VOL. 12, NO. 19, OCTOBER 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PERFORMANCE STUDY AND CHALLENGES FOR ALGORITHMS MINING RARE AND CORRELATED ITEMS IN VIDEO DATASET

K. Kumar[1] and P. Sudhakar[2]
[1]Ponnaiyah Ramajayam Institute of Science and Technology University, Tamil Nadu, India
[2]Department of Computer Software Engineering, Annamalai University, Tamil Nadu, India
Email: kkumarmoorthy@gmail.com

ABSTRACT

Data mining research is much occupied with Association rule mining (ARM) wherein these rules attempts to mine frequent items. However, in recent years, there has been an increasing demand for mining the infrequent or rare or minimal correlated items. The point is that interesting relationship among infrequent items has not been discussed much in the literature. In this paper, we conduct a comparative performance study on three such algorithms namely AprioriRare, AprioriInverse and CORI. After studying their pros and cons, we suggest how they can be applied in mining the video transaction datasets.

Keywords: data mining, association rule, frequent items, infrequent items, correlation.

1. INTRODUCTION

The aim of association rule mining is to discover relationships among set of items in a transactional database including video databases sequence represented numerically. An application of association rule mining is market basket analysis. Association rule is an implication of the form X⇒YX⇒Y, where XX and YY are frequent itemsets in a transaction database and X∩Y=∅X∩Y=∅. The rule X⇒YX⇒Y can be interpreted as "if itemset XX occurs in a transaction, then itemset YY will also likely occur in the same transaction". By such information, market personnel can place itemsets XX and YY within close proximity, which may encourage the sale of these items together and develop discount strategies based on such association/correlation found in the data.

2. RELATED LITERATURE

Association rule has been extensively studied in the literature since Agrawal *et al*. first introduced it in [1] [2]. Agrawal and Imielinski discussed mining sequential patterns in [3], as well as mining quantitative association rules in large relational tables in [4], while Bayardo considered efficiently mining long patterns from a database in [5] and Dong and Li studied efficient mining of emerging patterns in [6]. On the other hand, Kamber *et al*. [7] proposed using data cubes to mine multi-dimensional association rules and Lent *et al*. used the clustering method in [8]. While most researchers focus on association analysis of rules [9] [10] [11] [12] [13] [14], Brin *et al.* analyzed the correlations of association rules in [15]. With the development of data mining techniques, quite a few researchers have worked on alternative patterns; for example, Padmanabhan *et al*. discussed unexpected patterns in [16], Liu *et al*. and Hwang *et al.* studied exception patterns in [17] [18] [19], and Savasere *et al*., Wu *et al*. and Yuan *et al*. discussed negative association in [20] [21] [22] respectively. Recently, there are some growing interests in developing techniques for mining association patterns without a support constraint [23] [24] [25]. The algorithms proposed in [23] are limited to dealing with identifying pairs of similar columns. The approaches presented in [24] and [25] employ a confidence-based pruning strategy instead of the support-based pruning adopted in traditional association rule mining. The mining of support-free association discovers rules in the patterns with high support, cross-support where items have widely differing support levels, and low support. The patterns with a high minimum support level often are obvious and well known; the patterns with cross-support level have extremely poor correlation and patterns with low support often provide valuable new insights. J. Ding discussed association rule mining among rare items in [26]. He designed a new disk-based data structure, called Transactional Co-Occurrence Matrix (TCOM) to store the data information. This structure combines the advantages of transactional oriented (horizontal) layout and item oriented (vertical) layout of the database. So any itemsets could be randomly accessed and counted without full scan of the original database or the TCOM.

3. ALGORITHMS

The algorithms provide valid rules by exploiting support and confidence requirements, and use a minimum support threshold to prune its combinatorial search space. Two major problems may arise when applying such strategies.

(1) If the minimum support is set too low, this may increase the workload significantly, such as the generation of candidate sets, construction of tree nodes, comparisons and tests. It will also increase the number of rules considerably, which causes the traditional problem of algorithms suffering from extremely poor performance.

