Mining Frequent Sequential Patterns and Top Rules from Large Uncertain Database

Jayshri Banpurkar[1], Amreen Khan[2]

[1]Research Scholar, Dept. of Computer Science, GHRCE, Maharashtra, India
[2] Assistant Professor, Dept. of Computer Science, Maharashtra, India

Abstract - *Data mining play very important role in mining useful knowledge from Wireless Sensor Network. In Today's real world environment, most of the databases such as weather database, RFID, stock market analysis are uncertain as it may contain random and noisy values because of some faulty nodes. The uncertainty in database can be handle with the help of U-PrefixSpan algorithm. This algorithm is applied on the database for finding out the frequent sequential patterns by providing the minimum support. As Database is very large, the number of frequent sequential patterns is also large which is difficult to study. Hence there is need to apply the Top K Rule Mining algorithm on the generated frequent sequential patterns to obtain the top K rules. According to the generated rules the prediction can be done. The result and analysis show that the U-PrefixSpan takes more time for execution than Top K rule mining algorithm.*

Key Words: Wireless Sensor Network, U-PrefixSpan, Top K Rule mining.

1. INTRODUCTION

Data Mining is the process of mining the useful knowledge from the database. The Data mining concepts useful for uncovering the interesting patterns hidden in large data sets. Nowadays data uncertainty is inherent in many real time application like wireless sensor network, RFID, Scientific experiments and web access pattern. Uncertain databases contains random values which are subtly different from each other. Consider the wireless sensor network in the lab where 54 sensors are deployed to collect the reading. The sensors nodes collect the reading having attributes such as date, time, temperature, light, humidity, voltage. The problem of finding the frequent sequential patterns from uncertain databases has attracted a lot of attention in the research area. For example, in biological research, FSP mining is helpful in determining the association among gene sequence and also useful in the grouping of moving objects.

Consider the sequence level uncertainty where the expected support is used to measure the frequentness of patterns such as frequent item sets [2] and frequent subsequences on uncertain data. The sequence level uncertain model is fundamental for a lot of real-life application of expected support unable to reflect pattern frequentness. An RFID location tracking system taking as example, where a set of RF reader are organized in an indoor Environment, and a user may be sensed by several near-by readers. In such application, user locations are uncertain. In Uncertain databases, the expected support unable to identify frequent patterns. In fact expected support may give rise to many infrequent patterns. Therefore, the evaluation of frequentness of sequential pattern observing to probability theory which gives idea of probabilistic frequentness.

In this paper, the frequent sequential patterns is generated by U-PrefixSpan [5] algorithm. The output of U-PrefixSpan is given to TopKRules algorithm which generate rules showing association among different attributes. There are four stages required for finding frequent sequential patterns such as conversion of text dataset into CSV format, pre-processing the dataset, creating the batches of ten thousand each and then finally applying the U-PrefixSpan algorithm. The Rules showing association among them is obtained using TopKRules algorithm.

The remaining part of this paper are organized as follows. Section II describes the literature review, the working of system model and proposed plan are discussed in section III, results and analysis are shown in section IV and paper is concluded in section V.

2. LITERATURE REVIEW

In [1], the author proposed the concept of applying U-Prefix Span for mining probabilistically frequent sequential patterns in large uncertain databases in which two uncertainty models are used such as sequence-level and element-level uncertainty model to handle the data uncertainty in real world application. The drawback of this system is to use expected support as the measurement of frequentness of pattern. It effectively avoids the problem of possible world explosion this is advantage of this system.

In [2], the author proposed the concept of mining of frequent itemsets on large uncertain databases, since an uncertain database consist an exponential number of possible worlds. To overcome this problem the incremental mining algorithm is used and poisson binomial distribution model is applied for mining. As it the frequent itemset mining, the tuple and attribute uncertainty models are used. Incremental mining algorithm enables the results to be revitalized by reducing the need of re-executing the mining algorithm on novel database which is more expensive and unnecessary.

