The Generation of Association Rules Through the Utilisation of Frequent Item Sets

***Shikha Dhall**

****Dr. Ashutosh Sharma**

Abstract

The two techniques were compared with regard to the number of nodes that were present in the T-tree structure, the number of updates that were had to be made to the T-tree in order to locate big item sets, and the amount of bytes that were used to store the T-tree. The purpose of this study is to demonstrate the comparative connection between the numerous parameters that were calculated using the Apriori and IAR algorithms with varying amounts of data. Time is, however, the most significant component to consider. There is never a timing difference between IAR and Apriori. Both algorithms with support levels of 20% and 30% were compared in terms of their time constraints. Compared to the typical Apriori method, these numbers provide a clear indication of the temporal performance of the IAR algorithm. Because the data is created in a random fashion, it is important to keep in mind that the amount of time required and other characteristics may vary from one run to the next. Additionally, the behaviour of IAR does not necessarily have to be the same for the various qualities that are supplied by the user. On the other hand, it always requires less time and storage than Apriori does. It is also important to note that IAR does not conduct exhaustive searches; rather, it searches for association rules that include the attributes specification that the user has provided.

Keywords: T-tree structure, Number of nodes, Number of updates, Bytes used for storage, Apriori algorithm, IAR algorithm

Introduction

The process of discovering association rules is what is meant by the generation of association rule mining. The process of generating association rules makes use of the frequent item sets discovered in the stage before this one. Each and every permutation and combination of the things that are included in the frequent itemsets are taken into consideration as potential candidates for using severe rules. Numerous rules will be produced as a result of this method. A rule is considered to be strong if it has a minimal confidence level, which may be determined using the Formula. The IAR method takes into account the user's attribute preference for the rules that are produced, which is the primary distinction between the Apriori algorithm and the IAR algorithm. The IAR then looks for rules on the LHS that include the user-specified attributes and denies other attributes in the

database. This process is repeated until the desired attributes are found. In the event that such a regulation holds a high level of trust, then it has the potential to be beneficial for the organization throughout the marketing environment. By doing so, a significant amount of time may be saved, and the user puts greater faith in the rules that have been identified.

Review of Literature

According to [Agrawal R. The topic of identifying association rules between items in a big database of sales transactions has been taken into consideration [et al., 1994]. Apriori Tid and AIS are two novel algorithms that have been introduced for the purpose of discovering association rules between objects in huge datasets. These algorithms travel in a radically different direction than the techniques that are already in use. The results of the empirical assessment shown that these algorithms significantly outperform the algorithms that are already in use by a factor that ranges from three for minor issues to more than an order of magnitude for big problems. In later stages of the investigation, the most advantageous aspects of the two suggested algorithms were merged together to form a hybrid algorithm that was given the name Apriori Hybrid. The ability of Apriori Hybrid to scale linearly with the amount of transactions was shown by trials that were scaled up. Along with having great scale-up qualities, Apriori Hybrid also has outstanding features with regard to the quantity of transactions and the number of entries in the database.

To Y. A user-guided association rules mining approach that takes into consideration the users' varying levels of focus on each item using fuzzy regions was introduced by (2014). Compared to the approaches that were previously used for association rules, this approach was more realistic and practical. The method that was presented allocated fuzzy areas to each category characteristic in order to describe it in linguistic language that was more flexible according to the circumstances, particularly when the number of categorical regions was high. By using degraded membership functions and setting all weights to "1," the suggested solution was also able to overcome the usual binary value issue. Consequently, the strategy that was suggested was more adaptable, natural, and more easily understood. In conclusion, the suggested technique was used in the study to conduct an analysis of data sets pertaining to appliance loans, and the performance of the proposed methodology was also proved.

The authors DeXing Wang et al. It has been claimed by (2005) that domain knowledge that is concealed inside a database may often direct and constrain the search for intriguing information, and that it can also play additional roles in leading the discovery of knowledge within databases. The concept lattice, which is used to represent knowledge on the Hass diagram along with the link between entities and attributes, has been used to represent knowledge that has a hierarchical structure in the Hass diagram. Additionally, the same concept lattice was utilised to describe association rules mining in databases. Regarding the mining of association rules on idea lattice, the study described how to make use of domain knowledge to steer the process. The findings have shown that association rules mining illustrates the unexpected discoveries in a manner that is more

visually appealing, lively, and succinct by nature.