(2) If the minimum support threshold is set too high, many interesting patterns involving items with low support are missed.

We prefer to observe the role of infrequent items from which rules can be derived and algorithms which are popular on that subject are discussed.

3.1 Apriori rare

This is an algorithm for mining minimal rare itemsets from a transaction database. It is an Apriori-based algorithm. It was proposed by Szathmary *et al.* (2007). The input is a transaction database and a threshold named minsup (a value between 0 and 100 %).

Let us consider the transaction dataset in Table-1,

Table-1. Transaction dataset.

Transaction id	Items
t1	{1, 2, 4, 5}
t2	{1, 3}
t3	{1, 2, 3, 5}
t4	{2, 3, 5}
t5	{1, 2, 3, 5}

A transaction database is a set of transactions. Each transaction is a set of items. For example, consider the following transaction database. It contains 5 transactions (t1, t2... t5) and 5 items (1, 2, 3, 4, and 5 which denote a scene sequence in a given video). For example, the first transaction represents the set of items 1, 2, 4 and 5. It is important to note that an item is not allowed appearing twice in the same transaction and that item are assumed to be sorted by lexicographical order in a transaction.

An itemset is a unordered set of distinct items. The support of an itemset is the number of transactions that contain the itemset divided by the total number of transactions. For example, the itemset {1, 2} has a support of 60% because it appears in 3 transactions out of 5 in the previous database (it appears in t1, t2 and t5). A frequent itemset is an itemset that has a support no less than the *minsup* parameter. A minimal rare itemset is an itemset that is not a frequent itemset and that all its subsets are frequent itemsets.

Description: modification of Apriori to find minimal rare itemsets (mRIs)
Input: dataset + min_supp
Output: all frequent itemsets + minimal rare itemsets

1) $C_1 \leftarrow$ {1-itemsets}
2) $i \leftarrow 1$
3) while ($C_i \neq \emptyset$) {
4) SupportCount(C_i)
5) $R_i \leftarrow \{r \in C_i \mid \text{support}(r) < min\_supp\}$ // *R - for rare itemsets*
6) $F_i \leftarrow \{f \in C_i \mid \text{support}(f) \geq min\_supp\}$ // *F - for frequent itemsets*
7) $C_{i+1} \leftarrow$ Apriori-Gen(F_i) // *C - for candidates*
8) $i \leftarrow i + 1$
9) }
10) $I_{MR} \leftarrow \bigcup R_i$ // *minimal rare itemsets*
11) $I_F \leftarrow \bigcup F_i$ // *frequent itemsets*

For example, if we run AprioriRare algorithm with minsup = 60 % and the previous transaction database, we obtain the following set of minimal rare itemsets:

Table-2. Minimal rare set.

Minimal Rare Itemsets	Support
{4}	20 %
{1, 3, 5}	40 %
{1, 2, 3}	40 %

The input file is defined as follows:
1 2 4 5
1 3
1 2 3 5
2 3 5
1 2 3 5
The output file is:
4 #SUP: 1
1 2 #SUP: 2
1 5 #SUP: 2

The output file here consists of three lines which indicates that the itemsets {4}, {1, 2} {1, 5} are perfectly rare itemsets having respectively a support of 1 transaction, 2 transactions and 2 transactions.

3.2 AprioriInverse algorithm

This algorithm mine perfectly rare itemsets. One reason is that it is useful for generating the set of sporadic association rules.

