ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

Visit: www.ijmrsetm.com

Volume 1, Issue 2, November 2014

Discovering Probabilistic Frequent Sequential Patterns in Uncertain Databases under Systolic Tree

D.Sugumar, P.Leveen Bose

Department of computer Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.

Department of Computer Science, V.S.B. Engineering College, Karur, Tamil Nadu, India

ABSTRACT: Uncertain data are intrinsic in many real-world applications such as mobile tracking and environment surveillance. Mining sequential patterns from imprecise data, such as those data arising from GPS trajectories and sensor readings are important for discovering hidden knowledge in such applications. We establish two uncertain sequence data models abstracted from many real-life applications involving uncertain sequence data, and formulate the problem of mining probabilistically frequent sequential patterns (or p-FSPs) from data that conform to our models. However, the number of possible worlds is extremely large, which makes the mining prohibitively expensive. Inspired by the famous systolic tree algorithm, we develop patterns that effectively avoids the problem of "possible worlds explosion", and when combined with our pruning and validating methods, achieves even better performance. We also propose a fast validating method to further speedup by enabling the pattern within the boundary.

KEYWORDS: Frequent patterns, systolic tree, possible world semantics, uncertain databases

I. INTRODUCTION

Data mining, is the process of extraction of hidden information from large databases, which is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools forecast future trends and behaviors, allowing businesses to make practical, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by demonstration tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They clean databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

With the rapid advance of wireless communication technology and the increasing popularity of powerful portable devices, mobile users not only can access worldwide information from anywhere at any time but also use their mobile devices to make business transactions easily, e.g., via digital wallet. Meanwhile, the availability of location acquisition technology, e.g., Global Positioning System (GPS), facilitates easy acquisition of a moving trajectory, which records a user movement history. Thus, we envisage that, in the coming future of Mobile Commerce (MCommerce) age, some m-commerce services will be able to capture the moving trajectories and purchase transactions of users. Take the recent announced Shop kick as an example, it gives mobile users rewards and offers when users check-in in stores and on items. Anticipating that some users may be willing to exchange their locations and transactions for good rewards and discounts, we expect more mobile commerce applications, whether they will bear a business model similar with Shop kick or not, will appear in the future. In this project aim at developing pattern mining and prediction techniques that explore the correlation between the moving behaviors and purchasing transactions of mobile users to explore potential M-Commerce features. Owing to the rapid development of the web 2.0 technology, many stores have made their store information, e.g., business hours, location, and features available online, e.g., via mapping services such as Google Map. Additionally, user trajectories can be detected by GPS-enabled devices, when users move around. When a user enters a building, the user may lose the satellite signal until returning to the outdoors. By matching user trajectories

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

Visit: www.ijmrsetm.com

Volume 1, Issue 2, November 2014

with store location information, a user's moving sequence among stores in some shop areas can be extracted. In this proposed system we consider the mining p-FSPs using Systolic tree which accelerates the speed of processing than the existing works. Data mining, which is the exploration of knowledge from the large set of data, generated as a result of the various data processing activities. Frequent Pattern Mining is a very important task in data mining. The previous approaches applied to generate frequent set generally adopt candidate generation and pruning techniques for the satisfaction of the desired objective. Frequent pattern mining in transaction database is one of the well-studied problem in data mining. One obstacle that limits the practical usage of frequent pattern mining is the extremely large number of patterns generated. Such a large size of the output collection makes it difficult for users to understand and use in practice. Even restricting the output to the border of the frequent itemset collection does not help much in alleviating the problem.

II. RELATED WORK

A comprehensive survey of traditional data mining problems such as frequent pattern mining in the context of uncertain data can be found in[4]. We only detail some concepts and issues arising from FSP mining .

