Multidimensional Sequential Pattern Mining

Priyanka Tiwari, Nitin Shukla

Shri Ram institute of Technology
Jabalpur (M.P.), India

Abstract- Data mining is the task of discovering interesting patterns from large amounts of data. There are many data mining tasks, such as classification, clustering, association rule mining, and sequential pattern mining. Sequential pattern mining is the process of finding the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences. It is a data mining task which finds the set of frequent items in sequence database. It is applicable in a wide range of applications since many types of data sets are in a time related format. Besides mining sequential patterns in a single dimension, mining multidimensional sequential patterns can give us more informative and useful patterns. Due to the huge increase in data volume and also quite large search space, efficient solutions for finding patterns in multidimensional sequence data are nowadays very important. In this paper, we discuss about sequential pattern mining, sequential pattern, methods used in sequential pattern mining and we will see how sequential pattern mining is not applicable for mining item set from multidimensional data. And why multidimensional pattern mining is necessary.

Index Terms- sequential pattern mining, sequential pattern, sequential methods, multidimensional pattern mining

I. INTRODUCTION

Data mining is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It is the extraction of the hidden predictive information from large databases is a powerful new technology with great potential to analyze important information in the data warehouse. data mining, a branch of computer science and artificial intelligence, is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

One of the data mining methods is sequential pattern discovery introduced in . Informally, sequential patterns are the most frequently occurring subsequences in sequences of sets of items. Among many proposed sequential pattern-mining algorithms, most of them are designed to discover all sequential patterns exceeding a user specified minimum support threshold. Frequent pattern mining is Sequences are an important kind of data which occur frequently in many fields such as medical, business, financial, customer behavior, educations, security, and other applications. In these applications, the analysis of the data needs to be carried out in different ways to satisfy different application requirements, and it needs to be carried out in an efficient manner .Sequence mining is a topic of Data Mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus Time series mining is closely related, but usually considered a different activity. Sequence mining is a special case of structured data mining. One of the most important data mining problems is discovery of frequently occurring patterns in sequence data. There are many domains where sequence mining has been applied, which include analysis of telecommunication systems, discovering customer buying patterns in retail stores, analysis of web access databases, and mining DNA sequences and gene structures.

In this paper we introduce sequential pattern mining and discuss how sequential pattern mining is not applicable for multidimensional information.

II. SEQUENTIAL PATTERN MINING

Frequently occurring patterns ordered by time are found by sequential pattern mining. Sequential pattern mining has wide application since the data has a time component attached with them. For example, medical domain can determine a correct diagnosis from the sequence of symptoms experienced; over customer data to help target repeat customers; and with web-log data to better structure a company's website for easy accessibility of most popular links.

There are several known methods for discovering general sequential patterns at present. Still in specific domain of web-log analysis more methods exist.

Sequential pattern is a sequence of item sets that frequently occurred in a specific order, all items in the same item sets are supposed to have the same transaction- time value or within a time gap. Usually all the transactions of a customer are together viewed as a sequence, usually called customer-sequence, where each trans- action is represented as an item sets in that sequence, all the transactions are list in a certain order with regard to the transaction-time. Sequences are an important kind of data which occur frequently in many fields such as medical, business, financial, customer behavior, educations, security, and other applications. In these applications, the analysis of the data needs to be carried out in different ways to satisfy different application requirements, and it needs to be carried out in an efficient manner. It is obvious that time stamp is an important attribute of each dataset, and it can give us more accurate and useful information and rules. A database consists of sequences of values or events that change with time are called a time series database. This type of database is widely used to store historical data in a diversity of areas. One of the data mining techniques which have been designed for mining time series data is sequential pattern mining. Sequential pattern mining is trying to find the relationships between occurrences of sequential events for looking for any specific order of the occurrences. In the other words, sequential pattern mining is aiming at finding the frequently occurred sequences to describe the data or predict future data or mining periodical patterns. To gain a better understanding of sequential pattern mining problem, let's start by looking at an example. From a shopping store database, we can find frequent sequential purchasing patterns, for example "70% customers who bought the TV typically bought the DVD player and then bought the memory card with certain time gap." It is conceivable that achieving this pattern has great impact to better advertisement and better management of shopping store.