Author proposed the system in [3], to reduce the search space and avoid redundant computation by applying efficient mining algorithm based on depth-first search method to obtain all probabilistic frequent closed itemsets.

Author proposed the system in [4], where sequential pattern mining is one of the challenging task since explosive number of possible subsequence patterns are generated. The Apriori reduce the number of combination but rises the problem when sequence database is large. To overcome this problem PrefixSpan algorithm is proposed which mines complete set of patterns but reduces the efforts of candidate sequence generation which leads to efficient processing.

In [5], author discusses about the concept of dynamic programming algorithm which embedded into candidate-generate and test approach and explore the pattern into breadth-First search and depth first search manner. In this system, the elimination of candidate sequences without calculating their support through probabilistic pruning.

In [6], author discusses about mining top k association rules which chooses the top k rules only and has advantage that the user can control the number of rules. The drawback of this system is that it generate so many results or too few results eliminating valuable data.

3. PROPOSED PLAN

Basically proposed work consist of mainly 4 stages:

1) Weather Dataset taken as input

2) Data Pre-processing.

3) FSPs generation using U- PrefixSpan algorithm

4) Applying the top k association Rule mining algorithm.

1) Weather Dataset taken as input:

The collection of dataset is very essential step in the process of data mining because the data mining techniques are applied over the databases which is chosen. The dataset may be of different type such as wireless sensor network data, e-commerce data, web log etc. The data collected from Intel Research Lab where 54 sensors are deployed. The dataset having the attribute like date, time, epoch, mote id, temperature, humidity, light, voltage. The dataset consist of 2.3 million entries where epoch is increasing series number from each mote. At the same time two readings from the same epoch number were produced. There are few missing epoch in the dataset. The range of mote changes from 1 to 54; there may be the chance that some data is missing. Temperature is measured in degrees Celsius. Humidity ranges from 0-100%. Light is measured in the Lux. Voltage is expressed in volts, ranging from 2-3.

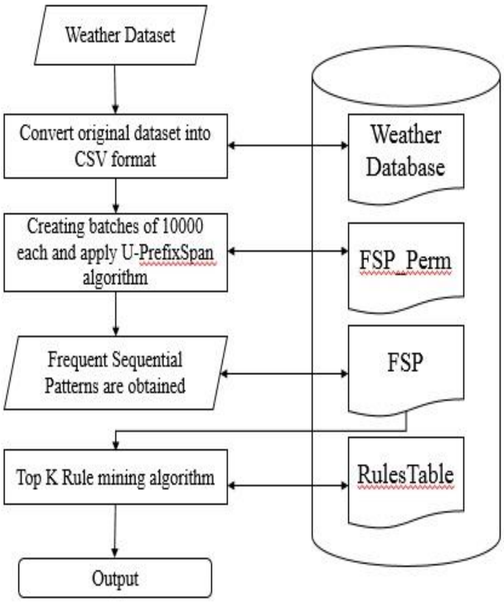

Weather Dataset
Convert original dataset into CSV format
Creating batches of 10000 each and apply U-PrefixSpan algorithm
Frequent Sequential Patterns are obtained
Top K Rule mining algorithm
Output
Weather Database
FSP_Perm
FSP
RulesTable

Fig. 1 Working of System Model

2) Data Pre-processing:

In Today's real-world databases are highly affected by noise. Missing and inconsistent data obtained due to their typically huge size and in fact data is collected from heterogeneous sources. It may happen that Low-quality data will lead to improper mining results hence there is need to apply the pre-processing techniques. There are many pre-processing techniques such as data cleaning, data transformation as well as integration of data etc. Among these pre-processing techniques, Data cleaning is most suitable for weather dataset to eliminate noise and correct irregularities in the data. Data cleaning helps to clean the records by filling in missing values.