A. *Frequent Sequential Pattern Mining on Uncertain Data*

The study is founded on two uncertain sequence data models that are fundamental to many real-life applications in which they propose two new U-PrefixSpan algorithms[1] to mine p-FSPs from data that conform to our sequence level and element-level uncertain sequence models. We also design three pruning rules and one early validating method to speed up pattern frequentness checking. These rules are able to improve the mining efficiency. The experiments conducted on synthetic and real datasets show that our two U-PrefixSpan algorithms effectively avoid the problem of "possible world explosion" and the approximation methods *PA* and *NA* are very efficient and accurate.The approach considers about the problem of mining FSPs in the context of uncertain sequence data. In contrast to previous work that adopts expected support to measure pattern frequentness, we propose to define pattern frequentness based on the possible world semantics.This approach leads to more effective mining of high quality patterns with respect to a formal probabilistic data model. We develop two uncertain sequence data models (sequence-level and element-level models) abstracted from many real-life applications involving uncertain sequence data. Based on the models we define the problem of mining probabilistically frequent sequential patterns (or p-FSPs).It is a difficult task to set rare association rules to handle unpredictable items since approaches frequent pattern-growth, a single minimum Support application based suffers from low or high minimum Support. If minimum support is set high to cover the rarely appearing items it will miss the frequent patterns involving rare Items since rare items fail to satisfy high minimum support. In The literature, an effort has been made to extract rare Association rules with multiple minimum supports. Some problems of the existing system are : Less performance on customer learning and Not adaptive as to customer needs.

III. PROPOSED WORK

With the increasing momentum of the development of technological features,ecommerce- oriented data mining will be a very promising area. It can automatically predict trends in customer spending, market trends which guide company to build personalized business intelligence using mobile commerce technology. In our task, we do not have a predefined class label. In fact, all items in the cart become attributes and the presence/absence of the other items has to be predicted. What are needed are a feasible rule generation algorithm and an effective method to use to this end the generated rules. For the prediction of all missing items in a cart, our algorithm called systolic tree approach that speeds up the computation by the use of the itemset trees (IT-trees) and then uses DS theoretic notions to combine the generated rules. The proposed rule generation algorithm makes use of the flagged IT-tree created from the training data set. The algorithm takes an incoming itemset as the input and returns a graph that defines the association rules entailed by the given incoming itemset.A systolic tree[5] is an arrangement of pipelined processing elements (PEs) in a multidimensional tree pattern. The goal of our architecture is to mimic the internal memory layout of the FP-growth algorithm while achieving a much higher throughput. The role of the systolic tree as mapped in FPGA hardware is then similar to the FP-tree as used in software. Ecommerce companies are faced with a wealth of data, lack of knowledge of discomfiture.To really take advantage of this domain, however, data mining must be integrated into the e-commerce systems with the appropriate data transformation bridges from the transaction processing system to the data warehouse and vice-versa.We proposed this work which provides solution for the best use of these rich data make the e-commerce more effective and analyzing about this data mining in order to fully understand customer preferences, buying patterns, design to meet the needs of different customer groups. Let us consider an example of Mobile Commerce where the

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

Visit: www.ijmrsetm.com

Volume 1, Issue 2, November 2014

users are allowed to use the application to extract the information that they need in considering the sales options. With the help of this application our proposed system could be exposed as shown in the Figure.1.It shows the process of pattern mining by p-FSPs that uses systolic tree approach to speedup the mining process as well as the techniques can be restricted using the pruning methods to specify only the required or relevant data regarding the specified constraints which makes mining effective in the purpose of commercial activities.

A. *Association Rule Learning:*

In data mining, **association rule learning** is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules,Agrawal et al. introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onions,potatoes}⇨{Burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.

B. *Systemflow Diagram*

Based on the request made by the registered client, the pattern mining is indulged to process through the functions such as requesting for the details available, for choosing among the lists provided by the system .The system refers with registered data from the databases seeking for solutions about particular domain, where frequent sequential mining is handled with proposed systolic tree approach to respond back to the request as fast as possible.The domain is considered by tracing the mobile environment through the GPS trajectory technique. The system flow is continued by preprocessing the database of the domains and the pattern is mined which provides required information from the system.The frequent and sequential itemset can be determined by considering the flow between preprocessing technique and frequent pattern mining Database is build and the details of further changes made by requestor is added which maintains the database consistency to look into for gathering updated information. Updated record is utilized for future pattern recognition.

