T. Sowmiya, V. Thangamani / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.354-358

Generating Similar Item Sets Of Temporal Databases Using Spamine Algorithm

T. Sowmiya[1], V. Thangamani[2]
[1]pg Scholar, Dept Of Cse, Vel Tech Dr.Rr & Dr.Sr Technical University, Chennai-62.
[2]pg Scholar, Dept Of It, Anna University, Guindy, Chennai-25.

ABSTRACT

Data mining is the process of extracting interesting like non-trivial, implicit, previously unknown and potentially useful information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Data mining is known as one of the core processes of Knowledge Discovery in Database (KDD). Association rule mining is a popular and well researched method for discovering interesting relationships among various data items in large databases. The existing system uses only frequent item sets for association rule mining. It may easily cause thrashing when dataset becomes large and sparse. To overcome this spamine algorithm is used here for computation of support values of all possible item sets at each time point and generates their support sequences and compares the generated support time sequences with a given reference sequence and finds similar item sets.

Keywords: Association rule mining, Support values, Similar item set, Candidate item sets.

1. INTRODUCTION:

Generating the support time sequences by reducing the candidate item sets, using the Association rules discover interrelation ships among various data items in transactional data. Similarity-profiled temporal association mining is to discover all associated item sets whose prevalence variations over time are similar to the reference sequence under a threshold. Similarity-profiled temporal association mining can reveal interesting relationships of data items that co occur with a particular event over time. Using the, time stamped transaction database and a user defined reference sequence of interest over time. One major area of data mining from these data is association pattern analysis. The discovery of association rules paid attention to temporal information, which is implicitly related to transaction data, e.g., the time that a transaction is executed, and discovered association patterns that vary over time. Association patterns can give important insight into many Application domains such as business, agriculture, earth science, ecology, and biology.

2. RELATED WORKS:

Many related works are applicable to this paper that works includes the induction by Bremen et al. and quinoa "classification rules" in 1984 and 1993, Spiegel halter et al. at "discovery of causal rules" in 1993, Muggleton and Feng at "learning of logical definitions" in 1992, Langley et al. at "fitting of functions to data" in 1987, Cheeseman et al. at "clustering" in 1988. The goal of association rule discovery is to find associations among items from a set of transactions, each of which contains a set of items.

Frequent item set generator:

Mohammed Saki discovers Charm frequent item set generator generates the closed frequent item sets and not the association rules on February 25, 2001. Jawed Han's research group at Simon Fraser University discovers FP-growth generates frequent Item sets for association rules from It generates all frequent item sets satisfying a given minimum support by growing a frequent pattern tree structure that stores compressed information about the frequent-growth can avoid repeated database scans and also avoids the generation of a large number of Candidate item sets. Jawed Han and Jean Pei discovers FP-growth implementation in February 5, 2001, it provides the improvement when compared to the earlier version of experimental results. Jawed Han's finds Closet; it is a frequent item set generator for association rules from research group on September 21, 2000 .Jawed Han and Jean Pei provided the closet implementation.

These implementations of FP-growth and Closet only generate the frequent item sets, and not the association rules. Geoff Webb implements the Magnum Opus on February 1, 2001. Magnum Opus directly generates association rules from a dataset based on a specified search preference. Magnum Opus is a unique technique for the search algorithm based on an efficient admissible algorithm for unordered search.

Discovering calendar based temporal association rules:

A temporal association rule is an association rule that holds during specific time

intervals. S.Ramaswamy, S.mahajan, and A.Silberschatz, "on the discovery of interesting patterns in association rules". B.Ozde, S.Ramaswamy, andA.Silberschatz at "cyclic association rules" was extended to approximately discover user-defined temporal patternsin association rules.the work in "on the discovery of interesting patterns in association rules" is more flexible and practical than "cyclic association rules" however it requires user-defined calendar algebraic expressions in order to discover temporal patterns. Y. Li, P. Ning, X.S. Wang, and S. Jajodia at "Discovering calendar based temporal association rules" J.Data and Knowledge Engg., Vol.15, no.2, 2003 uses calendar schema for better result. As a result this implements less priori then "on the discovery of interesting patterns in association rules" and "cyclic association rules". S.Ramaswamy, S.mahajan, and A.Silberschatz, at "on the discovery of interesting patterns in association rules" which discover temporal association rules for one user defined temporal pattern Y. Li, P. Ning, X.S. Wang, and S. Jajodia at "Discovering calendar based temporal association rules" J.Data and Knowledge Engg., Vol.15,no.2,2003 shows all possible temporal patterns in calendar schema. It can potentially discover more temporal association rules.

3. METHODS:

Association rule mining is a popular and well researched method for discovering interesting relationships among various data items in large databases. The existing apriori based temporal association rule mining. Find all frequent item sets which occur at least as frequent as a pre-determined minimum support count and Generate strong association rules from the frequent item sets which satisfy minimum confidence. The problem of mining the temporal pattern is formulated with a similarity-profiled subset specification, which consists of a reference time sequence.

