Sequential Pattern Mining – A Study

S.Vijayarani
Assistant professor
Department of computer science
Bharathiar University

S.Deepa
M.Phil Research Scholar
Department of Computer Science
Bharathiar University

ABSTRACT
Data mining is the process of identifying the valid information from large databases. There are many different tasks in data mining such as classification, clustering, prediction, time series analysis, sequence pattern mining, etc. The Sequential Pattern Mining is used to find sequential patterns that occur in large databases. It also identifies the frequent subsequences as patterns from a sequence database. In real world, as massive amount of data are needed to be collected continuously and stored in the databases. Many industries are becoming interested in mining sequential patterns from these databases. This paper provides a systematic study on sequential pattern mining methods. It also deals with the analysis of various research problems and challenges in sequential pattern mining.

Keywords
Sequential Pattern Mining, Apriori Methods, Pattern Growth methods,Time Interval Sequence Pattern, Closed sequential pattern mining, Target oriented sequential pattern

1. INTRODUCTION
Sequential Pattern Mining is the method of finding interesting sequential patterns among the large databases. It also finds out frequent subsequences as patterns from a sequence database. Enormous amounts of data are continuously being collected and stored in many industries and they are showing interests in mining sequential patterns from their database. Sequential pattern mining has broad applications including web-log analysis, client purchase behavior analysis and medical record analysis.

Sequential or sequence pattern mining is the task of finding patterns which are present in a certain number of instances of data. The identified patterns are expressed in terms of sub-sequences of the data sequences and expressed in an order that is the order of the elements of the pattern should be respected in all instances where it appears. If the pattern is considered to be frequent if it appears in a number of instances above a given threshold value, usually defined by the user, then it is considered to be frequent.

There may be huge number of possible sequential patterns in a large database. Sequential pattern mining identifies whether any relationship occurs in between the sequential events. The sequential patterns that occur in particular individual items can be found and also the sequential patterns between different items can be found. The number of sequences can be very large, and also the users have different interests and requirements. If the most interesting sequential patterns are to be obtained, usually a minimum support is pre-defined by the users. By using the minimum support, sequential patterns which are not so important is taken out and hence the mining process will be more efficient. However some sequential patterns that do not satisfy the support threshold are still interesting.

The rest of the paper is organized as follows. In Section 2, the basic concepts of sequential pattern mining are discussed. Section 3 describes various methods that are in the mining of sequential patterns and the use of various algorithmic techniques. Section 4 describes other time related patterns. Section 5 describes the sequence pattern mining problems. Section 6 describes application areas of sequential pattern mining. Section 7 describes the research challenges that are involved in the sequential mining process.

2. BASIC CONCEPTS OF SEQUENTIAL PATTERN MINING [1]

1. Let $X = \{i_1. . . i_n\}$ be a set of items, each being associated with a possible set of attributes. The value of an attribute A of item i is denoted by *i.A*. An itemset consists of a non-empty subset of items and an itemset with k items is called a k-itemset .

2. A sequence $\alpha = <A_1 \cdot \cdot \cdot A_n>$ is an ordered list of itemsets. An itemset A_i $(1 \leq i \leq l)$ in a sequence is called a transaction. A transaction A_i may have a special attribute, time-stamp, denoted by $A_i.time$, which registers the time at which the transaction gets executed. For a sequence $\alpha = <A_1 \cdot \cdot A_n>$, assume $A_i.time<A_j.time$ for $1 \leq i < j \leq l$.

3. The length of a sequence is denoted by the number of transactions that are present in a sequence. A sequence with length L is called an L-sequence. For a L-sequence β, length of β is denoted by len(β)=l. Then jth itemset can be denoted by β[j]. Maximum an item can occur one time in an itemset, but can also occur multiple times in various itemsets in a sequence.

4. A sequence α = <A1 . . . An> is called a subsequence of another sequence β = <B1 . . .Bm> (n ≤ m), and β a super-sequence of α, if there exist integers $1 \leq i_1 < . . < i_n \leq m$ such that $X_1 Y_{i1}$, . . . , X_n Y_{in}.

5. A sequential database is a set of 2-tuples (sid, β), where sid is a sequence-id and β is the sequence. A tuple (sid, β) in a sequence database is said to contain a sequence λ if λ is a subsequence of β. The number of tuples in a sequence database containing sequence λ is called the support of λ, denoted by sup (λ). Given a positive integer minimum_sup as the support threshold, a sequence λ is considered to be a sequential pattern in sequence database SDB if sup (λ) ≥ minimum_sup. The sequential pattern mining problem is used to find the complete set of sequential patterns with respect to a given sequence database and a support threshold minimum_support[1].