(1) Generate inverted index I of (item, [TID-list]) from D.
(2) Generate sporadic itemsets of size 1:
$S_1 = \emptyset$
for each item $i \in I$ do begin
if count$(I,i)/|D|$ < maximum support and
count(I,i) > minimum absolute support
then $S_1 = S_1 \cup i$
end
(3) Find S_k, the set of sporadic k-itemsets where $k \geq 2$:
for $(k = 2;\ S_{k-1} \neq \emptyset;\ k++)$ do begin
$S_k = \emptyset$
for each $i \in$ {itemsets that are extns of S_{k-1}} do begin
if all subsets of i of size $k-1 \in S_{k-1}$
and count(I,i) > minimum absolute support
then $S_k = S_k \cup i$
end
end
return $\bigcup_k S_k$

The output of AprioriInverse is the set of all perfectly rare itemsets in the database such that their support is lower than *maxsup* and higher than *minsup*. A perfectly rare itemset (sporadic itemset) is an itemset that is not a frequent itemset and that all its proper subsets are also not frequent itemsets. Moreover, it has to have a support higher or equal to the *minsup* threshold. With the same example Table-1, running the AprioriInverse algorithm with minsup = 0.1 % and maxsup of 60 % and this transaction database, we obtain,

Table-3. Rare Itemset.

Perfectly Rare Itemsets	Support
{3}	60 %
{4}	40 %
{5}	60 %
{4, 5}	40 %
{3, 5}	20 %

The input file is defined as follows:
1 2 4 5
1 3
1 2 3 5
2 3 5
1 2 3 5
The output file is:
3 #SUP: 3
4 #SUP: 2
5 #SUP: 3
3 5 #SUP: 1
4 5 #SUP: 2

The output file consists of five lines which indicate that the itemsets {3}, {4}, {5}, {3, 5}, {4, 5} are perfectly rare itemsets having respectively a support of 3, 2, 3 1 and 2 transactions.

3.3 CORI algorithm

An algorithm for mining rare correlated itemsets. It is an extension of the ECLAT algorithm. It uses two measures called the *support* and the *bond* to evaluate if an itemset is interesting and should be output. CORI discover itemsets that are rare and correlated in a transaction database (rare correlated itemsets). A rare itemset is an itemset such that its *support* is no less than a *minsup* threshold set by the user. The *support* of an itemset is the number of transactions containing the itemset.

Dataset D

A minimal correlation threshold minbond of the anti-monotone constraint.

A minimal conjunctive support threshold minsup of the monotone constraint.

Result: The RCP set of rare correlated patterns.
begin
A Scan the dataset D once to build the transformed dataset D∗
B. Initialization of the tree-data structure
a) Computing the conjunctive support of the items and sorting them in an ascendant order of their support value.
b) Rare items are printed in the output set.
c) The sorted items are added to the tree structure.
C. Recursive processing of each item in order to extract the rare correlated itemsets
D. Memory liberation end

A correlated itemset is an itemset such that its *bond* is no less than a *minbond* threshold set by the user. The *bond* of an itemsets is the number of transactions containing the itemset divided by the number of transactions containing any of its items. The bond is a value in the [0, 1] interval. A high value means a highly correlated itemset. Note that single items have by default a bond of 1.

For example, if CORI is run on the transaction database (as shown in Figure-1) with a *minsup* = 80% and *minbond* = 20%, CORI outputs the following rare correlated itemsets:

Table-4. Items bond and support.

itemsets	bond	support
{1}	1	3
{4}	1	1
{1, 4}	0.33	1
{3, 4}	0.25	1
{1, 3, 4}	0.25	1
{1, 2}	0.4	2
{1, 2, 3}	0.4	2
{1, 2, 5}	0.4	2
{1, 2, 3, 5}	0.4	2
{1, 3}	0.75	3
{1, 3, 5}	0.4	2
{1, 5}	0.4	2
{2, 3}	0.6	3
{2, 3, 5}	0.6	3
{3, 5}	0.6	3

