Mining Frequent Item Sets for Association Rule Mining in Relational Databases: An Implementation of SETM Algorithm Using Super Market Dataset

D. Kerana Hanirex* and K. P. Kaliyamurthie

Department of Computer Science Engineering, Bharath University, Chennai - 600073, Tamil Nadu, India; keranahanirex.cse@bharathuniv.ac.in, kaliyamurthie.cse@bharathuniv.ac.in

Abstract

Association rule mining is one of the recent data mining research. Mining frequent itemsets in relational databases using relational queries give great attention to researchers nowadays. This paper implements set oriented algorithm for mining frequent itemsets in relational databases. In this paper the sort and merge scan algorithm SETM is implemented for super market data set. This paper finds out the frequent itemset and its execution time and the results are compared with the traditional Apriori algorithm.

Keywords: 5-6 words, Drawn from title, Word representing the work.

1. Introduction

Data mining is the process of finding the hidden information from the database. Since large amounts of information are stored in companies for decision making the data need to be analyzed carefully. This process is known as Data mining or knowledge discovery in databases. Data mining consists of various tasks such as classification, clustering, association rule mining, sequence analysis. Classification is the task of classifying the data into predefined classes where as in clustering data are grouped together based on some common characteristics. In association rule mining one can find the frequent item set. An item set is a frequent item set if the number of occurrence is above the minimum threshold. Various work has been proposed in association rule mining. Various algorithms has been proposed for mining frequent item sets. That algorithms are feasible only for small data sets. Performance is a problem when dealing with large data sets. Paper proposes association rule discovery in large databases. This system proposes association rule discovery in database systems. Paper proposes set oriented approach in database systems. This set oriented approach integrate the association rule discovery with relational database systems.

This paper describes the set oriented mining in section 2. In section 3 explanation for set oriented algorithm SETM which includes simple sorting and simple merge scan join operations is given. Section 4 describes about the dataset description. Section 5 describes the data pre-processing. Section 6 includes the results and discussions. Section 7 presents our conclusion.

2. Set Oriented Mining

This paper uses supermarket data set for finding frequent item sets. This data set consists of collection transactions...
which involves the purchase of items by a customer. The frequent item set can be generated by repeated joins with TRANS. TRANS is a table which stores all the transactions of the customers from the supermarket data set with the schema TRANS(trans_id, item).

An association rule is of the form \( X \Rightarrow Y \) where \( X \) is the antecedent and \( Y \) is the consequent of the rule. Support of an itemset can be defined as the ratio of the number of transactions supporting that transaction to the total number of transactions in the database. Confidence measure can be defined as the ratio of the number of transactions that occur together to the number of transactions in the antecedent. An item set is a frequent item set if the number of occurrence of the item set is above the minimum support value. We are interested with association rules whose support value is greater than the minimum support value. For any pattern of length \( k \), the association rules are generated by combining all possible combinations of \( k-1 \) items in the antecedent. For each antecedent and consequent of the rule we have to check if the confidence exceeds the minimum confidence factor.

The frequent item set are generated by performing repeated joins with TRANS table. To generate all pattern of exactly 2 items we can use this query

\[
\begin{align*}
\text{SELECT } s_1.\text{trans_id}, & s_1.\text{item}, s_2.\text{item} \\
\text{FROM TRANS } s_1, & \text{TRANS } s_2 \\
\text{WHERE } s_1.\text{trans_id} = & s_2.\text{trans_id} \\
& s_1.\text{item} <> s_2.\text{item}
\end{align*}
\]

All patterns of 3 item sets are generated by joining the 2 item pattern with the TRANS table. We can repeat this joining to find the largest frequent item set. This doesn’t give the good result based on its performance.

### 3. Setm Algorithm

This algorithm is proposed by Houtsma and Swami. This algorithm uses two sort operations and one merge-scan join. The first sort is used to implement merge scan join and the second sort is used in order to generate support count efficiently.

#### Algorithm:

\[
k = 1 \\
\text{Sort } S_1 \text{ on item} \\
C_1 = \text{Generate counts from } S_1 \\
\text{repeat} \\
k = k + 1
\]

### 4. Data Set Description

This algorithm is implemented using super market data set available in UCI data repository with 4627 instances and 217 attributes.

### 5. Data Preprocessing

Data need to be processed in order to improve the quality of the data. The various tasks of data mining are data transformation, data integration, data discretization, data cleaning and data reduction. Data cleaning involves removing noisy data, incomplete data, inconsistent data. Data integration combines data from multiple sources. Data transformation task contains data aggregation, generalization, data smoothing, and normalization. Data reduction includes data aggregation, high dimensionality reduction, data compression and discretization.