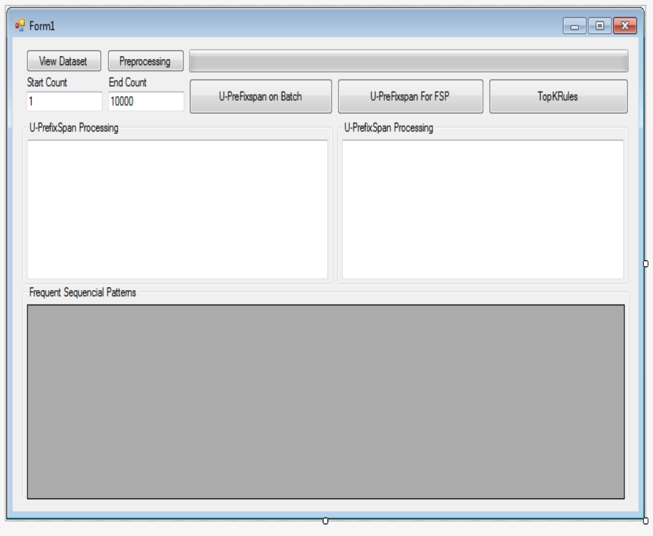
3) Mining frequent sequential pattern using U-Prefixspan algorithm:

The mining of frequent sequential pattern (FSPs) from probabilistic databases using U-PrefixSpan algorithm possess the model called Seq-UPrefixSpan which effectively solves the problem of possible world explosion. The various pattern mining algorithm like GSP, SPADE, FreeSpan and PrefixSpan having interestingness measure as expected support that unable to find some probabilistically frequent pattern. Expected support may give rise to some probabilistically infrequent patterns as a result. The U-PrefixSpan algorithm uses the idea of probabilistic frequentness which is able to capture the complex relationship between uncertain sequences.

4) Mining Rules using TopKRules algorithm:

The mining of frequent sequential patterns is not enough as large number of patterns are obtained because large uncertain database is taken as input. It is essential to obtain the top K rules by applying Top K association Rule mining algorithm. Since frequent sequential patterns consumes more memory. TopKRules algorithm added after U-PrefixSpan gives precise result based on which prediction can be done.

4. RESULT AND ANALYSIS



Start Count: 1
End Count: 10000

Fig. 2 GUI showing main window

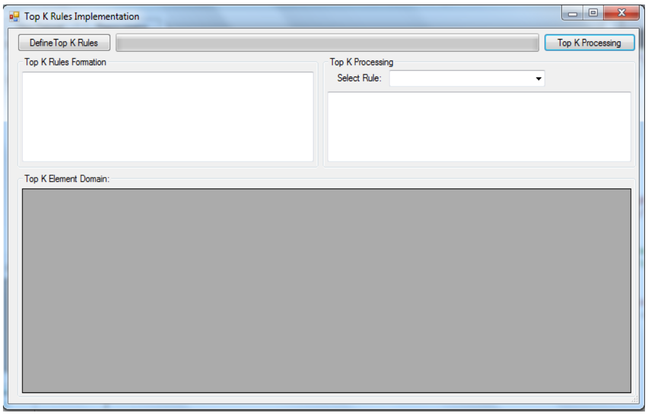
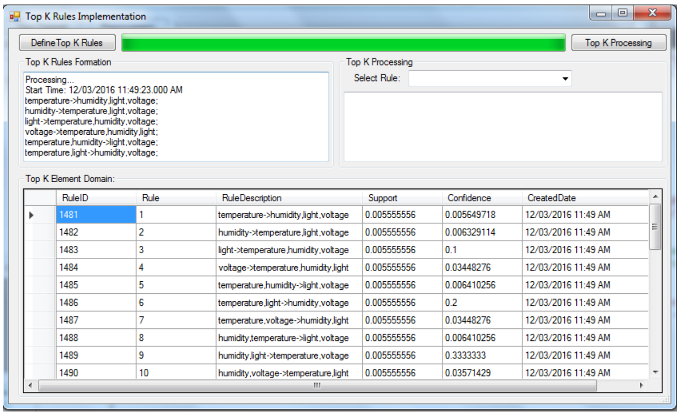


