IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 11-14

A Survey on Discovery of Sequential Pattern

Prof. AnkitaBidwaikar
G.H. Raisoni Institute of Information Technology,Nagpur

Abstract: *Sequential pattern mining is one of the essential technique in data mining which is concerned with finding statistically related patterns between the examples of data, where the values are delivered in a sequence. It is presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. The discovery of sequential Pattern is a special case of structured data mining. There are several key traditional computational problems addressed within this field including indexes for sequence information and building efficient databases, extracting the occurring patterns frequently, comparing sequences for similarity, and recovering missing sequence members. The sequence mining problems can be classified as string mining which is typically based on string processing algorithms and itemset mining which is typically based on association rule learning. Local process models extend sequential pattern mining to more complex patterns that can include choices, loops, and concurrency constructs in addition to the sequential ordering construct.*

Keywords: *String Mining, Itemset Mining, Association Rule Learning*

I. Introduction

Sequential pattern mining is one of the essential technique in data mining which is concerned with finding statistically related patterns between the examples of data, where the values are delivered in a sequence[1]. It is presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. The discovery of sequential Pattern is a special case of structured data mining. There are several key traditional computational problems addressed within this field including indexes for sequence information and building efficient databases, extracting the occurring patterns frequently, comparing sequences for similarity, and recovering missing sequence members. The sequence mining problems can be classified as string mining which is typically based on string processing algorithms and itemset mining which is typically based on association rule learning. Local process models[2] extend sequential pattern mining to more complex patterns that can include choices, loops, and concurrency constructs in addition to the sequential ordering construct.

String mining: String mining typically deals with a limited alphabet for items that appear in a sequence, but the sequence itself may be typically very long. Examples of an alphabet can be those in the ASCII character set used in natural language text, nucleotide bases 'A', 'G', 'C' and 'T' in DNA sequences, or amino acids for protein sequences. In biology applications analysis of the arrangement of the alphabet in strings can be used to examine geneand protein sequences to determine their properties. To Know the sequence of letters of a DNA or a protein is not an ultimate goal in itself. The major task is to understand the sequence, in terms of its biological function and Structural Function. This is first acheived by identifying individual regions or structural units within each sequence and then assigning a function to each structural unit. In many cases this requires comparing a given sequence with previously studied ones. When insertions, deletions and mutations occur in a string,the comparison between the strings becomes complex.
A survey of the key algorithms for sequence comparison for bioinformatics is presented by Abouelhoda&Ghanem (2010), which include:[3]

- **Repeat-related problems:** that deal with operations on single sequences and can be based on exact string matching or approximate string matching methods for finding dispersed fixed length and maximal length repeats, finding tandem repeats, and finding unique subsequences and missing which is un-spelled subsequences.
- **Alignment problems:** that deal with comparison between strings by first aligning one or more sequences; examples of popular methods include BLAST for comparing a single sequence with multiple sequences in a database, and ClustalW for multiple alignments. Alignment algorithms can be based on either exact or approximate methods, and can also be classified as global alignments, semi-global alignments and local alignment.

Itemset Mining: Some of the problems in the discovery of sequence lend themselves discovering frequent itemsets and the order they appear, for example, one is seeking rules of the form "if a {customer buys a car}, he /she is likely to {buy insurance} within 1 week", or in the context of stock prices, "if {Nokia up and Ericsson up}, it is likely that {Motorola up and Samsung up} within 2 days". The itemset mining is used in the applications of marketing for discovering regularities between frequently co-occurring items in large transactions. For example, by analysing transactions of customer shopping baskets in a supermarket, one can produce a rule which reads "if a customer buys potatoes and onions together, he/she is likely to also buy hamburger meat in the same transaction".A survey of the key algorithms for item set mining is presented by Han et al. (2007).[4]The two common methods that are applied to sequence databases for frequent itemset mining are the influential apriori algorithm and the FP-growth method.