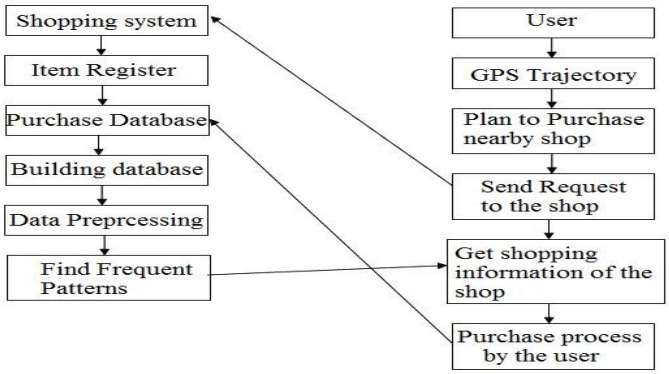

Fig:1 System flow diagram

C. *Mobile Movement Process*

Manage information about objects moving in two- (or higher) dimensional spaces are important for several emerging applications including traffic supervision, flight control, mobile computing, etc. In order to avoid frequent location updates, the database stores the motion function **o**(t) of each object o, which returns its location. M-commerce services will be able to capture the moving trajectories and purchase transactions of users.Mobile trajectories predictions can be used by nonlinear models. The nonlinear models capture objects' movements with sophisticate regression functions. Thus, their prediction accuracies are higher than those of the linear models. Recursive Motion Function (RMF) is the most accurate prediction method in the literature based on regression functions.

D. *Frequent Item Creation*

In this phase, we mine the frequent transactions (FTransactions) for each user by applying the systolic tree algorithm using the mobile transaction database. At first, the support of each (store, item) pair is counted for each user. The patterns of frequent 1-transactionsare obtained when their support satisfies the user-specified minimal support threshold TSUP.A candidate 2-transaction, indicating that two items are purchased together in the transaction, is

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

Visit: www.ijmrsetm.com

Volume 1, Issue 2, November 2014

generated by joining two frequent 1-transactions where their user identifications and stores are the same. For example, the candidate 2-transaction is generated by joining because the user identifications and purchased stores of them both are U1 and F, respectively. Thus, we keep the patterns as frequent 2-transactions, when their support is larger than TSUP.Finally, the same procedures are repeated until no more candidate transaction is generated. We use an item mapping table to re-label item sets in order to present F-Transactions for each unique item set, we use a symbol LIi (Large Item set i) to represent it, where i indicates a running number. The mapping procedure can reduce the time required to check if a mobile commerce pattern is contained in a mobile transaction sequence.

E. *Systolic Tree Process*

A systolic tree is an arrangement of pipelined processing elements (PEs) in a multidimensional tree pattern. The goal of our architecture is to mimic the internal memory layout of the Systolic tree algorithm while achieving a much higher throughput. The role of the systolic tree as mapped in FPGA hardware is then similar to the *U*-PrefixSpan as used in software.It is not always practical or efficient to directly translate a software algorithm into a hardware architecture. Our approach is to build the tree based on the maximum node degree estimation. When the actual node degree at some point in the tree exceeds the estimated node degree, some frequent item set will not be found. Suppose t he number of items in the database is n, t he estimated maximum node degree estimation is K and the estimated depth of the systolic tree is W . Each node in the static tree structure has K children. The total number of nodes in the tree is

$$K^W + K^{W-1} + \cdots + K^1 = \frac{K(K^W - 1)}{K - 1}$$

When K is large, the number of children for each node is large which in turn requires each node to have a large number of interfaces. This will make the inner structure of each node very complex. To simplify the complexity of the node, we assign two instead of K interfaces to each node. One of the two interfaces is dedicated to the connection with its first child, the other one is connected to its nearest sibling. In our systolic tree structure there is only one path tracing back from any PE to the control PE since each PE has a unique parent. The main principle of dictation is that any path containing the queried candidate itemset will be reported to the control node. Note that such path may contain more items than the queried itemset. To clarify the dictation algorithm, we deem there are two doors in each PE. The right door is always open. The bottom door is locked when no data should be sent to the children Fig:2 shows the static systolic structure where *K*=2 and *W*=3.

Each node in the systolic tree architecture is also referred to as a processing element (PE). Each PE has its local data structure and corresponding operations upon receiving signals from outside.. There are three kinds of processing elements in this figure. The root PE is the control node discussed above. The PEs in the rightmost column are the counting nodes which are specifically used for frequent itemset dictation, which we will talk about later. The third kind of processing elements are the general PEs.Each PE in the systolic tree has three modes: WRITE mode,SCAN mode and COUNT mode.