To overcome this problem this paper adopts the Similarity-Profiled temporal Association Mining method (SPAMINE)[1] algorithm divides the mining process into two separate phases.

- The first phase computes the support values of all possible item sets at each time point and generates their support sequences.
- The second phase compares the generated support time sequences with a given reference sequence and finds similar item sets.

A SPAMINE algorithm is developed here for Temporal patterns are searched with a user defined numeric reference sequence and consider the prevalence similarities of all possible item sets, not only frequent item sets. The SPAMINE algorithm will reduce the search space. The computation cost is reduced. Both transaction and time are taken under consideration.

4. MODULE DESCRIPTION:

4.1 Data Preprocessing

The temporal dataset given by the user can have any type of item sets. The proposed system is modeled such that the association rule mining algorithm could process only numeric item sets. So the dataset given by the user is converted into numeric format and then the numeric data is split into different time stamps to which the SPAMINE algorithm can be applied.

4.2 Calculation of Support Time Sequences

Support time sequences can be constructed by two different ways in scanning the time stamped transaction data set.

1. Lattice-dominant scan
2. Snapshot-dominant scan

Lattice-dominant scan: The lattice-dominant scan method reads a whole transaction data set from time slot t_1 to time slot t_{en} for candidate item sets of each size and generates their support time sequences.

Snapshot-dominant scan: The snapshot-dominant scan method repeats the scanning of transactions at each time slot. First, it counts the supports of all candidate item sets of different sizes in the first time slot, and then it moves to the next time slot and repeats the process. This method incrementally generates the support time sequences with the processed time slots.

SPAMINE algorithm uses **Lattice-dominant scan** to read the whole Transaction dataset from time slot t_1 to time slot t_{en} for candidate item sets and generate their support time sequences. The similarity measure used in this algorithm is Euclidean distance.

Euclidean distance between two points (x1,y1) and (x2,y2) is calculated using the

- **$D= [(x2-x1)^2+ (y2-y1)^2]^{1/2}$**

Similarity measure (**Euclidean distance)** is applied to each candidate item set to calculate the following:

1. Upper and lower bound sequence
2. Upper lower-bounding distance

Generating the support time sequences of item sets is the core operation in similarity-profiled association mining algorithm. The operation, however, is very data intensive and sometimes can produce the sequences of all combinations of items. A different way is explored to estimate support time sequences without examining an input data set.

Lower bounding distances are explored to find item sets whose support sequences could not possibly be a best match with a reference sequence under a dissimilarity threshold. In this concept of lower bounding distance, if the lower bounding distance of an item set does not satisfy the dissimilarity threshold, its true distance also does not satisfy the threshold.

Thus, the lower bounding distance can be used to prune item sets early without the computation of the true distances. Lower bounding distance is defined with upper and lower bound support time sequences.

Support time sequences:

Let $D=D_1U\ldots Du_n$ be a set of disjoint transactions. $J=\{J_1, \ldots, J_{oke}\}$ be a set of all size *k*-1 subsets of a size *k* item set I.

Upper bound sequence:

Upper bound support time sequence of item set I, $U_1 =< u_1 , \ldots, u_n>$ is calculated using

- $\mathbf{u_1}$**=*min*{*support* ($\mathbf{J_1}$, $\mathbf{D_1}$), …., *support* ($\mathbf{K_J}$, $\mathbf{D_1}$)}**

Lower Bounding Distance

The upper lower-bounding distance between Rand U, D_{ub}(R, U) is defined to

- $\mathbf{D(R^U , U^L)}$

4.3 Generation of Similar Item Sets

The algorithm compares the calculated distances with the dissimilarity threshold in order to prune the dissimilar item sets and results in similar item sets. The algorithm uses all the calculated values

SYSTEM FLOW DIAGRAM:

- $\mathbf{u_n}$ **=*min*{*support* ($\mathbf{J_1}$, $\mathbf{D_{NS}}$), …., *support* ($\mathbf{K_J}$, $\mathbf{D_{NS}}$)}**

Lower bound sequence:

Lower bound support time sequence of item set I, $L_1 =< l_1 , \ldots, i_n >$ is defined by

- $\mathbf{l_1}$ **=*max*{(*support* ($\mathbf{J_1}$, $\mathbf{D_1}$)+*support* (I- $\mathbf{J_1}$, $\mathbf{D_1}$)-1), …, (*support*($\mathbf{J_{oke}}$, $\mathbf{D_1}$)+*support* (I- $\mathbf{J_{oke}}$, $\mathbf{D_1}$)-1),0 }**
- $\mathbf{i_n}$**=*max*{(*support* ($\mathbf{J_1}$, $\mathbf{D_{en}}$)+*support*(I- $\mathbf{J_1}$, $\mathbf{D_{en}}$)-1), …, (*support* ($\mathbf{J_{oke}}$, $\mathbf{D_{en}}$)+*support*(I- $\mathbf{J_{oke}}$, $\mathbf{D_{en}}$)-1),0 }**