Example- An example of a sequential pattern is "A positive response in trade-in option is expected from a customer who purchased a new Ford Explorer two years ago". If X is the clause "purchased a new Ford Explorer" and Y be the clause "a positive response in trade-in". Then notice that the pattern XY, is different from pattern YX which states that "A customer shall purchase a Ford Explorer now, who responded positively to a trade-in two years ago". The order in which X and Y appear is important, and hence XY and YX are mined as two separate patterns.

III. METHODS FOR SEQUENTIAL PATTERN MINING

There are two types of methods used in sequential pattern mining. First is Apriori based approaches and second is Pattern growth based approaches.

Apriori based approaches- There are two algorithms in this approach GSP and SPADE.

A) GSP

GSP Algorithm (Generalized Sequential Pattern algorithm) is an algorithm used for sequence mining. The algorithms for solving sequence mining problems are mostly based on the a priori (level-wise) algorithm. One way to use the level-wise paradigm is to first discover all the frequent items in a level-wise fashion. It simply means counting the occurrences of all singleton elements in the database. Then, the transactions are filtered by removing the non-frequent items. At the end of this step, each transaction consists of only the frequent elements it originally contained. This modified database becomes an input to the GSP algorithm. This process requires one pass over the whole database.

B) SPADE

SPADE is a fundamentally different sequential pattern algorithm. In place of repeated database scans, this method uses lattice-search techniques and simple join operations to discover all sequence patterns. First a vertical id-list is created to associate with each item, a list of the sequences in which it occurs, along with the appropriate time-stamps.

Pattern Growth based approaches- There are two algorithms in this approach Free SPAN and Prefix Span.

C) Free Span

FreeSpan was developed to substantially reduce the expensive candidate generation and testing of Apriori, while maintaining its basic heuristic. In general, FreeSpan uses frequent items to recursively project the sequence database into projected databases while growing subsequence fragments in each projected database. Each projection partitions the database and confines further testing to progressively smaller and more manageable units. The trade-off is a considerable amount of sequence duplication as the same sequence could appear in more than one projected database. However, the size of each projected database usually (but not necessarily) decreases rapidly with recursion.

D) Prefix Span

Prefix Span was developed to address the costs of Free Span. Its general idea is that, instead of projecting sequence databases by considering all the possible occurrences of frequent subsequences, the projection is based only on frequent prefixes because any frequent subsequence can always be found by growing a frequent prefix.

IV. WHY SEQUENTIAL PATTERN MINING IS NOT ENOUGH

The answer is very simple current sequential pattern algorithms mine only one dimension. For example in case of mining for user-access patterns from a web-log we track for the in time order accessed pages. For date there is no sequential algorithm that can allow efficient mining of ordered and unordered data together, along with the elimination of the trivial solution of simply creating a larger set of individually specific items.

Example- Presume one of the sequential items is web-page A, and there is extra information in the form of IP addresses 1, 2 and 3. Then inclusion of the IP dimension can be done by way of creating the new items A1, A2 and A3. This is one of the ways, but it unnecessarily increases the number of items sequential mining, and since it is already expensive method, this is not efficient.

V. NEED OF MULTIDIMENSIONAL PATTERN MINING

A**s** we have seen above Sequential pattern mining uses only single dimensional data set. All the methods used in sequential pattern mining works on only single dimensional data set. Because of this drawback multidimensional sequential pattern mining is necessary for mining item sets from multidimensional data set.

VI. MULTIDIMENSIONAL PATTERN MINING

When one or more dimensions of information is mined and the order of the dimension values is not important, it is known as Multi-dimensional sequential pattern mining. Besides mining sequential patterns in a single dimension, which means while finding the frequent patterns we only consider one attribute together with time stamps, mining multiple dimensional sequential patterns can give us more informative and useful patterns. For example we may get the frequent pattern from the supermarket database that most people who buy package 1 also buy package 2 in a defined time interval by employ in general sequential pattern mining. However using multiple dimensional sequential pattern mining we can further end different groups of people have different purchase patterns. For example, students always buy A within a week after they buy B, while those sequential rules do not hold for other groups. Multi-dimensional sequential pattern mining is first introduced in . In multidimensional sequential pattern mining, different attributes of the transaction ID were introduced and formed a multi-dimensional sequential data sets as shown in Table VI. The aim of this special mining is to get more interesting sequential patterns with different dimensional attributes. One direct extension of Prefix Span algorithm and two approaches combined with Prefix Span and BUC (Bottom up Computation)-like algorithms will be briefly introduced.

