Transition in Time Series Data Mining on Correlated Items

D. Sujatha
Aurora's Technological and Research Institute
Parvathapur,Uppal
Hyderabad

PritiChandra
Scientist,Advanced System Laboratory
Hyderabad

B. L. Deekshatulu
Distinguished Fellow
IDRBTHyderabad

ABSTRACT
We are given a large database of customer transactions, where each transaction consists of transaction-id, the items bought in the transaction and the transaction time. The whole set of transaction is divided into a number of segments called durations (intervals) based on transaction time. And the dividing standard can be monthly, quarterly or yearly. We introduce the problem of mining strong association rules between consecutive durations using FP-tree and correlation coefficient, which is used to quantitatively describe the strength and sign of a relationship between two variables. This paper deals with the changes in the correlation between any two itemsets at the transition of the consecutive duration. Milestone is a change over point between durations. The transition may be positive or negative which are time points at which the pattern is either positively or negatively correlated. Also the method provides rare items, whose support is poor but are highly correlated.

General Terms
Data Mining.

Keywords
Association Rule mining; support; Itemsets; Frequent Patterns; FP-Tree; Correlation; Correlation Coefficient

1. INTRODUCTION
The process of extracting useful information from large quantities of data is known as Data mining. Such useful information is hidden and various techniques are applied to the data to obtain it. Association rule mining is one such method of discovering knowledge hidden in databases.
In general, an association rule denotes a relationship between two items in the same database. It is written in the form X®Y, where X and Y are items and XÇY=Æ. It means that item Y is purchased or taken along with item X. The strength or validity of such rule is shown by two terms namely Support and Confidence. Support specifies the frequency with which items X and Y occur together in the database, while confidence is a measure of strength of the association rule which is basically the number of times purchase of item X has resulted in purchase of item Y. Rules mined from the database are pruned using these two attributes.
Association Rule Mining has been under study for a long time and still efforts to determine association rules or frequent patterns in a flexible, efficient and with minimum mathematical assumptions keep coming. In its early day's association rules were determined by using Apriori algorithm [1] where rules were generated by finding candidates and verifying that their support and confidence meet a predefined minimum support and confidence. This approach was greatly limited because of its redundancy in generating candidates and multiple database scans hence its performance was greatly affected by the size of the database. FP-Growth algorithm [2] followed the Apriori, and it overcame the drawback by eliminating the candidate generation phase and multiple database scans. As time went on many methods and approaches have come up to improve association rule mining. However, association rules generated by such logical means could not prove to be strong. So mathematical approaches have come up such as correlation and principal component analysis to analyze or validate association rules [3].
Correlation mainly, promises to find meaningful association rules by providing information on how an association between two items behaves, i.e. whether the two items are positively correlated (purchase of one items results mostly in the purchase of the other) or negatively correlated (purchase of one items hinders the possibility of purchase of the other) or not correlated (items are independent).

2.RULE EFFECTIVENESS
A frequent pattern is not always static throughout the database. Every pattern in a real-world dataset has a dynamic behavior. The frequencies of the items in the pattern increase/decrease dramatically at some points in the database. There may be many milestones throughout the time period of the database, identifying a significant milestone which is a point in the database where there is a significant change in the frequencies of the items in the itemset. Statistically, plotting the frequencies of an item can give an overview of how it appears throughout the database. But this is not efficient as it requires a massive data interval which is not feasible to analyze.
In order to observe how the two items in a pattern vary with each other at different time points in the database, in [9] method to find lightly supported Boolean association rule is proposed by dividing the data into subsets according to time and then apply Apriori to mine these subsets and get the initialized rule set. Then finds support and confidence of every rule in other time periods and forms the rule matrix. In [6] a method to find significant milestones for a transitional pattern, which are time points at which the frequency of the pattern changes most significantly. In this paper we employ correlation between the two items to identify their behavior. A pattern that is found to be positively correlated after finding its correlation coefficient using the entire database might not always be positively correlated as we proceed through the

database, the items can be negatively correlated during some time interval of the database. Hence finding time points in the database where such changes occur can be quite helpful data.
Milestones provide a wide range of applications. For instance, in the market basket scenario, business owners can identify what combinations of products have become more popular and what combinations of products have lost popularity, identifying this they can plan their business accordingly and position products to in their retail environment. Another application in the medical domains, where data collected from a group of patients with similar disease administered with a new drug, can help identify at what point of time certain symptoms occur and the drug responsible. This way the side effects of a new drug can be indicated

3. CORRELATION

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

3.1.This paragraph is a repeat of 3.1

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

3.2.Title and Authors

The title (Helvetica 18-point bold), authors' names (Helvetica 12-point) and affiliations (Helvetica 10-point) run across the full width of the page – one column wide. We also recommend e-mail address (Helvetica 12-point). See the top of this page for three addresses. If only one address is needed, center all address text. For two addresses, use two centered tabs, and so on. For three authors, you may have to improvise.

3.3.Subsequent Pages

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

Table 1. Table captions should be placed above the table

Graphics	Top	In-between	Bottom
Tables	End	Last	First
Figures	Good	Similar	Very well

3.4.Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4.FIGURES/CAPTIONS

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7").
Captions should be Times New Roman 9-point bold. They should be numbered (e.g., "Table 1" or "Figure 2"), please note that the word for Table and Figure are spelled out. Figure's captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

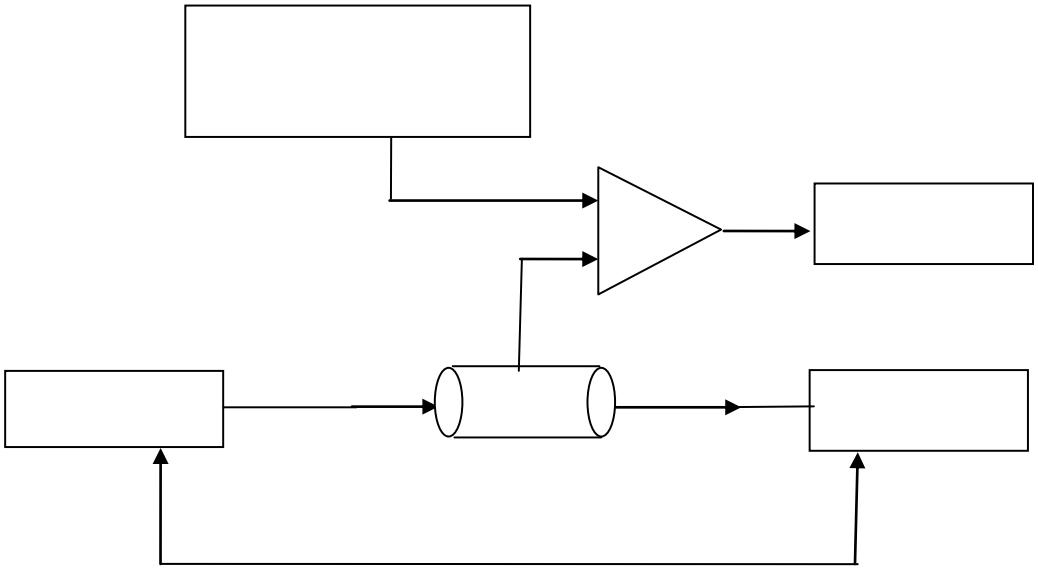

Fig 1: If necessary, the images can be extended both columns

5.SECTIONS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.

5.1.Subsections

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

1.1.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

1.1.1.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

1.1.1.2 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

7. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender