Journal of Advances in Science and Technology

Vol. IV, No. VII, November-2012, ISSN 2230-9659

REVIEW ARTICLE

GENERATION OF ASSOCIATION RULES USING FREQUENT ITEM SETS

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, No. VII, November-2012, ISSN 2230-9659

Performance Evaluation of IAR Algorithm in Discovering Frequent Item Sets

Jasvir Singh

Research Scholar, CMJ University, Shillong, Meghalaya, India

---------------------------♦-----------------------------

INTRODUCTION

The generation of Association rule mining is to discover the association rules. The frequent itemsets found in the previous step are used to generate association rules. All the permutations and combinations of the items present in the frequent itemsets are considered as candidates for strong rules. A lot of rules will be generated in this way. A strong rule is one that has minimum confidence which is computed by the Formula The main difference between Apriori and IAR algorithm is that IAR algorithm takes user's attribute preference for the resulting rules. Thereafter, the IAR searches for rules that contain the user specified attributes on the L.H.S. and derive other attributes in the database. If such a rule possesses high confidence level then it could be valuable in the marketing context for the organisation. In this way a lot of time can be saved and the user trusts more in the discovered rules.

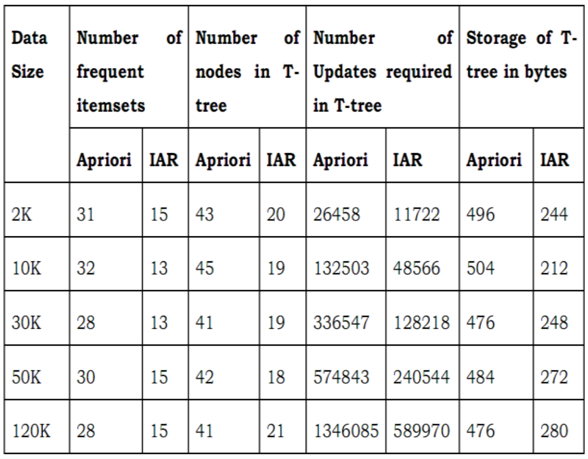
THE EXPERIMENT AND RESULTS

For the purpose of performance evaluation of IAR algorithm in discovering frequent itemsets, both Apriori and IAR have been run on the same platform under same conditions. Various parameters were computed for the purpose of comparison and the results have been shown in Tables 5.4 and 5.5, and Figure 5.4. The experimental runs have been conducted with two support levels and different sized datasets. It has been found that the IAR algorithm always takes less time and storage space than the standard Apriori. The interesting information can be mined in a shorter time. The test dataset has 7 attributes. The data was generated by artificial transactions to evaluate the performance of the algorithm over a range of data characteristics. The attributes are numbered starting from 1 and going in sequence. Any database of real world can be used with this algorithm by converting the attribute names to 1, 2, 3 and so on.

The algorithms use T-tree[1] data structure to store frequent item set information. The storage requirement for each node (representing a frequent item set) in the T-tree is 12 bytes i.e. a) reference to T-tree node structure (4 Bytes), b) support count field in T-tree node structure (4 Bytes) and c) reference to child array field in T-tree node structure (4 Bytes).

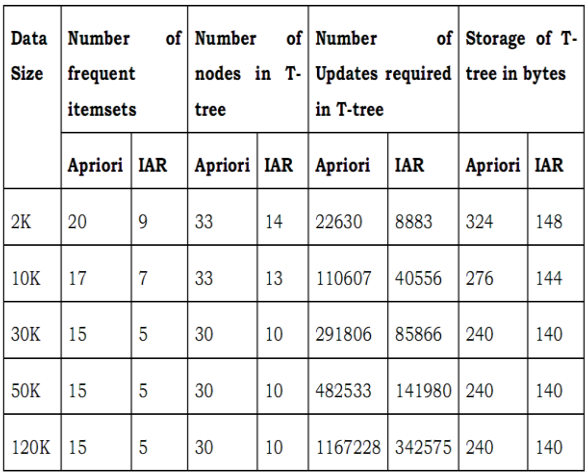
Both the algorithms were compared with respect to the number of nodes in the T-tree structure, updates required to in T-tree to find large itemsets and the storage of T-tree in bytes. Table 5.4 and Table 5.5 show the comparative relationship of the various parameters as computed in Apriori and IAR algorithms with different data sizes. However the most important factor is time. IAR always takes less time than Apriori. The time comparison of both the algorithms with support level 20% and 30% is shown in Figure 1 and Figure 2. These figures clearly indicate the time performance of IAR over the standard Apriori algorithm. It must be noted that the time taken and other parameters may differ for different runs as the data is generated randomly. Also the behaviour of IAR need not be the same for different attributes specified by the user. But it always takes less time and storage than Apriori. It must also be noted that IAR does not do exhaustive search instead it finds association rule containing the attributes specification given by the user.

