An Effective Approach to Mine Frequent Sequential Pattern over Uncertain dataset

Kshiti S Rana[#1], Hiren V Mer[*2]

[#]Research scholar, Parul Institue of Technology,Limda, Waghodia, Vadodara, India
[*]Asst. Prof. Parul Institute of Tehnology, Limda, Waghodia, Vadodara, India

***Abstract*— In recent years, due to the wide applications of uncertain data, mining frequent itemsets over uncertain databases has attracted much attention. In uncertain databases, the support of an itemset is a random variable instead of a fixed occurrence counting of this itemset. There are several application in which the data mining on uncertain data is useful like sensor network monitoring, moving object search etc. In this paper we are focusing on mining frequent sequential pattern using SeqU-PrefixSpan algorithm. Using this algorithm we can find the frequent sequential pattern from an uncertain database. We are proposing one incremental approach for this algorithm, which will also find the patterns and reduce the time of execution by dividing the data into parts. Data input will be in parts which are done randomly. Due to this the execution time will be less as compared to existing algorithm.**

***Keywords*— Sequential pattern mining, uncertain data**

I. Introduction

Before going on the topic let's first understand what is data mining and its aspects. Data mining is the practice of examining large pre-existing databases in order to generate new information. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It includes task like classification of data, make clusters of data, data preprocessing etc. it is very useful for finding any kind of information between dozens of data. For example, in any shopping mall if we want find how much pack of soap is being purchased in any month we can find is using any data mining software like weka, orange etc.

Now next is mining frequent pattern, it is probably most important and useful concept of data mining. The pattern which occurs maximum time in the transaction is called frequent pattern. These days researchers are focusing more on sequential patterns. Sequential Pattern 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. [4]

Sequential pattern mining is most widely searched topic in current time. It is useful in many area like finding DNA sequences, in stock market, find the location of the moving object, genetically structure of any living object, weather forecasting etc. For example, if we want to know the list of items which are being purchased together we can monitor baskets of all customers and find out like customer who buys jeans buys a top and matching accessories like belt, shoes within 1 2 days. Using sequential pattern mining algorithm we can mine this kind of useful patterns in very effective manner. That's why it is widely used in real life applications.

Uncertain data is the type of data which posses some amount of uncertainty. In recent years, many advance technologies have been developed to store and record large quantities f data in continues manner. This generates the problem of uncertain data. Uncertain data can't not be representing like simple data we need probability distribution to represent each and every value of data. In current scenario many data has some amount of uncertainty present in it. We can say that Sensor network is the best example of uncertain dataset. Consider the case of wireless sensor network (WSN) where each sensor records continues reading for weather temperature and humidity within certain detection range of it. In such a case, the reading can be inherently noisy due to temperature effect, and can be can be associated with a confidence value determined by, for example, the stability of the sensor.

Sequential pattern mining is one of the important tasks in data mining area. Moving object can be found by tracking its sequence of presence in all the location step by step. A trajectory of a moving object consists of time-stamped location data across a sequence of ordered timestamps. [6] This type of data is said to be the uncertain data. There is certain amount of noise is present in uncertain data. Various factors are responsible for data uncertainty; it includes incompleteness of data sources, the addition of artificial noise in privacy-sensitive applications and, most importantly, uncertainty arising from imprecision in measurements and observations.

II. Related Work

Recently researchers are focusing more on mining frequent sequential pattern from uncertain database. Here we provide some related work for it.

The algorithm which is solves the problem of mining sequential pattern are PrefixSpan, SPADE, GSP, Freespan.

PrefixSpan

PrefixSpan stands for **Prefix**-*projected* **S***equential* **pa***tter***n [1]** *mining* which searches for prefix projection in the sequential database. This algorithm mines the whole databse and reduces the effort of candidate subsequence generation. It is projection based approach and it shrinks the database after scanning the database.

Fig 1 shows the step by step execution of PrefixSpan This algorithm scans whole database. This is called level-1 sequential patterns. Once the first level patterns are found step is to find level-2 sequential patterns which are patterns with length 2. Respectively whole database is to list of pattern. First and main advantage is that there is to generate candidate sequence. And it reduces the size of as it is projection based approach. The major disadvantage algorithm is it requires major cost of constructing the database.

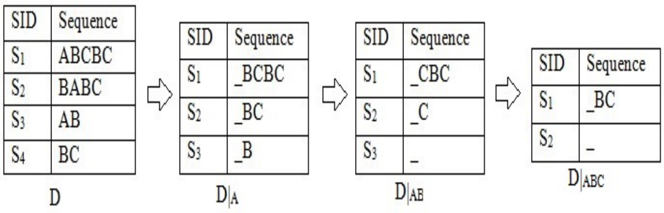

Fig 1: Step by step execution of PrefixSpan

SPADE

SPADE stands for **S**equential **PA**ttern **D**iscovery using **E**quivalence classes [2]. This algorithm uses vertical format sequential pattern mining method. Whole databse is mapped into huge set of (SID, EID). Lattice-theoretic approach can be used to crumble the original search space (lattice) into smaller pieces (sub-lattices) which can be processed independently in main-memory. It is used to effectively reduce number of databse scans and minimize cost of execution. Here searching is done by id-list. It is a three pass scanning process. Time required converting horizontal database to vertical database and storage space too large than the original sequence database.[2]

There are many other algorithms also to find frequent sequential pattern from uncertain data like GSP [2], Freespan [4]. But the PrefixSpan algorithm is most widely used algorithm.

