B.S. and Rukmani, K.V., 2010, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms" International Journal of Advanced Networking and Applications, Vol. 1, Issue 6, pp. 400-404.

[5] Martinez-Romo, J. and Araujo, L., 2010, "Analyzing Information Retrieval Methods to Recover Broken Web Links", In Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010, Milton Keynes, UK, pp. 26-37.

[6] Zheng, Z., Ma, H., Lyu, M. R., & King, I. (2009, July). Wsrec: A collaborative filtering based web service recommender system. In Web Services, 2009. ICWS 2009. IEEE International Conference on (pp. 437-444). IEEE.

[7] Das, R., Turkoglu, I. and Poyraz, M., 2007, "Analyzing of System Errors for increasing a web server performance by using web usage mining", Journal of electrical & electronics engineering, Vol. 7, No. 2, pp. 379 – 386.

[8] Glance, N. S. (2005). U.S. Patent No. 6,947,922. Washington, DC: U.S. Patent and Trademark Office.

[9] Avesani, P., Massa, P., & Tiella, R. (2005, March). A trust-enhanced recommender system application: Moleskiing. In Proceedings of the 2005 ACM symposium on Applied computing (pp. 1589-1593). ACM.

[10] Shani, G., Heckerman, D., & Brafman, R. I. (2005). An MDP-based recommender system. Journal of Machine Learning Research, 6(Sep), 1265-1295.

[11] Avesani, P., Massa, P., & Tiella, R. (2005, March). A trust-enhanced recommender system application: Moleskiing. In Proceedings of the 2005 ACM symposium on Applied computing (pp. 1589-1593). ACM.

[12] Miller, B. N., Konstan, J. A., & Riedl, J. (2004). PocketLens: Toward a personal recommender system. ACM Transactions on Information Systems (TOIS), 22(3), 437-476.

[13] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003, January). MovieLens unplugged: experiences with an occasionally connected recommender system. In Proceedings of the 8th international conference on Intelligent user interfaces (pp. 263-266). ACM.

[14] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003, April). Is seeing believing?: how recommender system interfaces affect users' opinions. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 585-592). ACM.

[15] Huang, Z., Chung, W., Ong, T. H., & Chen, H. (2002, July). A graph-based recommender system for digital library. In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries (pp. 65-73). ACM.

[16] Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. Expert systems with Applications, 23(3), 329-342.

[17] O'connor, M., Cosley, D., Konstan, J. A., & Riedl, J. (2001). PolyLens: a recommender system for groups of users. In ECSCW 2001 (pp. 199-218). Springer Netherlands.

[18] Andrassyova, E. and Paralic, J., 2000, "Knowledge Discovery in Databases: A Comparison of Different Views", In Journal of information and organizational sciences, Varazdin, Croatia, Vol. 23, No. 2, pp. 95 - 102.

[19] Tang, C., Lau R.W.H., Li, Q., Yi, H., Li, T. and Kilis, D., 2000, "Personalized Courseware Construction Based on Web Data Mining", In Proceeding of the First International Conference on Web

Information Systems Engineering (WISE 2000), vol. 2, pp. 204-211.

[20] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of dimensionality reduction in recommender system-a case study (No. TR-00-043). Minnesota Univ Minneapolis Dept of Computer Science.

[21] Kleinberg, J.M., 1999, "Authoritative sources in a hyperlinked environment", Journal of the ACM, Vol. 46, No. 5, pp. 604-632.

[22] Resnick, P., & Varian, H. R. (1997). Recommender systems.Communications of the ACM, 40(3), 56-58.

[23] Hedberg, S.R., 1996, "Searching for the mother lode: tales of the first data miners", IEEE EXPERT, Vol. 11, No. 5, pp. 4-7.

[24] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, pp. 27-34.

[25] Brachman, R.J., Anand, T., 1996, "The Process of Knowledge Discovery in Databases", Advances in Knowledge Discovery & Data Mining, Fayyad, U.M. - Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds. AAAI/MIT Press, Cambridge, Massachusetts, pp. 37-57.

[26] Agrawal, R., Imielinski, T. and Swami, A., 1993, "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., pp 207-216.