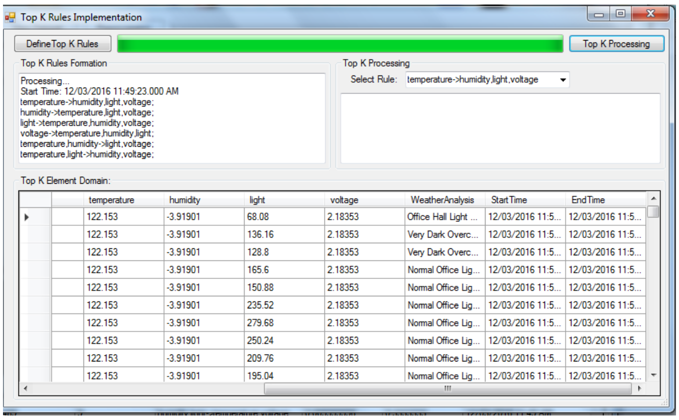
Fig. 3 GUI obtained after Top K Rule Mining Button



Processing...
Start Time: 12/03/2016 11:49:23.000 AM
temperature->humidity,light,voltage;
humidity->temperature,light,voltage;
light->temperature,humidity,voltage;
voltage->temperature,humidity,light;
temperature,humidity->light,voltage;
temperature,light->humidity,voltage;

RuleID	Rule	RuleDescription	Support	Confidence	CreatedDate
1481	1	temperature->humidity,light,voltage	0.005555556	0.005649718	12/03/2016 11:49 AM
1482	2	humidity->temperature,light,voltage	0.005555556	0.006329114	12/03/2016 11:49 AM
1483	3	light->temperature,humidity,voltage	0.005555556	0.1	12/03/2016 11:49 AM
1484	4	voltage->temperature,humidity,light	0.005555556	0.03448276	12/03/2016 11:49 AM
1485	5	temperature,humidity->light,voltage	0.005555556	0.006410256	12/03/2016 11:49 AM
1486	6	temperature,light->humidity,voltage	0.005555556	0.2	12/03/2016 11:49 AM
1487	7	temperature,voltage->humidity,light	0.005555556	0.03448276	12/03/2016 11:49 AM
1488	8	humidity,temperature->light,voltage	0.005555556	0.006410256	12/03/2016 11:49 AM
1489	9	humidity,light->temperature,voltage	0.005555556	0.3333333	12/03/2016 11:49 AM
1490	10	humidity,voltage->temperature,light	0.005555556	0.03571429	12/03/2016 11:49 AM

Fig. 4 GUI obtained after Clicking Define Top K Rules button



Processing...
Start Time: 12/03/2016 11:49:23.000 AM
temperature->humidity,light,voltage;
humidity->temperature,light,voltage;
light->temperature,humidity,voltage;
voltage->temperature,humidity,light;
temperature,humidity->light,voltage;
temperature,light->humidity,voltage;

Select Rule: temperature->humidity,light,voltage

	temperature	humidity	light	voltage	WeatherAnalysis	StartTime	EndTime
	122.153	-3.91901	68.08	2.18353	Office Hall Light ...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	136.16	2.18353	Very Dark Overc...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	128.8	2.18353	Very Dark Overc...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	165.6	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	150.88	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	235.52	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	279.68	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	250.24	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	209.76	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...
	122.153	-3.91901	195.04	2.18353	Normal Office Lig...	12/03/2016 11:5...	12/03/2016 11:5...