To deal with mining uncertain data we have discussed p-FSE and SeqU-prefix span algorithm. In [5] probabilistic frequent serial episodes are mined from uncertain sequence data which relates to many real-world applications like sensor networks as well as customer purchase sequence. There are mainly three approaches to mine frequent sequential pattern from uncertain data. They are:

1. An **exact approach** which calculates accurate frequentness probabilities of episodes.[5]
2. An **approximate approach** which approximates the frequency of episode using probability models.[5]
3. An **optimized approach** which efficiently prunes candidate episodes by estimating an upper bound of its frequentness probability using approximation techniques.[5]

Next algorithm which is based on uncertain data is discussed in [6]. In [6] authors have focus on mining frequent pattern on spatio temporal object database with gap constraints. This kind of pattern is useful in locating any moving object in space. Breadth first search and depth first search is explored for frequent pattern mining. [6] The main challenge in mining moving object is extraction of objects co-occurring at a minimum number of time stamps. Objects are said to be close if they occur in same timestamp. We consider objects are together is they belongs to same cluster based on its closeness measure.

Next algorithm is combine approach of frequent pattern mining over uncertain data. To deal with uncertain data [7], the U-Apriori algorithm[15] was proposed in PAKDD 2007. Similar to Apriori algorithm U-Apriori scans database multiple times. To reduce the number of scans new approach is used called UF-growth algorithm [16] was proposed in PAKDD 2008.

III. EXISTING WORK

The existing work is Sequence level U-PrefixSpan which is modified version of the original PrefixSpan which finds sequential pattern from uncertain database. The earlier work uses expected support which measures frequentness of pattern and it is unable to mine high quality sequential pattern. U-PrefixSpan overcomes the challenges of p-FSP algorithm which is to confirm data to the sequence level U-PrefixSpan. In the existing algorithm they use threshold support and threshold probability count to mine the frequent pattern. The main problem with existing work is that the take whole dataset as input so cost of computation is increased. so to overcome this problem we proposes one Extended SeqU-PrefixSpan algorithm.

In the existing algorithm of SeqU-PrefixSpan. Here there are mainly 3 algorithm running first is of finding *PMFcheck* which is compare the element and heck the frequency of the element. The second is of *Prune* which is used to prune the infrequent element and the the last is combination of both which will at the end generate frequent sequential patterns. At the end of the algorithm we will get frequent sequences present in the database.

IV. PROPOSED WORK

The above existing algorithm works well with the dataset but when the size of dataset increased the computation cost is increased due to large number of data. So to overcome this problem we propose incremental approach in which the dataset is scan in the parts and one list of frequent item is preserved along with its probability count. When the new element is it is appended to the frequent item list and if the element is already presents its probability count is increased. Here we are not focusing on the accuracy of the algorithm we are currently focusing on the time complexity of the algorithm. We will supply the data as a input in passes so that overall time can be reduced. There will be one list which will maintain the list of all the frequent items and at the end all the frequent sequences are generated based on it probability count.

V. CONCLUSION

In this paper we have discussed what is sequential pattern mining as well as uncertain data mining. In current time uncertainty is very common in all kind of databases. To address the problem of uncertainty we have discussed algorithm related to it. We have focused on the algorithm which mines sequential pattern from uncertain database. Previous work uses support count as a basis to solve the problem. PrefixSpan is most widely used algorithm to solve the problem. We have discussed multiple algorithms like sequence level U-PrefixSpan, p-FSE, Element level U-PrefixSpan etc. all this algorithm is very useful to mine sequential pattern from uncertain database. And at the end we have discussed our proposed work to reduce the execution time of the existing algorithm.

REFERENCES

[1] Pei, Jian, et al. "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth." *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, 2001.

[2] Zaki, Mohammed J. "SPADE: An efficient algorithm for mining frequent sequences." *Machine learning* 42.1-2 (2001): 31-60.

[3] Han, Jiawei, et al. "FreeSpan: frequent pattern-projected sequential pattern mining." *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.

[4] Motegaonkar, Vishal S., and Madhav V. Vaidya. "A Survey on Sequential Pattern Mining Algorithms." International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (2) , 2014, 2486-2492

[5] Wan, Li, Ling Chen, and Chengqi Zhang. "Mining frequent serial episodes over uncertain sequence data." *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013.

[6] Li, Yuxuan, et al. "Mining Probabilistic Frequent Spatio-Temporal Sequential Patterns with Gap Constraints from Uncertain Databases." *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013.

[7] Leung, Carson Kai-Sang, and Syed Khairuzzaman Tanbeer. "PUF-tree: a compact tree structure for frequent pattern mining of uncertain data." *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013. 13-25.

[8] Aggarwal, Charu C., and Philip S. Yu. "A survey of uncertain data algorithms and applications." *Knowledge and Data Engineering, IEEE Transactions on*21.5 (2009): 609-623.

[9] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data."*Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[10] Tong, Yongxin, et al. "Mining frequent itemsets over uncertain databases. "*Proceedings of the VLDB Endowment* 5.11 (2012): 1650-1661.

[11] Zhao, Zhou, Da Yan, and Wilfred Ng. "Mining probabilistically frequent sequential patterns in uncertain databases." *Proceedings of the 15th international conference on extending database technology*. ACM, 2012.

[12] Zhao, Zhou, Da Yan, and Wilfred Ng. "Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases." (2013): 1-1.

[13] Muzammal, Muhammad, and Rajeev Raman. "Mining sequential patterns from probabilistic databases." *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2011. 210-221.

[14] M. Chau, R. Cheng, B. Kao, and J. Ng, “Uncertain data mining: An example in clustering location data,” in PAKDD. Springer, 2006, pp. 199–204.

[15] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data."*Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.