In this paper we have applied preprocessing using weka3.6.4 tool for this Supermarket data set. The following figure Figure 1 shows CSV file in Excel for Supermarket Dataset. Here \( t \) represents that item present in the corresponding transaction.

The above figure Figure 2 shows the supermarket dataset. The data preprocessing filters such as Replace MissingValues, Normalization are not suitable for this data set.

Here we applying one of the data preprocessing technique called selection of attributes. Here the attributes are selected using Weka3.6.4 tool using Attribute Evaluator OneRAttributeEval and the search method we are using here is the Ranker. The following figure Figure 3 shows the selection of attributes in supermarket marketdataset. Here 211 attributes are selected instead of 216 using this OneRAttributeEval.

In this paper the accuracy is generated by finding the number of association rules generated for different threshold values. The effectiveness of the association rule mining is measured by considering the time taken to
generate the association rules from databases. The execution time and accuracy generated by SETM algorithm is compared with our traditional Apriori algorithm\textsuperscript{12}.

**Apriori Algorithm using super market dataset**

The following tables 1 and 2 show the performance evaluation of Apriori algorithm for its efficiency & accuracy by varying its confidence and support\textsuperscript{13}.

### Table 1. Efficiency(Apriori algorithm)

| Confidence | Support 10 | Support 20 | Support 30 | Support 40 | Support 50 |
|------------|------------|------------|------------|------------|------------|
| Confidence | 10         | 20         | 30         | 40         | 50         |
| 90         | 109        | 115        | 120        | 121        | 123        |
| 80         | 115        | 120        | 121        | 123        |            |
| 70         | 120        | 121        | 123        |            |            |
| 60         | 121        | 123        |            |            |            |
| 50         | 123        |            |            |            |            |

### Table 2. Accuracy(Apriori algorithm)

| Confidence | Support 10 | Support 20 | Support 30 | Support 40 | Support 50 |
|------------|------------|------------|------------|------------|------------|
| Confidence | 10         | 20         | 30         | 40         | 50         |
| 90         | 458        | 8071       | 24570      | 43580      | 69369      |
| 80         | 8071       | 24570      | 43580      | 69369      |            |
| 70         | 24570      | 43580      | 69369      |            |            |
| 60         | 43580      | 69369      |            |            |            |
| 50         | 69369      |            |            |            |            |

**SETM Algorithm using super market dataset**

The following tables table3&table4 show the performance evaluation of SETM algorithm for its efficiency & accuracy by varying its confidence and support\textsuperscript{14}.
Execution time in Secs(SETM Algorithm)

| Support | Confidence |
|---------|------------|
| 10      | 80         |
| 20      | 30         |
| 30      | 15         |
| 40      | 10         |
| 50      | 10         |

Table 3. Efficiency(SETM algorithm)

| Support | Confidence |
|---------|------------|
| 10      | 105        |
| 20      | 30         |
| 30      | 15         |
| 40      | 10         |
| 50      | 10         |

Number of association rule generated (SETM Algorithm)

Table 4. Accuracy(SETM algorithm)

| Support | Confidence |
|---------|------------|
| 10      | 464        |
| 20      | -          |
| 30      | -          |
| 40      | -          |
| 50      | -          |

From the above table we can identify that SETM algorithm has higher efficiency and accuracy than our traditional Apriori algorithm.[15]. This algorithm reveals some additional hidden rules and the execution time is less when compared with Apriori algorithm.

The following figure Figure 4 shows the performance evaluation of Apriori and SETM algorithm by varying confidence value with support=10 for its efficiency.

Figure 5. Comparison of Number of Rules Generated Between Apriori and SETM Algorithm for Supermarket Dataset with Support=20 and by Varying Confidence. From the above results we can understand SETM algorithm works well and the number of rules generated gets increased when the confidence value get increased while with increasing support value the number of rules generated gets decreased. The execution time gets increased when the confidence value gets decreased and gets decreased with increasing support value[17].

The following figure Figure 5 shows the performance evaluation of Apriori and SETM algorithm by varying confidence value with support=20 for its accuracy[16].