The input file is defined as follows:
1 3 4
2 3 5
1 2 3 5
2 5
1 2 3 5
The output file is:
1 #SUP: 3 #BOND: 1.0
4 #SUP: 1 #BOND: 1.0
4 1 #SUP: 1 #BOND: 0.3333333333333333
4 3 #SUP: 1 #BOND: 0.25
4 1 3 #SUP: 1 #BOND: 0.25
1 2 #SUP: 2 #BOND: 0.4
1 2 3 #SUP: 2 #BOND: 0.4
1 2 5 #SUP: 2 #BOND: 0.4
1 2 3 5 #SUP: 2 #BOND: 0.4
1 3 #SUP: 3 #BOND: 0.75
1 3 5 #SUP: 2 #BOND: 0.4
1 5 #SUP: 2 #BOND: 0.4
2 3 #SUP: 3 #BOND: 0.6
2 3 5 #SUP: 3 #BOND: 0.6
3 5 #SUP: 3 #BOND: 0.6

The output file here consists of 15 lines. Consider the last line. It indicates that the itemset {3, 5} is a rare correlated itemset having a support and bond of respectively 3 and 0.6.

4. MINING CHALLENGES IN VIDEO DATASET

There were also studies needed to improve the speed of finding large itemsets with hash table, map, and tree data structures [14] [27]. Compare to the numerous works done to search for better algorithms to mine large itemsets in a video sequence database, the qualifying criterion - support threshold - and the mechanism behind it - counting, temporal distance between the items - have received much less attention. The problem with support threshold is that it requires expert knowledge, which is subjective at best, to set this parameter in the system. Setting it arbitrarily low will qualify itemsets that should be left out, vice versa. Moreover, as database size increases, the support threshold may need to be adjusted [28] [29].

5. CONCLUSIONS

AprioriRare and AprioriInverse are based on Apriori, it suffers from the same fundamental limitations i.e. it may generate too much candidates and it may generate candidates that do not appear in the database.

The CORI algorithms use a tree structure with a specific node structure and an optimized construction and pruning strategies. Two main features constitute the thrust of Cori algorithm:

i) Only one scan of the database is performed to build the new transformed dataset. This helps to optimize the time and the space needed to support computing.

ii) It offers a resolution for the problem of handling both rarity and correlation constraints

Comparatively, CORI seems to perform better and must be improved to apply it widely on a video dataset.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami. 1993. Mining association rules between sets of items in large databases, in: Proceedings of the Association for Computing Machinery-Special Interest Group on Management of Data, ACM-SIGMOD. pp. 207-216.

[2] R. Agrawal, R. Srikant. 1994. Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Databases, VLDB, September. pp. 487-499.

[3] R. Agrawal, R. Srikant. 1995. Mining sequential patterns, in: International Conference on Data Engineering, ICDE. pp. 85-93.

[4] R. Srikant, R. Agrawal. 1996. Mining quantitative association rules in large relational tables, in: Proceedings of the Association for Computing Machinery-Special Interest Group on Management of Data, ACM SIGMOD. pp. 1-12.

[5] R.J. Bayardo. 1998. Efficiently mining long patterns from database, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of data. pp. 85-93.

[6] G. Dong, J. Li. 1999. Efficient mining of emerging patterns: Discovering trends and differences, in: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD. pp. 43-52.

[7] M. Kamber, J. Han, J.Y. Chiang. 1997. Metarule-guided mining of multi-dimensional association rules using data cubes, in: Proceeding of 3rd International Conference on Knowledge Discovery and Data Mining, KDD. pp. 207-210.

[8] B. Lent, A. Swami, J. Widom. 1997. Clustering association rules, in: Proceeding of International Conference Data Engineering, ICDE. pp. 220-231.

[9] R. Agrawal, T. Imielinski. 1993. A. Swami Database mining: A performance perspective IEEE Transactions on Knowledge and Data Engineering. 5(6): 914-925.

[10] J. Han, J. Pei, Y. Yin. 2000. Mining frequent pattern without candidate generation, in: Proceeding of ACM SIGMOD International Conference Management of Data, ICMD. pp. 1-12.

[11] M. Chen, J. Han. 1996. P. Yu Data mining: An overview from a database perspective IEEE Transactions on Knowledge and Data Engineering. 8(6): 866-881.

