ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 12, December 2013

A Review of Literature on Delta Closed Patterns and Noninduced Patterns from Sequences

Vishal B. Rathod, Prof. A.V.Deorankar, Dr. P.N.Chatur

***Abstract*—Discovering patterns from sequence data has significant impact in many aspects of science and society. In the real world, usually a large set of patterns could be discovered yet many of them are redundant, thus degrading the output quality. To improve the output quality by removing two types of redundant patterns. First, the notion of delta tolerance closed itemset is employed to remove redundant patterns that are not delta closed. Second, the concept of statistically induced patterns is proposed to capture redundant patterns which seem to be statistically significant yet their significance is induced by their strong significant subpatterns. This approach produce a relatively small set of patterns which reveal interesting information in the sequences.**

Keywords-closed frequent itemset, delta closed patterns, statistically induced patterns, suffix tree

I. INTRODUCTION

The sequential pattern mining, which discovers frequent subsequences as patterns in a sequence database, is an important data mining problem with broad applications, including the analyses of customer purchase behavior, Web access patterns, scientific experiments, disease treatments, natural disasters, DNA sequences, and so on.Over the years, many algorithms have been proposed to discover frequent patterns i.e. frequent subsequences. However, most of them overlook the output quality, producing a very large set of patterns, many of which are highly redundant. It has been observed that frequent itemsets (FIs) often contain redundancies. Thus, closed frequent itemsets (CFIs) are proposed as a concise representation of FIs. Although CFIs contain fewer redundancies than FIs. Hence, CFIs are further extended to the delta-tolerance closed itemsets aiming at giving a more concise representation.

Delta closed itemsets provide a controllable tight lossy approximation to the closed itemsets. By allowing tunable tolerance in delta closed itemsets, a great number of redundant itemsets are pruned while important information is retained. Statistical significance is evaluated through statistical hypothesis test which measures how much the frequency of a pattern deviates from the expected one given the random model. It is hoped that patterns occurring with significantly higher frequency will correspond to the functional units inherent in the sequences. Furthermore, the assessment of statistical significance can help in ranking output patterns, enabling experts to assess the result. However, among the statistically significant patterns some are actually statistically redundant. They are considered as significant merely because they contain very strong significant subpatterns. Suffix tree is used to identify the proper superpatterns and subpatterns and hence are able to discover delta closed patterns and noninduced patterns in linear time.

Main objective of Discovery of Delta Closed Patterns and Non induced Patterns from Sequences is that it improves the output quality by removing two types of redundant patterns. First, the notion of delta tolerance closed itemset is employed to remove redundant patterns that are not delta closed. Second, the concept of statistically induced patterns is proposed to capture redundant patterns which seem to be statistically significant yet their significance is induced by their strong significant subpatterns. It is computationally intense to mine these non redundant patterns (delta closed patterns and non induced patterns). To efficiently discover these patterns in very large sequence data, two efficient algorithms have been developed through innovative use of suffix tree.

II. RELATED WORK

A. Finding Sequential Patterns

It is the problem in database mining which is motivated by the decision support problem faced by most large retail organizations. An example of such pattern is customer is going for shopping of computer first then printer and there after camera within period of two months. So the problem is "what items are bought together in a transaction".While related, the problem of finding what items are bought together is concerned with finding intra-transaction patterns, whereas the problem of finding sequential patterns is concerned with inter-transaction patterns. A pattern in the first problem consists of an unordered set of items whereas a pattern in the latter case is an ordered list of sets of items. The problem of finding all sequential patterns is solved in five phases: i) sort phase, ii) litemset phase, iii) transformation phase, iv) sequence phase, and v) maximal phase[3]. There are two algorithm Two of the algorithms, AprioriSome and AprioriAll, have comparable performance, although AprioriSome performs a little better for the lower values of the minimum number of customers that must support a sequential pattern[3].