The lower lower-bounding distance between R and L, D_{ib} (R, L) is defined to

- $\mathbf{D(R^U , L^L)}$

The lower-bounding distance,

- $\mathbf{D_{ab}(R, U, L) = D_{ub} (R, U)+ D_{ib}(R, L)}$

such as upper bound support time sequence, lower bound support time sequence and lower bounding distance along with the user defined reference sequence.

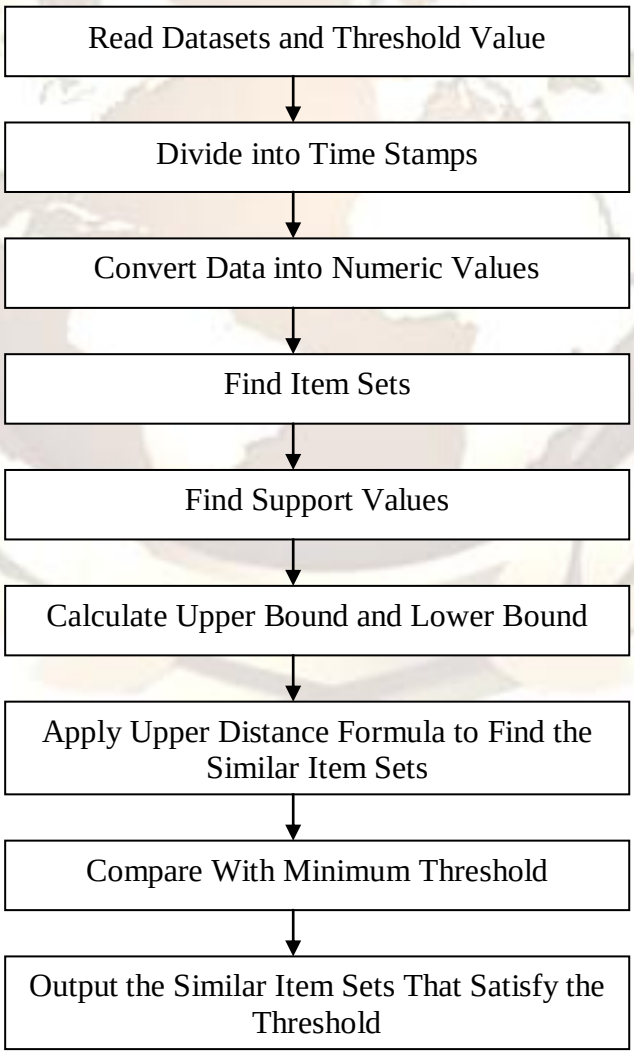

FIGURE 1: FINDING SIMILAR ITEM SETS

here is depends on data distribution,

ESULT:

Experimental results on fake and real data sets showed that the SPAMINE algorithm used here is computationally efficient and can produce meaningful results from real data. The similarity-profiled temporal association mining method algorithm uses the lower bounding distance of the bounds of support sequences, and the monotonicity property of the upper lower bounding distance without compromising the correctness and completeness of the mining results. For the significant reduction of search space by pruning candidate item sets However, the pruning scheme used this type of reference sequence, dissimilarity threshold.

The future improvement may be of different similarity models for temporal patterns can be explored. The current similarity model using a L_p norm-based similarity function is a little rigid in finding similar temporal patterns. It may be interesting to consider not only a relaxed similarity model to catch temporal patterns that show similar trends but also phase shifts in time. For example, the sale of items for cleanup such as chain saws and mops would increase after a storm rather than during the storm. The current project considered whole sequence matching for the similar temporal patterns. Subsequence matching models may be more flexible for the patterns.