The sample example for sequential pattern mining:

Given a certain set of sequences, each sequence contains a list of elements and each element contains set of items and given a user-specified minimum_support threshold, sequential pattern mining is used to find all the frequent subsequences whose occurrence frequency in the set of sequences in no less than minimum_support.

A sequence database

Sequence id	Sequence
10	<p(pqr)(pr)d(ru)>
20	<(ps)r(qr)(pt)>
30	<(tu)(pq)(su)rq>
40	<tv(pu)rqr>

<p(qr)sr> is a subsequence of <p(pqr)(pr)s(rv)> Given support threshold minimum support=2,((pq)r> is a sequential pattern.

3. SEQUENTIAL PATTERN MINING METHODS

There are two main research issues in sequential pattern mining [1].

1. The first issue lies in improving the efficiency in process of sequential pattern mining.
2. The mining of sequential pattern may be extended to other time-related patterns.

3.1 Improving the efficiency by formulating diverse algorithms:

The algorithms for Sequential Pattern Mining mainly differ in two ways:

(1) The algorithms may differ in the way in which candidate sequences are generated and stored. The main goal of these algorithms is to reduce the number of candidate sequences generated in order to minimize I/O cost.

(2) They may also differ in the way in which the support is counted and how the candidate sequences are tested for frequency. The key strategy here is to eliminate any database or data structure that has to be maintained all the time for support of counting purposes only.

Based on these conditions sequential pattern mining can be divided broadly into three parts:

•Apriori based

•Pattern growth based

3.1.1 APriori-Based Approaches:

The Apriori method of sequences states that if a sequence S is not frequent, then the subsequences of S are also not frequent. It is also described as anti monotonic [3] property (or downward-closed).

The first pass of the algorithm simply counts the occurrences of the items to determine the frequent itemsets. A subsequent pass k consists of two steps. First the frequent itemsets L k-1 found in the (k-1)th pass are taken to generate the candidate itemsets Ck using apriori candidate generation. Then the scanning of the database is performed and the support of candidates in Ck is counted. The set of candidate itemsets are then pruned out to ensure that all the subsets of the candidate sets are already known to the frequent itemsets. If a sequence fails in the minimum support test, then the entire subsequences will also fail in the test.

Key features of Apriori-based algorithm are[16]:

(1)Breadth-first search: The algorithms in apriori-based approach are described as breath-first search algorithms because they construct all the k-sequences, in kth iteration of the algorithm, as they traverse the search space.

(2)Generate-and-test: Algorithms based on this feature display an inefficient pruning method and produces an explosive number of candidate sequences and then tests each one by one until some user specified constraints are satisfied. This method consumes a lot of memory in the early stages of mining.

(3)Multiple scans of the database: The original database is scanned to check whether a long list of generated candidate sequence is frequent or not. It is a very undesirable characteristic of most apriori-based algorithms and requires a lot of processing time and I/O cost.

In A priori based approach, there are three algorithms namely GSP, SPADE and SPAM.

GSP Algorithm:

GSP Algorithm (Generalized Sequential Pattern Algorithm) is used in solving various issues in sequence mining. It is a level wise algorithm where all the frequent items are found level-wise and all singleton items are counted. Then the non-frequent items are eliminated. Finally, each transaction contains only the frequent items that it originally contained. This database is given as the input to the GSP algorithm. GSP makes several passes over the database. In the first pass, all single items are (1-sequence) are counted. From the frequent items, a set of candidate 2-sequences are computed and another pass is made to identify their support. The frequent 2-sequences are used to produce the candidate 3-sequences and this process is repeated until no more frequent sequences are identified.

SPADE Algorithm:

Sequential Pattern Discovery using Equivalence classes is a form of vertical format sequential pattern mining method that uses lattice-search techniques and simple join operations to find all sequence patterns instead of repeated database scans. A vertical id-list associated with each item is created first along with time stamps [10]. Spade appears to be an efficient method for frequent sequence mining.

SPAM Algorithm:

SPAM approach is an integration of GSP, SPADE and Free Span algorithms. This algorithm is very useful when the database is too large. A depth first method is used to generate the candidate sequences and various pruning methods are used with the idea of reducing the search space. SPAM traverses the sequence tree in depth-first search fashion [1] and checks the support of each sequence-extended or itemset extended child against minimum_sup recursively for efficient support-counting.