B. Finding Frequent closed itemset:

It is challenging since one may need to examine a combinatorially explosive number of possible subsequence patterns. Most of the previously developed sequential pattern mining methods follow the methodology of Apriori which may substantially reduce the number of combinations to be examined[2][1]. However Apriori still encounters problems when a sequence database is large and when sequential patterns to be mined are numerous and long. They propose a novel sequential pattern mining method, called Prefixspan (i.e., Prefix-projected-Sequential Pattern mining), which explores prefix projection in sequential pattern mining[2]. Prefixspan mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. Moreover; prefix-projection substantially reduces the size of projected databases and leads to efficient processing. Their performance study shows that Prefixspan outperforms both the Apriori-based GSP algorithm and another recently proposed method; Frees pan, in mining large sequence data bases[2].

C. The PrefixSpan Approach:

Sequential pattern mining is an important data mining problem with broad applications. However, it is also a difficult problem since the mining may have to generate or examine a combinatorially explosive number of intermediate subsequences. Most of the previously developed sequential pattern mining methods, such as GSP, explore a candidate generation-and-test approach to reduce the number of candidates to be examined. However, this approach may not be efficient in mining large sequence databases having numerous patterns and/or long patterns. In this paper, they propose a projection-based, sequential pattern-growth approach for efficient mining of sequential patterns[4]. In this approach, a sequence database is recursively projected into a set of smaller projected databases, and sequential patterns are grown in each projected database by exploring only locally frequent fragments. Based on an initial study of the pattern growth-based sequential pattern mining, FreeSpan, we propose a more efficient method, called PSP, which offers ordered growth and reduced projected databases. To further improve the performance, a pseudoprojection technique is developed in PrefixSpan[4].Furthermore, this mining methodology can be extended to mining sequential patterns with user-specified constraints. The high promise of the pattern-growth approach may lead to its further extension toward efficient mining of other kinds of frequent patterns, such as frequent substructures[4].

D. Ukkonen's Edit Distance Calculating Algorithm:

Edit distance measures the similarity between two strings (as the minimum number of change, insert or delete operations that transform one string to the other). An edit sequence s is a sequence of such operations and can be used to represent the string resulting from applying s to a reference string.They present a modification to Ukkonen's edit distance calculating algorithm based upon representing strings by edit sequences[5]. We conclude with a demonstration of how using this representation can improve mitochondrial DNA query throughput performance in a distributed computing Environment[5].

E. δ -Tolerance Closed Frequent Itemsets:

In this paper, they have given an inherent problem of mining Frequent Itemsets (FIs): the number of FIs mined is often too large. The large number of FIs not only affects the mining performance, but also severely thwarts the application of FI mining. However, the number of CFIs is still too large in many cases, while MFIs lose information about the frequency of the FIs. To address this problem, relax the restrictive definition of CFIs and propose the δ-Tolerance CFIs (δ- TCFIs). Mining δ-TCFIs recursively removes all subsets of a δ-TCFI that fall within a frequency distance bounded by δ[2]. They propose two algorithms, CFI2TCFI and MineTCFI, to mine δ-TCFIs. CFI2TCFI achieves very high accuracy on the estimated frequency of the recovered FIs but is less efficient when the number of CFIs is large, since it is based on CFI mining. MineTCFI is significantly faster and consumes less memory than the algorithms of the state-of-the-art concise representations of FIs, while the accuracy of MineTCFI is only slightly lower than that of CFI2TCFI[2].

F. Delta Closed Patterns and Noninduced Patterns from Sequences:

Discovering patterns from sequence data has significant impact in many aspects of science and society, especially in genomics and proteomics. Here they consider multiple strings as input sequence data and substrings as patterns. In the real world, usually a large set of patterns could be discovered yet many of them are redundant, thus degrading the output quality[6]. Their paper improves the output quality by removing two types of redundant patterns. First, the notion of delta tolerance closed itemset is employed to remove redundant patterns that are not delta closed. Second, the concept of statistically induced patterns is proposed to capture redundant patterns which seem to be statistically significant yet their significance is induced by their strong significant subpatterns[6]. It is computationally intense to mine these nonredundant patterns (delta closed patterns and noninduced patterns). To efficiently discover these patterns in very large sequence data, two efficient algorithms have been developed through innovative use of suffix tree. Three sets of experiments were conducted to evaluate their performance.[6]