As stated in [Zhang S. assessment indexes of association rule mining methods have been provided by the authors of [et al., 2006]. A novel knowledge discovery approach of enhanced Apriori-based high-rise intelligent form selection was developed as the next step in the research process. An investigation into the high-rise structural case library has been carried out, which has resulted in the mining of many characteristics. These attributes include the height of the main building above the ground floor, the ratio of length to width, the ratio of height to width, the maximum wind pressure, performance, fortification intensity, site class, major structure form, and so on. Using this technique, a novel strategy was developed for mining knowledge information in engineering situations, directing the design of structural form selection, and enhancing the quality, efficiency, and intelligence level of structural design.

I am Xiaonan Ji X. et al. sequential patterns have been used by (2013) in order to acquire information in a broad variety of software applications. On the other hand, the quality of the pattern may be poor in many situations because of short lengths or low supports. Furthermore, when dealing with dense datasets like proteins, the majority of the sequential pattern mining algorithms produced an extremely high number of patterns, which were difficult to process and assess. Through the relaxation of the definition of frequency and the allowance of some mismatches, the article put out the hypothesis that it was feasible to find patterns of greater quality. Frequent Approximate Substrings, also known as FAS-patterns, were the names given to these patterns, and an algorithm known as FAS-Miner was developed in order to carry out the mining process in an effective manner. In the trials conducted on real-world protein and DNA datasets, it was shown that FAS-Miner was able to uncover patterns that were much longer in length and had greater supports than the conventional sequential mining methods.

Material and Method

Both the Apriori and IAR algorithms have been executed on the same platform under the same circumstances in order to evaluate the performance of the IAR method in terms of identifying frequent item sets. For the purpose of comparison, a number of parameters were calculated, and the results are shown in Tables 1 and 2, as well as Figure 1. The experimental runs have been carried out with two distinct degrees of assistance and datasets of varying sizes. It has been discovered that the IAR method always requires less time and storage space than the Apriori approach (which is the standard). It is possible to mine the interesting information with a shorter amount of energy. 7 characteristics are included in the test dataset. For the purpose of evaluating the effectiveness of the algorithm across a variety of data characteristics, the data was produced via the use of false transactions. All of the characteristics are numbered, beginning with one and proceeding in the order listed. This approach may be used to any database that is based on the actual world by translating the attribute names to the numbers 1, 2, 3, and so on respectively.

The methods make use of a T-tree data structure in order to store information on frequent item sets.

The storage need for each node in the T-tree that represents a frequent item set is 12 bytes. This includes the following: a) a reference to the structure of the T-tree, which takes up four bytes; b) support for the count field in the T-tree node structure, which takes up four bytes; and c) a reference to the T-tree node structure four bytes are allocated for the reference to the child array field in the T-tree node structure.

The two approaches were compared with regard to the number of nodes that were present in the T-tree structure, the number of updates that were necessary to locate big item sets inside the T-tree, and the amount of bytes that were used to store the T-tree. Both Table 1 and Table 2 show the comparative connection between the various parameters that were estimated using the Apriori and IAR algorithms with varying amounts of data. However, time is the most significant aspect to consider. There is never a timing difference between IAR and Apriori. The comparison of the amount of time required by both algorithms with support levels of 20% and 30% is shown in Figure 1 and Figure 2, respectively. Compared to the typical Apriori method, these numbers provide a clear indication of the temporal performance of the IAR algorithm. Because the data is created in a random fashion, it is important to keep in mind that the amount of time required and other characteristics may vary from one run to the next. Additionally, the behaviour of IAR does not necessarily have to be the same for the various qualities that are supplied by the user. On the other hand, it always requires less time and storage than Apriori does. It is also important to note that IAR does not conduct exhaustive searches; rather, it searches for association rules that include the attributes specification that the user has provided.

