Probabilistic Frequent Sequential Patterns Analysis Using Apriori of Unsure Databases

Madhavi G. Patil[1], Ravi Patki[2]

[1] PG Student, Dr. D. Y. Patil College of Engineering, Pune, Maharashtra, India

[2] Project Guide, Dr. D. Y. Patil College of Engineering, Pune, Maharashtra, India

Abstract: ***Frequent item-set mining in uncertain transaction databases semantically and computationally differs from traditional techniques applied on standard (certain) transaction databases. Data uncertainty is inherent in number of real-world applications. Uncertain transaction databases consist of sets of existentially uncertain items. The uncertainty of items in transactions makes traditional techniques inapplicable. Mining serial patterns from inaccurate data, like those data arising from detector readings is incredibly necessary for locating hidden information in such applications. In applications such as natural habitat monitoring, web data integration, the values of the underlying data are inherently deafening or vague. PrefixSpan tend to propose to reside pattern frequentness supported the achievable globe linguistics. It tend to establish two unsure sequence information models abstracted from a number of real-life applications involving uncertain sequence information, and plan the subject of removal probabilistically frequent sequential patterns from information that adapt to developed models.***

Keywords: frequent item-set, serial patterns, uncertain databases, prefixspan algorithm

1. Introduction

Data mining examinations the data gathered from diverse sources and gather valuable data from it. It discovers relationships or designs among handfuls of fields in substantial social databases. Consider a wireless sensor network (WSN) system, where each sensor continuously collects readings of environmental parameters, such as temperature and humidity, within its detection range. In such a case, the readings are inherently noisy, and can be associated with a confidence value determined by, for example, the stability of the sensor.

In data mining Sequential design mining will be one of the critical errands. A direction of a moving item comprises of time-stamped area data crosswise over an arrangement of requested timestamps. This sort of data will be ordered as indeterminate dataset. Questionable dataset have notable measure of clamor present. Different elements add to data vulnerability, including deficiency of data sources, the expansion of manufactured commotion in protection touchy applications and, in particular, instability emerging from imprecision in estimations and perceptions.

Sequential design will be basically talked point in current time. Applications of consecutive design mining will be Medical medicines, common calamities, science and designing methodologies, securities exchanges, DNA arrangements, and quality structures. For illustration, Customer shopping arrangements: First purchase PC, then CD-ROM, and after that advanced cam, inside 3 months. Consecutive design mining calculations give this sort of valuable designs in exceptionally compelling way. So it is generally acknowledged truth be told application. In order to satisfy the increasing needs of the above applications, this envision that novel, correct, and scalable methods for managing uncertain data need to be developed.

In order to satisfy the increasing needs of the above applications, this paper envision that novel, correct, and scalable methods for managing uncertain data need to be developed. To achieve this goal, this paper is leading for following steps:

Step 1: Develop a practical database system that incorporates uncertain data as a first-class citizen, in order to facilitate the development of the above applications; and

Step 2: Investigate the issues of data uncertainty in data mining, ambiguity removal, and data integration.

2. Related Work

PrefixSpan (**Prefix** projected **S**equential **pa**tter**n)** mines entire record of sequences and shrink the applicant subsequence creation try. In addition, prefix-projection significantly reduces the amount of probable databases and leads to efficient processing. It is projection based approach and it shrinks the database after scanning the database. This technique memory can be saved successfully. PrefixSpan cover numerous advantages. First and most important advantage is that there is no require generating applicant string. And it reduces the size of database by projection based approach. SPADE (**S**equential **PA**ttern **D**iscovery using **E**quivalence) is used to reduce I/O by reducing database scan as well as lessen cost of totaling by using well-organized search method. In SPADE, searching is done by id-list transactions. The whole process is three passes of database scanning.