G. Synthesis and Recognition Of Sequences:

A string or sequence is a linear array of symbols that come from an alphabet. Due to unknown substitutions, insertions,and deletions of symbols, a sequence cannot be treated like a vector or tuple of a fixed number of variables. The synthesis of an ensemble of sequences is a "sequence" of random elements that specify the probabilities of occurrence of the diff. Symbols at corresponding sites of sequences[7]. The synthesis is determined by a hierchical sequence synthesis procedure (HSSP), which returns not only taxonomic hierchy of the whole ensemble of sequences but also alignment and synthesis of group of sequences at each level of hierchy[7]. The HSSP does not require the ensemble of sequences to be presented in the form of tabulated array of data, hierchical information of data or assumption of a stochastic process. This correspondence present the concepts of sequence synthesis & the applicability of HSSP as a supervised classification procedure as well as an unsupervised classification procedure[7].

III. CONCLUSION AND FUTURE SCOPE

In this paper we review the existing techniques of pattern mining . We discussed a variety of pattern mining method techniques such as PrefixSpan Approach , Delta Closed Patterns and Noninduced Patterns from Sequences. For each technique we have provided a detailed explanation of the techniques which are used for finding the patterns.It is observed that for finding frequent pattern there is problem of fake pattern & redundant. So to solve this problem there is Delta Closed Patterns and Noninduced Patterns algo which gives good output in minimum time. They produce a relatively small set of patterns which reveal interesting information in the sequences.

Mining of patterns using Discovery of Delta Closed Patterns and Noninduced Patterns using suffix tree gives patterns and proposes the notion of statistically induced patterns to capture redundant patterns. Here efficient algorithms for discovering delta closed patterns and noninduced patterns from large sequence data given. Two algorithms that use a generalized suffix tree in an innovative manner, assisting the identification of these patterns effectively in linear time.The proposed approach is very useful to give interesting patterns at the end.

REFERENCES

[1] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," *Proc. Seventh Int'l Conf. Database Theory, pp. 398-416, 1999.*

[2] J. Cheng, Y. Ke, and W. Ng, "δ -Tolerance Closed Frequent Itemsets," *Proc. Sixth Int'l Conf. Data Mining, pp. 139-148, 2006.*

[3] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. 11th Int'l Conf. Data Eng., pp. 3-14, 1995.*

[4] J. Pei and J. Han, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," *Proc. 17th Int'l Conf. Data Eng., pp. 215-224, 2001.*

[5] Bruce Johnson "A Bioinformatics-Inspired Adaptation to Ukkonen's Edit Distance Calculating Algorithm and Its Applicability Towards Distributed Data Mining," *2008 International Conference on Computer Science and Software Engineering*

[6] Andrew K.C. Wong, Fellow, IEEE, Dennis Zhuang, Gary C.L. Li, Member, IEEE, and En-Shiun Annie Lee Discovery of "Delta Closed Patterns and Noninduced Patterns from Sequences," *IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 8, August 2012*

[7] S.C. Chan and A.K.C Wong, "Synthesis and Recognition of Sequences," *IEEE Trans. Pattern Analysis Machine Intelligence, vol. 13, no. 12, pp. 1245-1255, Dec. 1991.*

AUTHOR BIOGRAPHY

Vishal B. Rathod completed his B.Tech. in CSE from Government college of Engineering Amravati, India in 2011 and now he is pursuing M.Tech in CSE branch from Government college of Engineering Amravati . His area of research includes data mining, network security, pattern recognition, and neural networks.

Prof. A. V. Deorankar has received his M.E.degree in Electronics Engineering from Govt. College of Engineering, Amravati, India. He has published nine national level papers and seven international papers. He is also patent of one paper. His area of research includes Computer Network, Web Mining. Currently he is working as an Associate Professor at Govt. college of Engineering, Amravati, India.

Dr. P. N. Chatur has received his M.E. degree in Electronics Engineering from Govt. College of Engineering, Amravati, India and PhD degree from Amravati University. He has published twenty national level papers and fifteen international papers. His area of research includes Neural Network, data mining. Currently he is head of Computer Science and Engineering department at Govt. College of Engineering, Amravati.