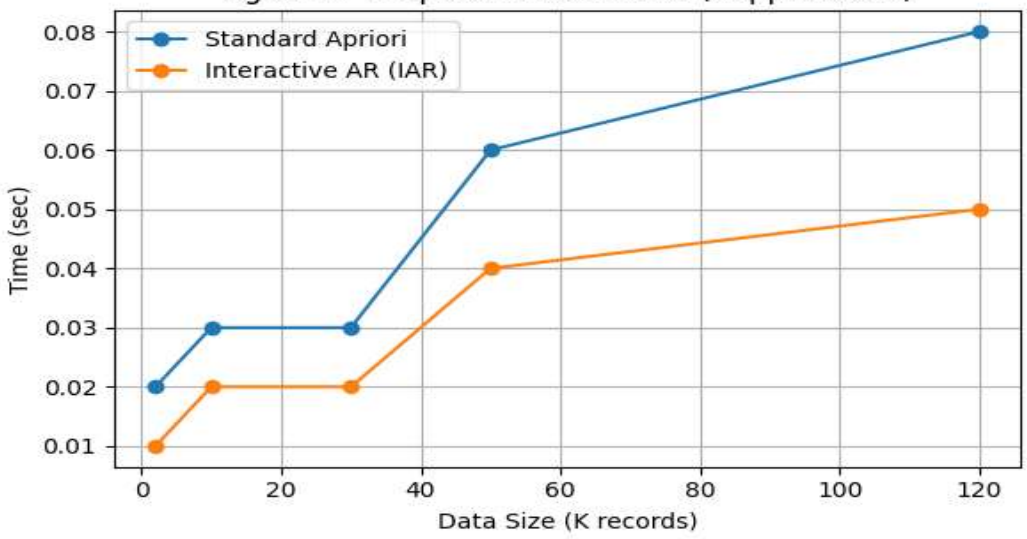
Table 1: Values of parameters with support level 20%

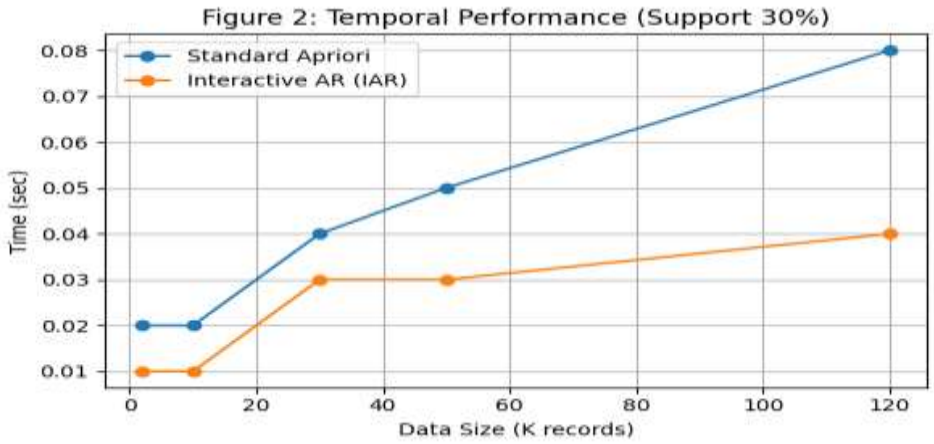
Data Size	Number of frequent itemsets Apriori	IAR	Number of nodes in T-tree Apriori	IAR	Updates required in T-tree Apriori	IAR	Storage of T-tree in bytes Apriori	IAR
2K	31	15	43	20	26458	11722	496	244
10K	32	13	45	19	132503	48566	504	212
30K	28	13	41	19	336547	128216	476	248
50K	30	15	42	18	574843	240544	484	272
120K	28	15	41	21	1346085	589970	476	280

Table 2: Values of parameters with support level 30%

Data Size	Number of frequent itemsets Apriori	IAR	Number of nodes in T-tree Apriori	IAR	Updates required in T-tree Apriori	IAR	Storage of T-tree in bytes Apriori	IAR
2K	20	9	33	14	22630	8883	324	148
10K	17	7	33	13	110607	40556	276	144
30K	15	5	30	10	291806	85866	240	140
50K	15	5	30	10	482533	141980	240	140
120K	15	5	30	10	1167228	342575	240	140

Figure 1: Temporal Performance (Support 20%)





CONCLUSION

When it comes to the different methods of data mining, rule-based approaches are the most suitable for incorporating human perspectives. This is because human thinking can be translated into rules in a reasonably straightforward manner. It is possible to include the user's recommendations and needs into the process of transferring domain knowledge. This may be accomplished either by supplying some implicit information to initiate the mining process or by including them into the findings. Because of this, the number of iterations that occur inside the knowledge discovery loop is reduced and lengthened.

The IAR approach that is being given is a modification of the normal Apriori algorithm. It was selected because it provides the user with the opportunity to have a part in discovering interesting associations between objects in a database. Comparisons are made between the two methods by using information of varying sizes and degrees of support. Based on the findings, it is clear that human engagement is a potentially fruitful area of data mining. In every situation, the IAR method performs better than the Apriori approach, and the performance improves as the amount of data rises. It is possible to demonstrate beyond a reasonable doubt that the knowledge associated with the domain user may play a significant role in the identification of sequences and patterns of interest.