3.1.2 Pattern Growth Based Approach

Pattern growth approaches can be considered as depth-first traversal algorithms since they recursively generate the projected database for each length-k pattern to find length-(k+1) patterns. They emphasize the search on a restricted portion of the initial database to avoid the expensive candidate generation and test step. Pattern-growth approach is a more incremental method in producing the possible frequent sequences and it uses the divide-and-conquer technique. Pattern-growth algorithms make projections of the database in an attempt to reduce the search space.

Key features of pattern growth-based algorithm are[16]:

(1)Search space partitioning: It allows partitioning of the generated search space of large candidate sequences for efficient memory management. There are different ways to partition the search space. After partitioning the search space, smaller partitions can be mined in parallel. Advanced techniques for search space partitioning include projected databases and conditional search, referred to as split-and-project techniques.

(2)Tree projection: Tree projection usually accompanies pattern-growth algorithms. Here, algorithms implement a physical tree data structure representation of the search space which is then traversed breadth-first or depth-first in search of frequent sequences and pruning is based on the apriori property.
(3) Depth-first traversal: This approach helps in the early pruning of candidate sequences as well as mining of closed sequences. The main reason for this performance is the fact that depth-first traversal utilizes much less memory, more directed search space and thus the candidate sequence generation is lower than the breadth-first or post-order which are used by some early algorithms.
(4) Candidate sequence pruning: Pattern-growth algorithms try to utilize a data structure that allows them to prune candidate sequences early in the mining process. This result in early display of smaller search space and maintain a more directed and narrower search procedure.

This approach consists of three algorithms namely FreeSpan, PrefixSpan and Wapmine.

FreeSpan:

In FreeSpan, frequent items are used to recursively project the sequence database into projected databases while finding subsequence fragments in each projected database. Each projection divides the database and performs further testing to progressively smaller and more manageable units [9].

PrefixSpan:

PrefixSpan is capable of dealing very large databases. PrefixSpan mainly employs the method of database projection to make the database for next pass much smaller and consequently make the algorithm speedier. Prefix Span finds the frequent items after scanning the sequence database one time. The database is then projected according to the frequent items into a number of smaller databases. Finally, a complete set of sequential patterns is identified by recursively growing the subsequence fragments in each projected database [8].

Wap Mine:

It is a tree structure-mining technique along with its WAP-tree structure. Here the sequence database is scanned twice to build the WAP-tree from frequent sequences along with their support. A header table [1] is maintained to point at the first occurrence for each item in a frequent itemset. Then it is later tracked in a threaded way to mine the tree for frequent sequences building on the suffix.

3.2 Extensions of Sequential Pattern Mining to Other Time-Related Patterns:

Sequential pattern mining has been widely studied in the recent years. There exist numerous algorithms for sequential pattern mining. Numerous extensions of the initial definition of algorithms have been found which may be related to other types of time-related patterns or to the addition of time constraints. These algorithms are to be used in special cases such as multidimensional, closed and constraint based sequential pattern mining.

3.3 Two Main Frameworks of Sequential Mining

Type1: Sequential pattern mining for a single data sequence

The mining of sequential patterns with single dimension data [2] means only one attribute along with time stamps is considered in the pattern discovery process.

Type2: Multidimensional Sequential Pattern Mining

Mining multiple dimensional sequential patterns [4] can give more informative and useful patterns. For example, a traditional sequential pattern may be obtained from the supermarket database that after buying product A most people also buy product B in a defined time interval. However, by using multidimensional sequential pattern mining, different groups of people with different purchase patterns can be found.

4. OTHER TIME RELATED PATTERNS

4.1Discovering Constraint Based Sequential Pattern

Although the efficiency of mining the complete set of sequential patterns has been developed, sequential pattern mining still have tough challenges in both effectiveness and performance in certain cases. There could be a huge number of sequential patterns in a large database. The users are interested in only a small subset of such patterns. It is very complex to present and understand a complete set of sequential patterns in the mining process.

To overcome this problem, the problem of using various constraints deep into sequential pattern mining using pattern growth methods was introduced. Constraint-based mining [3] may overcome the difficulties of effectiveness and efficiency since constraints usually represent user's interest and focus, which limits the patterns to be found to a particular subset satisfying some strong conditions.