The applications and use with a great variation of products and the behaviors of user buying, shelf on which products are being displayed is one of the most essential resources in retail environment. Retailers can not only increase their profit but, also decrease cost by proper management of shelf space allocation and the display of the products. To solve this problem, George and Binu (2013) have proposed an approach to user buying patterns using PrefixSpan algorithm and place the products on shelves based on the order of mined purchasing patterns.[5]

II. Methodology

The various methods or techniques and algorithms involved are as follows:

I. GSP Algorithm
II. Sequential Pattern Discovery using Equivalence Classes
III. FreeSpan
IV. PrefixSpan

GSP algorithm which is also known as Generalized Sequential Pattern algorithm is an algorithm used for sequence mining or discovery of sequential Patterns. 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. In a simple manner it 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.
The Algorithm are as follows:

F_1 = the set of frequent 1-sequence
  k=2,
do while F_{k-1} != Null;
    Generate candidate sets C_k (set of candidate k-sequences);
    For all input sequences s in the database D
do
      Increment count of all a in C_k if s supports a
    End do
Fk = {a $\in C_k$ such that its frequency exceeds the threshold}
    k = k+1;
  End do
  Result = Set of all frequent sequences is the union of all F_k's

This Algorithm makes multiple database passes.
The above algorithm looks like the Apriori algorithm. Let us assume that:
A → B and A → C
are two frequent 2-sequences. The items are involved in these sequences are (A, B) and (A,C) respectively. The candidate generation in a usual Apriori style would give (A, B, C) as a 3-itemset, but in the present context it get the following 3-sequences as a result of joining the above 2- sequences
A → B → C, A → C → B and A → BC
The candidate–generation phase takes this into account. The GSP algorithm discovers frequent sequences, allowing for time constraints such as maximum gap and minimum gap among the elements of the sequences. It also supports the notion of a sliding window of a time interval within which items are observed as belonging to the same event, even if they originate from different events.

III. Performance Evaluation

A sequence alignment and a Pairwise Alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.[6] Aligned sequences of nucleotide or amino acid residues are represented as rows within a matrix. The Gaps are inserted between the residues so that similar characters are aligned in successive columns. The Sequence alignments are also used for non-biological sequences, such as calculating the edit distance cost between strings in a natural language or in the financial data.

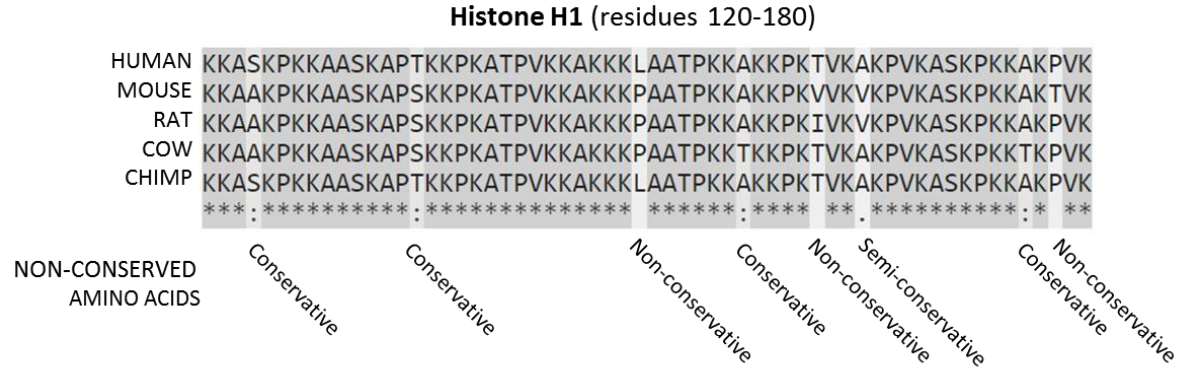

Fig.(1) A sequence and pairwise alignment

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. It also refers to the process of aligning such a set of sequence.