Fig. 5 GUI obtained after clicking Top K Processing button

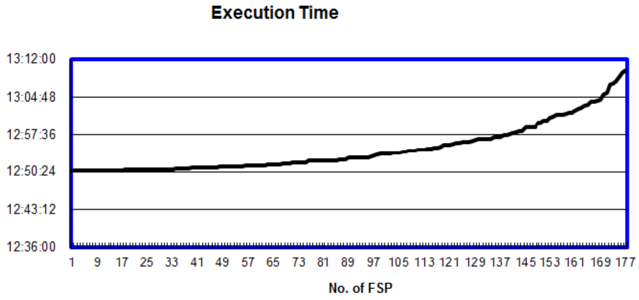


Fig. 6 Graph Showing execution time required for FSP generation

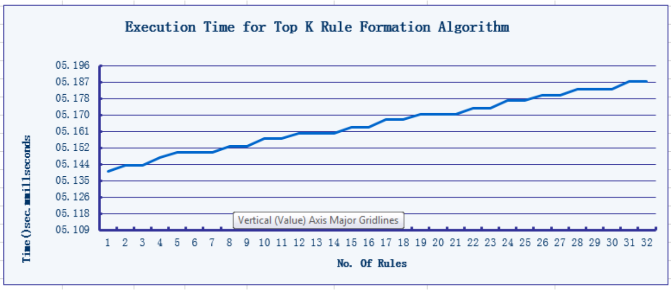


Fig. 7 Graph displaying execution time (in milliseconds) for generation of Top K Rules.

5. CONCLUSION

The research work include combining the two different algorithms to improve the efficiency of system in terms of time and iteration. The U-PrefixSpan algorithm applied on large uncertain database to obtain frequent sequential patterns. Since database is large uncertain, the large number of patterns are generated which is difficult to study hence there is need of top k rule mining algorithm for obtaining precise results. TopKRules algorithm generate 32 rules based upon that the prediction of weather can be done which help to know the climate based on the ranges of attributes.

REFERENCES

[1] Zhou Zhao, Da Yan and Wilfred Ng, "Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases, IEEE transactions on knowledge and data engineering," Vol. 26, 2014 pp. 1171-1184.

[2] L. Wang, D. Cheung, R. Cheng, S. Lee, and X. Yang, "Efficient Mining of Frequent Itemsets on Large Uncertain Databases," IEEE Transaction on Knowledge and Data Engineering, Vol. 24, No. 12, pp. 2170-2183, Dec. 2012.

[3] Y. Tong, L. Chen and B. Ding, "Discovering Threshold-based Frequent Closed Itemsets over Probabilistic Data," 2012 IEEE 28th International Conference on Data Engineering, Washington, DC, 2012, pp. 270-281..

[4] Jian Pei et al., "Mining sequential patterns by pattern-growth: the PrefixSpan approach," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1424-1440, Nov. 2004 .

[5] M. Muzammal and R. Raman, "Mining Sequential patterns from probabilistic database," 15th Pacific-Asia Conference on Knowledge Discovery, 2011.

[6] Bi-Ru Dai, Hung-Lin Jiang and Chih-Heng Chung, "Mining Top-K Sequential Patterns in the Data Stream Environment," International Conference on Technologies and Applications of Artificial Intelligence, 2010, pp. 142-149.

[7] K. Rana et al, "An Effective Approach to Mine Frequent Sequential Pattern over Uncertain dataset," International Journal of Computer Science and Information Technologies, Vol. 6 , 2015 pp. 3242-3244.

[8] C. C. Aggarwal and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 609-623, May 2009.

[9] C. K. S. Leung, C. L. Carmichael and B. Hao, "Efficient Mining of Frequent Patterns from Uncertain Data," Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE, 2007, pp. 489-494.

[10] Jayshri Banpurkar, Amreen Khan, "Review on Mining Sequential Patterns from Large Uncertain Databases", IEEE 3rd International Conference On Electronics & Communication Systems (ICECS- 2016) , 25-26 February 2016, Coimbatore, India.