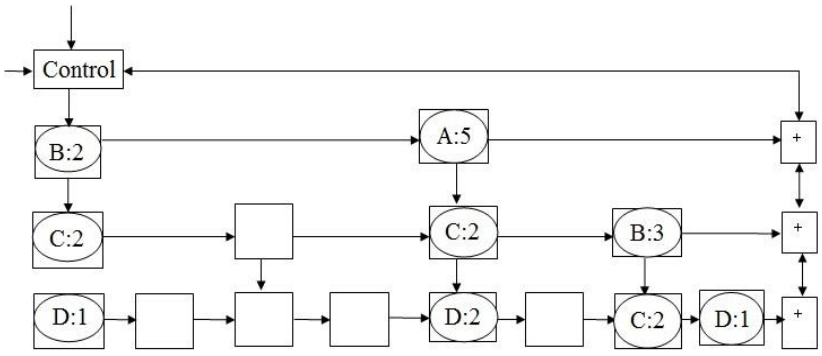

Fig:2 Systolic tree architecture

The design principle of the WRITE mode algorithm is that the built-up systolic tree should have a similar layout with the FP-tree given the same transactional database. The software sends a candidate pattern to the systolic tree. After some clock cycles, the systolic tree sends the support count of the candidate pattern back to the software. The software compares the support count with the support threshold and decides whether the candidate pattern is frequent or not. After all candidate patterns are checked with the support threshold in software, the pattern mining is done. The approach to get the support count of a candidate pattern is called candidate item set (pattern) matching. The SCAN

mode is utilized in determining the dictation process of candidate itemset. The approach used in systolic tree architecture is what we call candidate itemset dictation. When we want to check whether a given itemset is frequent or not, it is sent to the systolic tree. The number of the itemset will be obtained in the output of the systolic tree after some clock cycles. The dictation must be performed after t he systolic tree is built. When the tree is in itemset dictation phase, PEs are in SCAN mode. In our systolic tree To clarify the dictation algorithm, we seem there are two doors in each PE. The right door is always open. The bottom door is locked when no data should be sent to the children. Once all items in a candidate itemset are sent to the systolic tree, a control signal signifying the COUNT mode is broadcasted to the whole systolic tree. The architecture of the systolic tree will change accordingly with response to the COUNT mode signal.

IV. CONCLUSION AND FUTURE WORK

In this paper, we study the problem of mining probabilistically frequent sequential patterns (p-FSPs) in uncertain databases. Our study is founded on two uncertain sequence data models that are fundamental to many real-life applications. We propose systolic tree algorithms to mine p-FSPs from data that conform to our sequence level and element-level uncertain sequence models. We also design a validating method to speed up pattern frequentness checking that can be restricted for designing within the boundary.. The rules implemented are able to improve the mining efficiency. We devise two approximation methods to verify the probabilistic frequentness of the patterns based on Poisson and Normal distributions. The experiments conducted on synthetic and real datasets show that our systolic tree algorithms effectively avoid the problem of "possible world explosion" and the approximation methods *PA* and *NA* are very efficient and accurate. Our preliminary experiments show that with the careful selection of the size of the systolic tree, the mining time can be greatly accelerated compared to current software approaches. The future work can be extended in determining the probability with user specification that authenticates the result with assurance about quality and recognition.

REFERENCES

1 Zhou Zhao, Da Yan, and Wilfred Ng "Mining probabilistically frequent sequential patterns in large uncertain databases",2014.

2 M. Muzammal and R. Raman, "Mining sequential patterns from probabilistic databases," in Proc. 15th PAKDD, Shenzhen, China, 2011.

3 H. Chen, W. S. Ku, H. Wang, and M. T. Sun, "Leveraging spatiotemporal redundancy for RFID data cleansing," in Proc. ACM SIGMOD, Indianapolis, IN, USA, 2010.

4 C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in *Proc. 15th ACM SIGKDD*, Paris, France, 2009.

5 Song Sun and Jo seph Z ambreno "Mining Association Rules With Systolic Trees",2008.

6 F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD*, San Jose, CUSA, 2007.

7 J. Pei *et al*., "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. 17th ICDE*, Berlin, Germany, 2001.