Table 1: Values of parameters with support level 20%.



Data Size	Number of frequent itemsets		Number of nodes in T-tree		Number of Updates required in T-tree		Storage of T-tree in bytes	
	Apriori	IAR	Apriori	IAR	Apriori	IAR	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	128218	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		Number of nodes in T-tree		Number of Updates required in T-tree		Storage of T-tree in bytes	
	Apriori	IAR	Apriori	IAR	Apriori	IAR	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

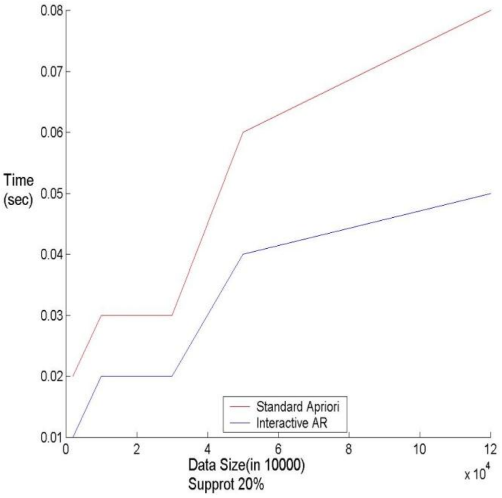


Figure 1: Temporal performance of Apriori (red - upper) and IAR (blue – lower) with Support level 20%.

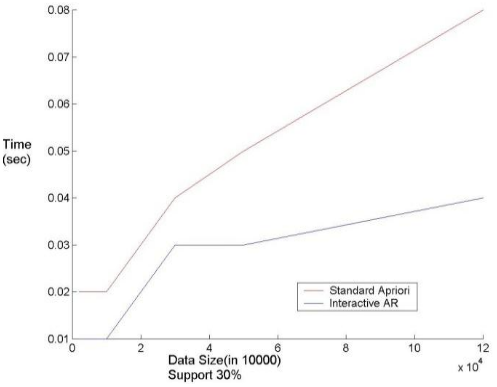


Figure 2: Temporal performance of Apriori (red - upper) and IAR (blue – lower) with Support level 30%.

CONCLUSION

Among various data mining techniques, rule based techniques are most appropriate for integrating human opinions, and human thoughts can be converted into rules relatively more easily. User's suggestions and demands can be incorporated in the process to transfer domain knowledge either by providing some implied information to instruct the mining process or by being merged into the results. This results in less and shorter iterations within the knowledge discovery loop.

The IAR algorithm presented is a variation of standard Apriori algorithm, and it was chosen to include user's role in finding interesting association among items in a database. The two algorithms are compared using different data sizes and support levels. The results show that human involvement is a promising field in data mining. The IAR algorithm always outperforms Apriori and the performance enhances as the data size increases. It can conclusively be made out that the domain user's knowledge may contribute a lot in the discovery of sequences and patterns of interest.

REFERENCES

Agrawal R. and Srikant R. (1994). Fast Algorithms for Mining Association Rule. **Proceedings of the 20th** International Conference on Very Large Databases (VLDB), 487 – 499.

Agrawal R., Faloutsos C. and Swami A. (1993). Efficient similarity search in sequence databases. Proceedings of the Fourth International Conference on Foundations of Data Organisation and Algorithms, Chicago, Vol. 730, 69-84.

Agrawal R., Imielinski T. and Swami A. (1993). Mining association rules between sets of items in large databases. **Proceedings of the 1993 ACM SIGMOD International Conference on Management of** Data, Washington DC, 207-216.