The next algorithm is GSP, which stands for Generalized Sequential Pattern mining algorithm. It is based on Apriori algorithm. GSP scans database multiple times. In the very first scanning all the items occurring in the database is counted and scheduled from the succession applicant 2-sequence is generated. Now in next step support count of this candidate 2-sequence is counted. This candidate 2-sequence will be the source for next candidate 3-sequence. This process is continual until no further recurrent sequence is found. There mainly two major steps of the algorithm:

1. Candidate generation- which will generate the candidate sequence and perform join operation to perform next pass.
2. Support Counting. Normally, a hash tree–based search is employed for efficient support counting.

Freespan (**Fre**quent pattern-projected Sequential **Patter n)** reduces the applicant making cost. It uses the recurrent items to iteratively plan the sequence database in predictable database while rising subsequence's repeatedly. Each projection partitions the database and restricts added testing to smaller units.

3. Proposed System

A. Problem Definition

Expected support as the measurement of pattern frequentness, which has inherent this weak-nesses with respect to the underlying probability model, and is therefore ineffective for mining high-quality sequential patterns from uncertain sequence databases.

B. System Design

In uncertain sequence data mining, many real-world applications like sensor networks as well as customer purchase sequence. To mine frequent sequential patterns from uncertain data three different approaches of p-FSE are proposed in this paper they are:

1) The uncertain data is collected from external data sources.
2) An approximate approach which approximates the frequency of episode using probability models.
3) An optimized approach which efficiently prunes candidate episodes by estimating an upper bound of its frequentness probability using approximation techniques.
4) The cleaning engine is applying Prefixspan algorithm to mine the data item-sets.

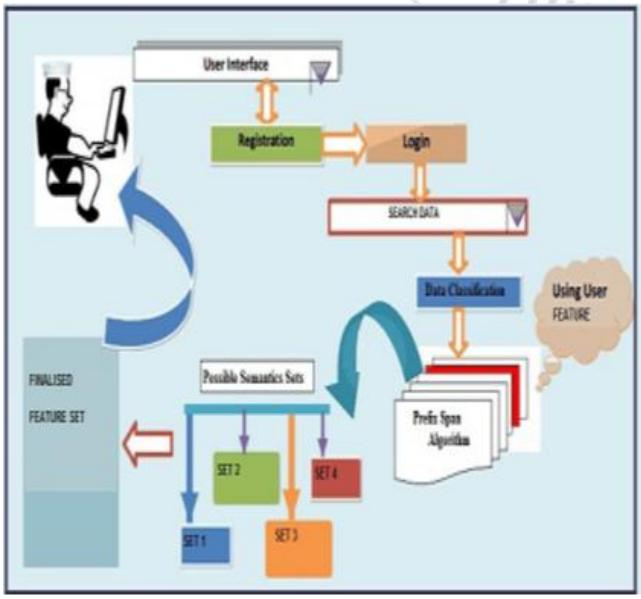


Figure 3.1: System Diagram

Two types of frequentness dimensions for an uncertain pattern are:

1. Expected support and
2. Probabilistic frequentness

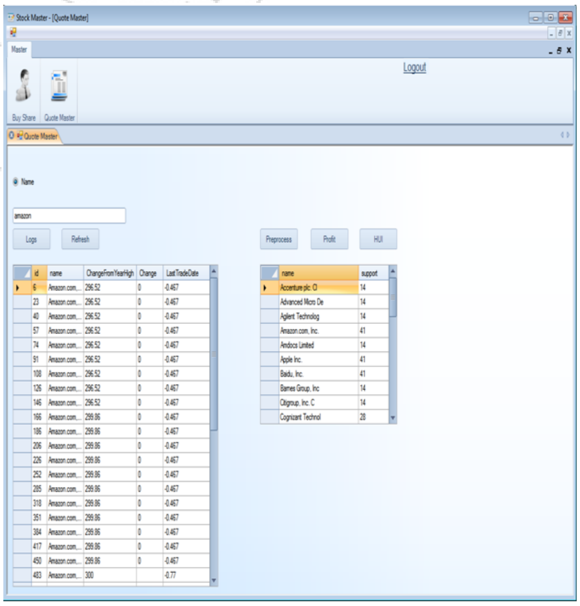
To estimate closeness of an item first applies uncertain clustering algorithm.