***Research Scholar**
****Department of Computer Science**
Singhania University, Rajasthan

References

1. Ankerest M. (2001). Human Involvement and Interactivity of the Next Generation's Data Mining Tools. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara, CA.

2. Bethel C.L., Hall L.O. and Goldgof D. (2006). Mining for Implications in Medical Data. Proceedings of the 18th International Conference on Pattern Recognition, Vol.1, 1212-1215.
3. Bifoldo S., Buccini A. and Di-Bello C. (2004). Marker Identification and Classification of Cancer Types Using Gene Expression Data and SIMCA. Germany: Methods of Information in Medicine, Vol. 43(1), 44-48.
4. Brause R.W. (2001). Medical Analysis and Diagnosis by Neural Networks. Computer Science Department, Frankfurt am Main, Germany.
5. Chattoraj N. and Roy J.S. (2007). Application of Genetic Algorithm to the Optimisation of Gain of Magnetised Ferrite Microstrip Antenna. Engineering Letters, Vol. 14(2).
6. Ding Q. and He X. (2004). k-Means Clustering via Principal Components Analysis. ACM Proceedings of the 21st International Conference on Machine Learning, Vol. 69, page 29.
7. Guillaume S. (2001). Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review. IEEE Transactions on Fuzzy Systems, Vol. 9, 426-443.
8. Han J. and Kamber M. (2001). Data Mining: Concepts and Techniques. San Francisco, Morgan Kaufmann Publishers. Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R. and Wu A.Y. (2002). An Efficient k-means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, 881-892.
9. Kaur H. and Wasan S.K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. The Journal of Computer Science 2 (2). Science Publications, 194-200.
10. Li C.H. and Yuen P.C. (2001). Semi-supervised Learning in Medical Image Database. Advances in knowledge discovery and data mining, Vol. 2035, 154-160.
11. Li H. (2005). Applications of Data Warehousing and Data Mining in the Retail Industry. Proceedings of International Conference on Services Systems and Services Management, 2005. Vol. 2.
12. Li J. and Yao J. (2007). Study on Method of Association Knowledge Discovery for Schema Design Based on Constraint. Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 3, 694-698.
13. Li Y. (2005). A User-Guided Association Rules Mining Method and Its Applications. Fifth International Conference on Computer and Information Technology, IEEE Computer Society, Washington, 150–156.
14. Matsuoka K., Yokoyama S., Watanabe K. and Tsumoto S. (2007). Analysis of Relationship between Blood Stream Infection and Clinical Background in Patients' Lactobacillus Therapy by Data Mining. IEEE International Conference on Data Mining, IEEE Computer Society, Washington, 175-180.
15. Matsushita M. and Kato T. (2001). Interactive Visualisation Method for Exploratory Data Analysis. Proceedings of the Annual Conference of JSAI, Vol. 15, 1-3.

16. Maulik U. (2009). Medical image segmentation using genetic algorithms. IEEE Transactions on Information Technology in Biomedicine, Vol. 13(2), 166-173.

17. Moody G.B. and Mark R.G. (1996). A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring. IEEE Computers in Cardiology, Vol. 23, 657-660.

18. Mounji, A. (1997). Languages and Tools for Rule-Based Distributed Intrusion Detection. PhD thesis, Faculties Universitaires Notre-Dame de la Paix Namur (Belgium).

19. Pei J., Upadhyaya S.J., Farooq F. and Govindaraju V. (2004). Data Mining for Intrusion Detection: Techniques, Applications and Systems. Proceedings of the 20th International Conference on Data Engineering, 877.

20. Piatetsky-Shapiro G. and Simoudis E. (1995). Commercial KDD Applications: The Secret Ingredients for Success. First International Conference on Knowledge Discovery and Data Mining (KDD-95), Panel Session, Montreal.

21. Sung S.Y., Li Z., Tan C.L. and Ng P.A. (2003). Forecasting Association Rules Using Existing Data Sets. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, 1448-1459.

22. Xue R., Ji L. and Streveler D.J. (2004). Microarray Gene Expression Profile Data Mining Model for Clinical Cancer Research. Proceedings of the 37th Hawaii International Conference on System Sciences, Vol. 6, 60137.1.

23. Yeh C., King-Jang Yang and Tao-Ming Ting. (2009). Knowledge Discovery on RFM Model using Bernoulli Sequence. Expert Systems with Applications, Vol. 36(3-2), 5866-5871.