Ahmed S.R. (2007). Applications of Data Mining in Retail Business. International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, Vol. 2.

Anand S.S., Bell D.A. and Hughes J.G. (1995). The Role of Domain Knowledge in Data Mining. **Proceedings of the Fourth International Conference on Information and knowledge management, 37-43.**

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.

Ankerest M., Ester M. and Kriegel H.P. (2000). Towards an Effective Cooperation of the User and the Computer for Classification. Proceedings of 6th International conference on Knowledge Discovery and Data Mining, Boston, MA.

Aruna P., Puviarasan N. and Palaniappan B. (2005). An Investigation of Neuro-Fuzzy Systems in

Psychosomatic Disorders. Expert Systems with Applications. Vol. 28, 673-679.

Bates J.H.T. and Young M.P. (2003). Applying Fuzzy Logic to Medical Decision Making in the Intensive Care Unit. American Journal of Respiratory and Critical Care Medicine, Vol. 167, 948-952.

Bayrak C., Kolukisaoglu H. and Chia-Chu Chiang .(2006). Di-Learn: Distributed Knowledge Discovery with Human Interaction. IEEE International conference on Systems, Man and Cybernetics, Taipei, Taiwan, Vol. 4, 3354 – 3359.

Berks G., Keyserlingk D.G.V., Jantzen J., Dotoli M. and Axer H. (2000). Fuzzy Clustering - A Versatile Mean to Explore Medical Databases. ESIT, Aachen, Germany, 453-457.

Berson A., Smith S. and Thearling K. (1999). **Building Data Mining Applications for CRM. First Edition,** McGraw-Hill Professional.

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.

Bicciato S., Luchini 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), 4-8.

Brause R.W. (2001). Medical Analysis and Diagnosis by Neural Networks. Computer Science Department, Franfurt a.M., Germany.

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).

Cheung Y.M. (2003). k-Means: A New Generalised k-Means Clustering Algorithm. N-H Elsevier Pattern Recognition Letters 24, Vol 24(15), 2883–2893.

Chiang I.J., Shieh M.J., Hsu J.Y.J. and Wong J.M. (2005). Building a Medical Decision Support System for Colon Polyp Screening by Using Fuzzy Classification Trees. Applied Intelligence, Vol. 22 n.1, 61-75.

Chung H. M. and Paul G. (1999). Special Section: Data Mining. Journal of Management Information Systems, Vol. 16(1), 11 – 16**.**

Cios K.J. (2000). Medical Data Mining and Knowledge Discovery. IEEE Engineering in Medicine and Biology, Vol. 19(4), 15-16.

Cios K.J. and Moore G.W. (2002). Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine, Vol. 26, 1-24.

D. Jiang, Pei J. and Zhang A. (2005). An Interactive Approach to Mining Gene Expression Data. IEEE Transactions on Knowledge and Data Engineering, Vol. 17(10) 1363-1378.

Ding C. 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.

Edelstein H.A. (1999). Introduction to Data Mining and Knowledge Discovery (3rd Edition) Potomac, MD: Two Crows Corp.

Ester M., Kriegel H.P. and Sander J. (2001). Algorithms and Applications for Spatial Data Mining. Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis.

Fawcett T. and Provost F. (1997). Adaptive Fraud Detection. Data Mining and Knowledge Discovery, Vol. 1(3):291-316.

Fayyad U. M., Piatetsky-Shapiro G. and Smyth P. (1996). From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data mining, AAAI Press, 1-34.

Fayyad U., Piatetsky-Shapiro G. and Smyth P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of ACM, Vol. 39, 27-34.

Forgionne G.A., Gagopadhyay A. and Adya M. (2000). Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems. Topics in Health Information Management, Proquest Medical Library, Vol. 21(1), 21-34.

Frank H., Klawonn F., Kruse R. and Runkler T. (1999). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. New York: John Wiley.

Frawley W.J., Piatetsky-Shapiro G. and Matheus C.(1996). Knowledge Discovery in Databases: An Overview. Knowledge Discovery in Databases, AAAI Press/MIT Press, Cambridge, MA., Menlo Park, C.A, 1-30.