It can be expressed with the help of an example. Consider a dataset D who purchase records for people who live in Canada, which shall include their province of residence, occupation and education level. On the base of this dataset, following query can be answered.

Question- What shall be the purchasing pattern for entrepreneurs who reside in BC and possess a graduate degree?
Example- Three dimensions in dataset D are relevant for this problem: Province (P), Occupation (O) and Education (E). These are called the task-relevant dimensions. Further assumption is that these dimensions have the following individual values.

- Province: BC (P_1) AB (P_2) ON (P_3)
- Occupation: Entrepreneurs (O_1) Op. Managers (O_2) Teachers (O_3)
- Education: Graduate (E_1) Under-graduate (E_2) High-school (E_3)

To find the suitable subset of records for mining task-related dimensions are used. One record stands for the series of purchase transactions made by one customer, over a precise time period (week/month/year/etc.) For example, $P_1 \cap O_1 \cap E_1$ represent the subset of records required to answer the above query. Assume that the chronological patterns mined from $P_1 \cap O_1 \cap E_1$ are < (Moore's suit, silk shirts)>: 7, <TV, VCR>: 11 … The first pattern can be interpreted as there are 7 customers, who bought a Moore's suit and a silk shirt together in one transaction. The second pattern interprets that 11 customers in this category bought a TV in first transaction, and then a VCR in later transaction.

Noticeable fact is that there is only one ordered dimension of purchase transactions themselves. All the other dimensions (P, O, and E) are unordered and has no relevance in specification given $P_1 \cap O_1 \cap E_1$ or $P_1 \cap E_1 \cap O_1$; it still describes the same set. When the transactions alone shall be mined would be sequential pattern mining. To mine the transactions along with the other unordered dimensions is multi-dimensional sequential pattern mining. User can gain much more about the conditions when these other unordered dimension values are present which enriches the discovered sequential patterns.

VII. Conclusion

The research area of sequential pattern mining has focused on multidimensional sequential pattern mining. In this paper we discussed about sequential pattern mining then what sequential pattern is, methods which used in sequential pattern mining GSP, SPADE, FreeSpan, Prefix Span. We have seen that sequential pattern mining and its entire methods only works on single dimension of information. Then we have discussed about the drawbacks of sequential pattern mining or we can say that why sequential pattern mining is not suitable for multidimensional data set. Then we have discussed about the need of multidimensional pattern mining and then multidimensional pattern mining.

References

[1] Dong G., and Pei J., Sequence Data Mining, Springer, 2007.

[2] Mahdi Esmaieli and Mansour Tarafdar Sequential Pattern Mining from Multidimensional Sequence Data in Parallel , 2010.

[3] Qiankun Zhao, Sourav S. Bhowmick, Sequential pattern mining 2008.

[4] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Academic Press, New York, 2001.

[5] J. Han, J. Pei, K. Wang, Q. Chen, and U. Dayal, Multi-dimensional sequential pattern mining, Conference on Information and Knowledge Management, pp: 81–88, 2001 .

[6] C.-C. Yu, and Y.-L. Chen, Mining Sequential Patterns from Multidimensional Sequence Data, IEEE Transactions on Knowledge and Data Engineering archive, Volume 17, Issue 1, pp: 136-140, 2005.

[7] R. Agrawal, R. Srikant, Mining Sequential Patterns, Proc. of the 11th Int'l Conference on Data Engineering, 1995.

[8] C. Lucchese, S. Orlando and R. Perego, "DCI-CLOSED: A Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets," In Proc. of IEEE ICDM workshop on Frequent Itemset Mininig Implemention (FIMI'04), Vol. 126, Brighton, UK, November 2004.

[9] www.google.com

[10] www.wickipedia.com

Authors

First Author – Priyanka Tiwari, M.E- C.S.E. IVth Semester, Shri Ram institute of Technology Jabalpur (M.P.), India.
Email id - Priyanka8684@gmail.com

Second Author – Nitin Shukla, Lecturer M.C.A. Dept, Shri Ram institute of Technology Jabalpur (M.P.), India.