[12] H. Mannila, H. Toivonen, A. Verkamo. 1994. Efficient algorithm for discovering association rules, in: Knowledge Discovery and Data Mining, KDD. pp. 181-192.

[13] A. Savasere, E. Omiecinski, S. Navathe. 1995. An efficient algorithm for mining association rules in large databases, in: Proceeding of the 21st International Conference on Very Large Databases, VLDB. pp. 432-444.

[14] R. Srikant, R. Agrawal. 1995. Mining generalized association rules, in: Proceedings of the 21th

International Conference on Very Large Data Bases, VLDB. pp. 407-419.

[15] S. Brin, R. Motwani, C. Silverstein. 1997. Beyond market basket: Generalizing association rules to correlations, in: Special Interest Group on Management of Data, SIGMOD. pp. 265-276.

[16] B. Padmanabhan, A. Tuzhilin: 2000. Discovering the minimal set of unexpected patterns, in: Proceeding of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD. pp. 54-63.

[17] H. Liu, H. Lu, L. Feng, F. Hussain. 1999. Efficient search of reliable exceptions, in: Proceedings of the Third Pacific–Asia Conference on Methodologies for Knowledge Discovery and Data Mining, PAKDD. pp. 194-204.

[18] F. Hussain, H. Liu, E. Suzuki, H. Lu. 2000. Exception rule mining with a relative interestingness measure, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD. pp. 86-97.

[19] S. Hwang, S. Ho, J. Tang. 1999. Mining exception instances to facilitate workflow exception handling, in: Proceedings of the Sixth International Conference on Database Systems for Advanced Applications, DASFAA. pp. 45-52.

[20] A. Savasere, E. Omiecinski, S. Navathe. 1998. Mining for strong negative associations in a large database of customer transactions, in: Proceedings of the Fourteenth International Conference on Data Engineering, ICDE. pp. 494-502.

[21] X. Wu, C. Zhang, S. Zhang. 2004. Efficient mining of both positive and negative association rules ACM Transactions on Information Systems. pp. 381-405.

[22] X. Yuan, B. Buckles, Z. Yuan, J. Zhang. 2002. Mining negative association rules, in: Proceedings of the Seventh International Symposium on Communications, ISCC. pp. 623-629.

[23] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. Ullman, C. Yang. 2000. Finding interesting associations without support pruning, in: Proceedings of the 16th International Conference on D Engg, ICDE. pp. 489-500.

[24] K. Wang, Y. He, D. Cheung, Y. Chin. 2001. Mining confident rules without support requirement, in: Proceedings of ACM International Conference on Information and Knowledge Management, CIKM. pp. 89-96.

[25] H. Xiong, P. Tan, V. Kumar. 2003. Mining strong affinity association patterns in data sets with skewed support distribution, in: Proceedings of the Third IEEE International Conference on Data Mining, ICDM. pp. 387-394.

[26] J. Ding. 2005. Efficient association rule mining among infrequent items. Ph.D. Thesis, University of Illinois at Chicago.

[27] Vijayakumar. V and Nedunchezhian. R. 2011. Recent Trends and Research Issues in Video Association Mining. The International Journal of Multimedia & Its Applications (IJMA). 3(4).

[28] László Szathmáry. 2014. Finding minimal rare itemsets with an extended version of the Apriori algorithm. Proceedings of the 9th International Conference on Applied Informatics Eger, Hungary. 1: 85-92 doi: 10.14794/ICAI.9.2014.1.85.

[29] Anil Kumar, Er. Varun Singla. 2015. Comparative analysis of Data Mining Algorithms for Frequent item set with enhance technique. International Journal of Applied Engineering Research, ISSN 0973-4562, 10(55).

[30] Zhiyong Ma, Juncheng Yang, Taixia Zhang and Fan Liu. 2016. An Improved Eclat Algorithm for Mining Association Rules Based on Increased Search Strategy. International Journal of Database Theory and Application. 9(5): 251-266.