4.2 Discovering Time-interval Sequential Pattern

A Time interval sequential pattern can provide more useful information than traditional sequence pattern mining methods.Although sequential patterns can tell what items are frequently bought together and the order in which they are bought, they cannot provide information about the time span between items for further decision support [12]. The solution to this problem is to generalize the mining problem into discovering time-interval sequential patterns [3], which tells not only the order of items but also the time intervals between successive items.

4.3 Closed Sequential Pattern Mining

The sequential pattern mining algorithms discussed so far shows good performance in databases consisting of short frequent sequences. In certain cases, when long frequent sequences are to be mined or when very low sustainable thresholds are used, the performance of such algorithms often degrade dramatically. Among several sequential patterns methods, closed sequential pattern is the considered to be most significant one since it holds all the information about the complete pattern set.

Assume that the database contains only one long frequent sequence <(a1) (a2) . . . (a100) >, it will generate 2100−1 frequent subsequence if the minimum support is one, although all of them except the longest one are redundant because they have the same support as that of <(a1) (a2) . . . (a100) >. So an alternative was proposed that provides successful solution. Instead of mining the complete set of frequent subsequence, frequent closed subsequences are only mined, i.e., those containing no super-sequence with the same support.

4.4 Target-Oriented Sequence Pattern

A target-oriented sequential pattern consists of a concerned itemset at the end of pattern. In decision making, when the users want to make efficient marketing strategies, they usually consider the happening order of concerned itemsets only and this makes the sequential patterns discovered by using traditional algorithms to be irrelevant and useless[13].

5. SEQUENCE PATTERN MINING PROBLEM

There are several computational problems within the sequence pattern mining field. These may include building efficient databases and indexes for sequence information and finding out the frequently occurring patterns. Then the sequences are compared for similarity and recovering missing sequence members. Sequence mining problems can be classified as string mining and itemset mining.

String Mining:
String mining typically deals with limited alphabets for items that appear in a sequence. In biological applications,the analysis on the arrangement of the alphabet in strings can be used to analyze the gene and protein sequences to determine their properties. The major task here is to understand the sequence in terms of its structure and biological function [11].

Itemset Mining:

Some problems in sequence mining make themselves to discover frequent itemsets and their order of occurrence. For example, one is seeking rules of the form "if a {customer buys a laptop}, he or she is likely to {buy mouse} within 1 week", or in the context of stock prices, "if {TCS up and WIPRO Up}, it is likely that {CTS up and IBM up} within 2 days". This method is applied in marketing applications for discovering regularities between frequently co-occurring items in large transactions.

6. APPLICATION AREAS

Sequential pattern mining is used in a variety of fields. The main goal of sequential pattern mining is to find the frequent subsequences in a dataset. Sequential pattern mining has numerous applications in the area of medicine, DNA sequencing, Web log analysis, computational biology. Large amount of sequence data have been and continue to be collected in genomic and medical studies, in security applications, in business applications, etc. 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 effective manner.

In the medical field, sequential patterns for symptoms and diseases exhibited by patients identify strong symptom or disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine. The sequential pattern mining of the medical records can reveal how much time a virus can take to cause a disease after it has infected somebody. Time-interval sequential patterns enable doctors to prevent their patients from becoming infected by other viruses. They can also help doctors to make good decisions when diagnosing their patients' illnesses [15].

In DNA sequencing, the DNA sequence pattern consists of four kinds of alphabets (A, G, T, and C). The pattern also includes a gap in the DNA sequence. By finding out the DNA sequence with gap, the unknown sequences can be found and classified into its corresponding DNA family and this can be used for further research in biological analysis [11].

Sequential Pattern mining is useful in the field of bioinformatics for predicting rules for organization of certain elements in genes and for protein function prediction. It is also used in gene expression analysis and for protein fold recognition [14].

Sequential Pattern mining can be used in the field of telecommunications for mining of group patterns from mobile user movement data and for predicting future location of a mobile user for location based services. It can also be used for customer behavior prediction and for mining patterns useful for mobile commerce [14].

Sequential pattern mining can be done on web logs. In Web log analysis [5], the exploring behavior of a user can be extracted from member records or log files. In this case the sequences could be sequences of WebPages visited by users on a website. Then a sequential pattern mining algorithm could use sequential pattern mining to discover sequences of web pages that are often visited by users. The website could use these patterns to generate suggestions to the user such as recommended links. Sequential data is collected periodically from web server logs, online transaction logs, performance measurement. This data helps in searching for a particular value or event and it may also assists in the analysis of the frequency of certain events or sets of related events.