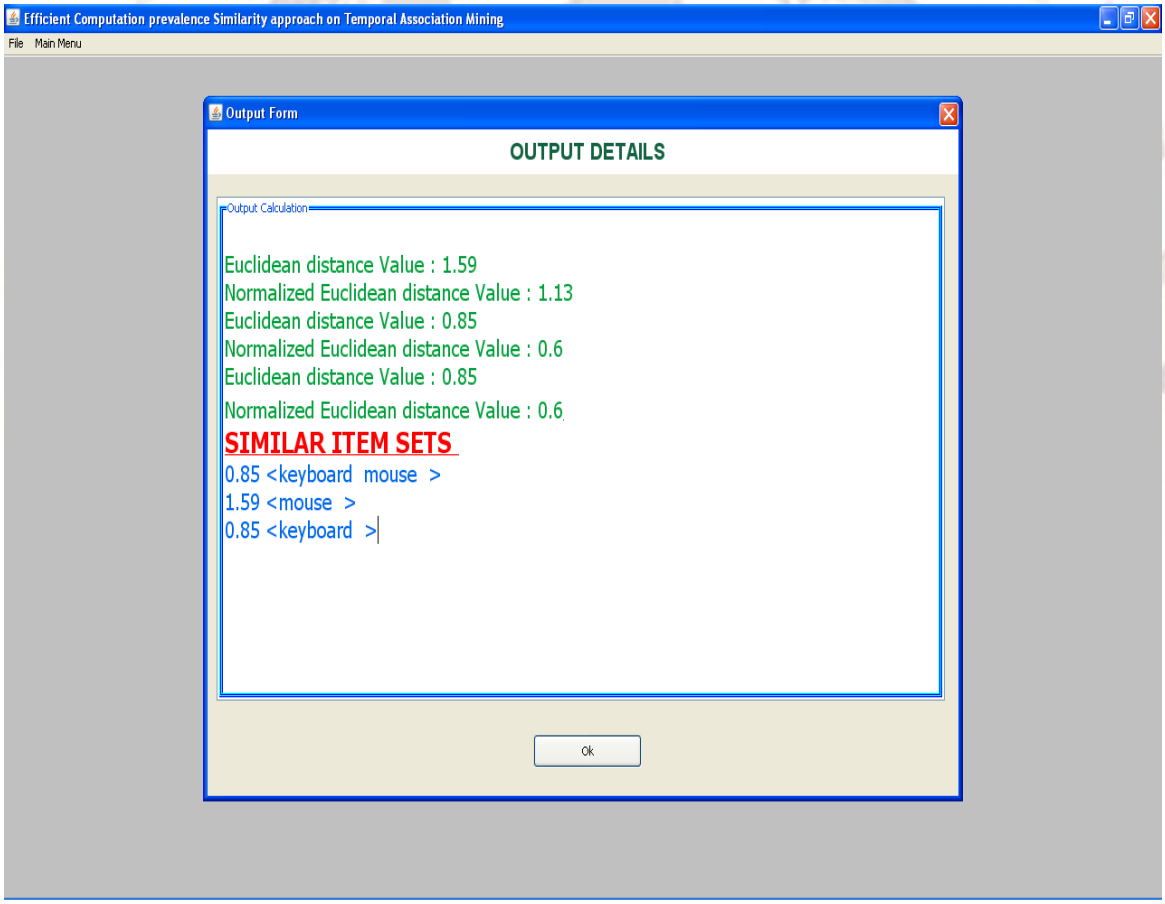

FIGURE 2: FABRICATION OF SIMILAR ITEM SETS

CONCLUSION:

The similarity-profiled temporal association mining method (SPAMINE)[1] algorithm substantially reduced the search space by pruning candidate item sets using the lower bounding distance of the bounds of support sequences, and the monotonicity property of the upper lower-bounding distance without compromising the correctness and completeness of the mining results. The mining problem similarity-profiled temporal association patterns are formulated and an algorithm is proposed to discover them. Experimental results on data sets showed that the SPAMINE algorithm is

computationally efficient and can produce meaningful results from real data.

REFERENCES:

1. Jin Soung Yoo and Shashi Shekhar, "Similarity-Profiled Temporal Association Mining", IEEE 2009.
2. Juan M.Ale and Gustavo H.Rossi R,"An approach to discovering temporal association rules", ACM SIGDD 2002.
3. Keshri Verma and O.P.Vyas, "Efficient calendar based temporal association rule", SIGMOD Record, Vol.34, No.3, 2005.
4. R. Agarwal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Int'l Conf. Very Large Databases (VLDB), 1994.
5. R.Srikant and R.Agrawal, "Mining Generalized Association Rules," Proc. Int'l Conf. Very Large Databases (VLDB), 1995.
6. C.Bettini, X.Wang, S.Jajodia, and J.Lin, "Discovering Frequent Event Patterns with Multiple Granularities in Time Sequences," IEEE Trans. Knowledge and Data Eng., vol. 10, no. 2, Mar.-Apr. 1998.
7. G.Dong and J.Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," Proc. ACM SIGKDD, 1999.
8. B. Liu, W.Hsu, and Y.Ma, "Discovering the Set of Fundamental Rule Change," Proc. ACM SIGKDD, 2001.
9. W.Teng, M.Chen, and P.Yu, "A Regression-Based Temporal Pattern Mining Scheme for Data Streams," Proc. Int'l Conf. Very Large Databases (VLDB), 2003.
10. Y. Li, P.Ning, X.S. Wang, and S.Jajodia, "Discovering Calendar Based Temporal Association Rules," J. Data and Knowledge Eng., vol. 15, no. 2, 2003.