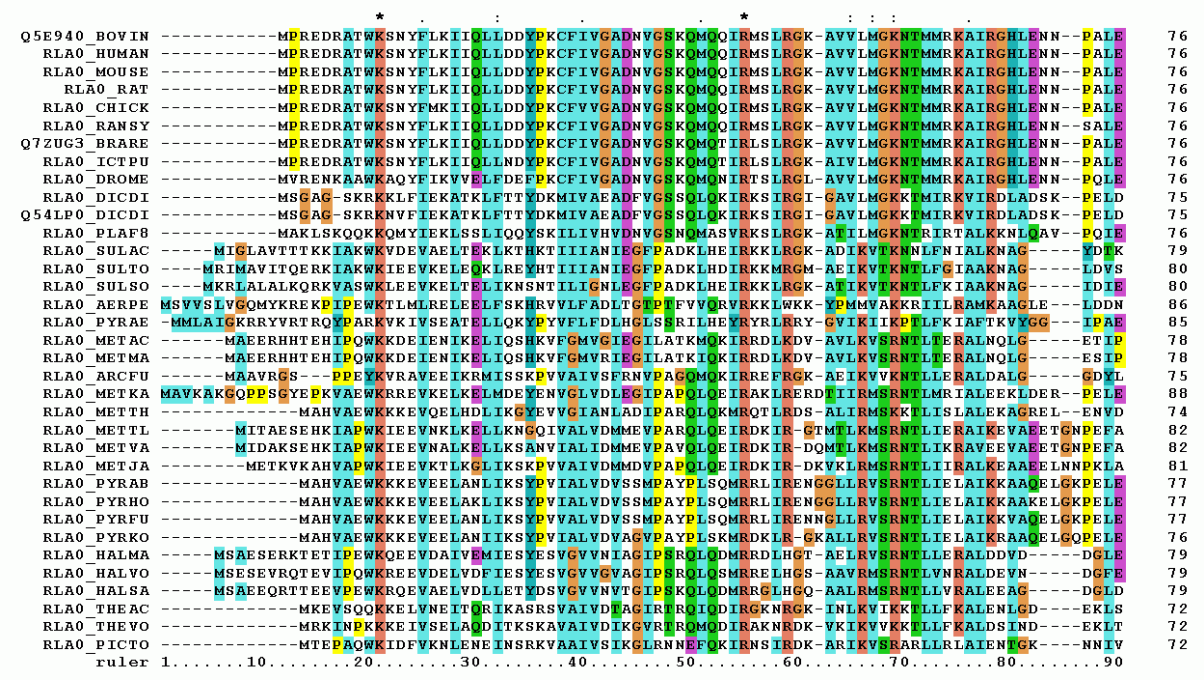

Fig.(2) A multiple sequence alignment

IV. Conclusion

It is concluded that it is very important to do the discovery or mining of sequential pattern by using the different methods, algorithms and some of the useful techniques to get the efficient output.

As we know that the mining plays an important role in discovering the sequential patterns. It depends on the input parameter to get the better results. This techniques which has used are very beneficial to use in a near future to build a new additional features in data mining to dicover more sequential patterns.

References

[1]. Mabroukeh, N. R.; Ezeife, C. I. (2010). "A taxonomy of sequential pattern mining algorithms". ACM Computing Surveys.
[2]. Tax, N.; Sidorova, N.; Haakma, R.; van der Aalst, Wil M. P. (2016). "Mining Local Process Models". Journal of Innovation in Digital Ecosystems.
[3]. Abouelhoda, M.; Ghanem, M. (2010). "String Mining in Bioinformatics". In Gaber, M. M. Scientific Data Mining and Knowledge Discovery. Springer.
[4]. Han, J.; Cheng, H.; Xin, D.; Yan, X. (2007). "Frequent pattern mining: current status and future directions". Data Mining and Knowledge Discovery.
[5]. George, A.; Binu, D. (2013). "An Approach to Products Placement in Supermarkets Using PrefixSpan Algorithm". Journal of King Saud University-Computer and Information Sciences.
[6]. *Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.*