A time interval sequential pattern mining has great advantage in various areas. In the field of retailing, it discovers the sequential patterns in the transaction record details of the customers. In Police department, time interval sequence pattern mining can be used to predict when a criminal may commit a crime again or which district may suffer such a crime.

Business enterprises make use of sequential pattern mining to study the customer behaviors. It is also used in the analysis of system performance and telecommunication network analysis. Sequential pattern mining is used to analyze the mutation patterns of different amino acids in computational biology.

7. RESEARCH CHALLENGES

Today several methods are available for efficiently discovering sequential patterns according to the initial definition. Such patterns are widely used for a large number of applications. But still there are various research challenges in this field of data mining. Some of the research challenges [1] are:

- Finding the complete set of patterns and satisfying the minimum support (frequency) threshold is a complex task. When the database is large, distributed sequential pattern mining is used for mining process which helps to increase the scalability.
- The ability to incorporate various kinds of user-specific constraints is a complex process. To add other useful constraints to the RFM patterns, for example, the constraint that the number of repetitions in a sequence must be no less than a given threshold.
- Constraints like frequency and Monetary constraints are difficult to be studied and checking their effect with respect to execution time, memory usage and scalability is also difficult.
- Algorithm should handle large search space. Repeated scanning of the database during the mining process must be reduced as much as possible. To introduce the concept of object-orientedness in sequential pattern

mining, by which there will be flexibility of mining only on the focused parts of the database.

- The target oriented sequential pattern mining and its application in real dataset is difficult to proceed. Various methods are used by which early candidate sequences are pruned and search space partitioning will be possible for efficient mining of patterns.
- There are many interesting problems especially in the development of specialized sequential pattern mining methods for particular applications such as DNA sequence mining[1] that may identify faults which in turn allows various insertions, deletions, and mutations in DNA sequences, and handling industry or engineering sequential process analysis are interesting issues for future research.
- The mining of multi-level time-interval sequential patterns are performed by using fuzzy time value.

8. CONCLUSION

Sequence pattern mining is gaining importance in today's world since it assists in finding the relationships among the data in an effective manner. In this paper we have discussed about sequential pattern mining, and the methods which are used in sequential pattern mining. Due to the continuous addition of large amount of data in the databases, the idea of sequential pattern mining is becoming popular. Various algorithms have been developed that are used for mining the sequential patterns in the data. These algorithms have proved to be more effective for smaller databases, but when the size of the database is increased, their performance may decline. Hence these methods have to be improved in order to perform the mining processes in a better way.

9. REFERENCES

[1] Chetna Chand, Amit Thakkar, Amit Ganatra- Sequential Pattern Mining: Survey and Current Research Challenges, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[2] Koji Iwanuma, Hidetomo Nabeshima - A Short Introduction to Sequential Data Mining, The First Franco-Japanese Symposium on Knowledge Discovery in System Biology, September 17, Aix-en-Provence.

[3] Rakesh Agrawal Ramakrishna Srikant, ―Mining Sequential Patterns‖, 11th Int. Conf. on Data Engineering, IEEE Computer Society Press, Taiwan, 1995 pp. 3-14.

[4] Priyanka Tiwari, Nitin Shukla- Multidimensional Sequential Pattern Mining, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012 1 ISSN 2250-3153

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

[6] Qiankun Zhao, Sourav S. Bhowmick-Sequential Pattern Mining: A Survey

[7]Helen Pinto- Multidimensional Sequential Pattern Mining, © Helen Pinto 2001, Simon Fraser University, April 2001

[8]Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu-PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth.

[9] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu-FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining.

[10] Mohanned J.Zaki- SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning, 42, 31–60, 2001

[11]Syeda Farzana, Byeong-Soo Jeong-A Fast Contiguous Sequential Pattern Mining Technique in DNA Data Sequences Using Position Information.

[12] Yen-Liang Chen, Mei-Ching Chiang, Ming-Tat Ko-Discovering time-interval sequential patterns in sequence databases, Expert Systems with Applications 25 (2003) 343–354.

[13] Hao-En Chueh-Mining Target-Oriented Sequential Patterns With Time-intervals, International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[14] Manish Gupta, Jiawei Han- Applications of Pattern Discovery Using Sequential Data Mining

[15] Yen-Liang Chen, Mei-Ching Chiang, Ming-Tat Ko-Discovering time-interval sequential patterns in sequence databases, Expert Systems with Applications 25 (2003) 343–354

[16] NIZAR R. MABROUKEH and C. I. EZEIFE, ‖A Taxonomy of Sequential Pattern Mining Algorithms‖, ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.