6. Conclusions

The SETM algorithm has been implemented to find the frequent item set using super market dataset. The efficiency and accuracy of the SETM algorithm is measured and it is compared with traditional Apriori algorithm. This algorithm works well and stable and provides similar performance with the traditional Apriori algorithm. The execution time is less when co

7. References

1. Houtsma M, SwamiA. Set oriented mining for association rules in relational databases. Proceeding of the 11th International Conference on Data Engineering; 1995 Mar 6-10; Taipei. USA: IEEE Press; 1995. p. 25–33.
2. Udayakumar R, Khanaa V, Saravanan T. Chromatic dispersion compensation in optical fiber communication system and its simulation. Indian Journal of Science and Technology. 2013 Jun; 6(S6):4762–6. ISSN: 0974-6846.
3. Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective. IEEE Trans. Knowledge and Data Engineering 1993; 5(6):914–25.
4. Kumar S, Das MP, Jeyanthi RL, Sharmila S. Isolation and identification of LDPE degrading fungi from municipal solid waste. Journal of Chemical and Pharmaceutical Research. 2013; 5(3):78–81. ISSN: 0975-7384.
5. Anwar TM, Beck HW, Navathe SB. Knowledge mining by imprecise querying: A classification-based approach. IEEE 8th Int Con J on Data Engineering; 1992 Feb 2-3; Phoenix, Arizona, USA. USA: IEEE Computer Society Press; 1992.
6. Udayakumar R, Khanaa V, Saravanan T. Analysis of polarization mode dispersion in fibers and its mitigation using an optical compensation technique. Indian Journal of Science and Technology. 2013; 6(S6):4767–71. ISSN: 0974-6846.
7. Fischer DH. Knowledge acquisition via incremental conceptual clustering. Machine Learning 2. Boston: Kluwer Academic Press; 1987. p. 139–72.
8. Sundar RM, Arkin VH, Adalarasu, Jagannath M. Nano-composites based on polymer and hydroxyapatite for drug delivery application. Indian Journal of Science and Technology. 2013; 6(S5):4653–8. ISSN: 0974-6846.
9. Michalski RS, Kerschberg L, Kaufman KA, Ribeiro JS. Mining for knowledge in databases: The INLEN architecture, initial implementation, and first results. J Intelligent Information Systems. 1992; 1:85–113.
10. Udayakumar R, Khanaa V, Saravanan T. Synthesis and structural characterization of thin films of SnO2 prepared by spray pyrolysis technique. Indian Journal of Science and Technology. 2013; 6(S6):4754–7. ISSN: 0974-6846.
11. Catlett J. Megainduction: A test flight. In: Birnbaum L, Collins G, editors. 8th Int Conf on Machine Learning; 1991; Evanston, Illinois, USA. Burlington, USA: Morgan Kaufman; 1991 Jun.
12. Cheeseman P. Autoclass: A bayesian classification system. In: Laird J, editor. Proceeding of 5th Int Conf on Machine Learning; 1988 Jun 12-14; Ann Arbor, Michigan, USA. Burlington, USA: Morgan Kaufman; 1988 Jun.
13. DeWitt DJ, Naughton JF, Schneider DA. Parallel sorting on a shared-nothing architecture using probabilistic splitting. In: Gray J, editor. Proc 1st Int Conf on Parallel and Distributed Information Systems; 1991. USA: IEEE; 1991 Dec. p. 280–91.
14. Krishnamurthy R, Imielinski T. Practitioner problems in need of database research: Research directions in knowledge discovery, ACM-SIGMOD Record. 1991 Sep; 20(3):76–8.
15. Houtsma M, Swami A. Set-oriented mining for association rules, Technical report. IBM Research Division; 1993 Oct; IBM Research Report RJ 9567.
16. DeWitt DJ, Naughton JF, Schneider DA. Parallel sorting on a shared-nothing architecture using probabilistic splitting. In: Gray J, editor. Proc. 1st Int Conf on Parallel and Distributed Information Systems; 1991. USA: IEEE; 1991 Dec. p. 280–91.
17. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S, editors. Proc ACM-S1GMOD Int Conf on Management of Data; Washington, DC; 1993 May 26-28. US: Association for Computing Machinery Press; 1993 Jun. p. 207–16.
18. Kimio T, Natarajan G, Hideki A, Taichi K, Nanao K. Higher involvement of subtelomere regions for chromosome rearrangements in leukemia and lymphoma and in irradiated leukemic cell line. Indian Journal of Science and Technology. 2012 Apr; 5(1):1801–11.
19. Cunningham CH. A laboratory guide in virology. 6th ed. Minnesota: Burgess Publication Company; 1973.
20. Sathishkumar E, Varatharajan M. Microbiology of Indian desert. Ecology and vegetation of Indian desert. In: Sen DN, editor. India: Agro Botanical Publ; 1990. p. 83–105.
21. Varatharajan M, Rao BS, Anjaria KB, Unny VKP, Thayagarajan S. Radiotoxicity of sulfur-35. Proceedings of 10th NSRP; India. 1993. p. 257–8.
22. 01 Jan 2015. Available from: http://www.indjst.org/index.php/vision