Sequence level U-PrefixSpan which is modification made on PrefixSpan to discover in order pattern from uncertain datasets. Preceding work is of predictable sustained which dealings pattern frequentness and it is not able to mine high quality sequential pattern. FSP is helpful in numerous conducts like for mobile tracking. It is used to categorize or to create group/cluster of items and in biological used for inherited sequential mining. UPrefixSpan overcome the challenges with the p-FSP algorithm which is to authenticate data to the sequence level UPrefixSpan. It uses PrefixSpan projection and pattern growth algorithm to handle the problem.

C. Algorithm

```
1) Scan S|α once,
   find the set of frequent items b such that:
   • b can be assembled to the last element of α to form a
     sequential pattern; or
   • <b> can be appended to α to form a sequential pattern.
2) For each frequent item b:
   • append it to α to form a sequential pattern α' and output
     α';
   • output α';
3) For each α':
   • construct α'-projected database S|α' and
   • call PrefixSpan(α', L+1, S|α').
```

4. Results



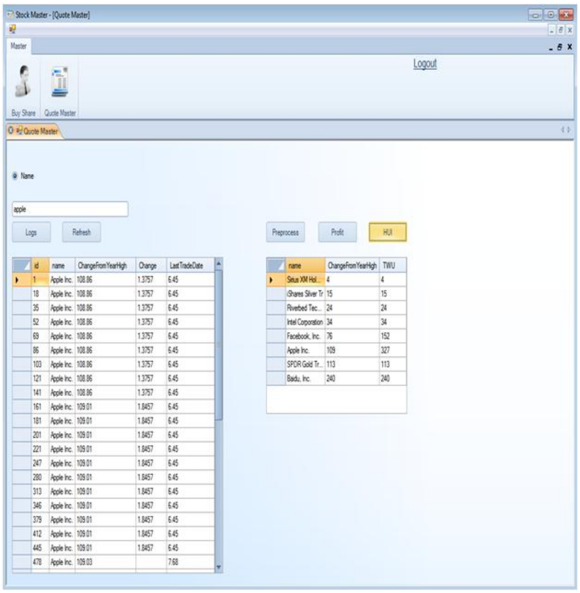
○ Name

amazon

id	name	ChangeFromYearHigh	Change	LastTradeDate
6	Amazon.com,...	296.52	0	-0.467
23	Amazon.com,...	296.52	0	-0.467
40	Amazon.com,...	296.52	0	-0.467
57	Amazon.com,...	296.52	0	-0.467
74	Amazon.com,...	296.52	0	-0.467
91	Amazon.com,...	296.52	0	-0.467
108	Amazon.com,...	296.52	0	-0.467
126	Amazon.com,...	296.52	0	-0.467
146	Amazon.com,...	296.52	0	-0.467
166	Amazon.com,...	299.86	0	-0.467
186	Amazon.com,...	299.86	0	-0.467
206	Amazon.com,...	299.86	0	-0.467
226	Amazon.com,...	299.86	0	-0.467
252	Amazon.com,...	299.86	0	-0.467
285	Amazon.com,...	299.86	0	-0.467
318	Amazon.com,...	299.86	0	-0.467
351	Amazon.com,...	299.86	0	-0.467
384	Amazon.com,...	299.86	0	-0.467
417	Amazon.com,...	299.86	0	-0.467
450	Amazon.com,...	299.86	0	-0.467
483	Amazon.com,...	300		-0.77

name	support
Accenture plc. Cl	14
Advanced Micro De	14
Agilent Technolog	14
Amazon.com, Inc.	41
Amdocs Limited	14
Apple Inc.	41
Baidu, Inc.	41
Barnes Group, Inc	14
Citigroup, Inc. C	14
Cognizant Technol	28

Figure 4.1: Pre-process Companies



apple

id	name	ChangeFromYearHigh	Change	LastTradeDate
1	Apple Inc.	108.86	1.3757	6.45
18	Apple Inc.	108.86	1.3757	6.45
35	Apple Inc.	108.86	1.3757	6.45
52	Apple Inc.	108.86	1.3757	6.45
69	Apple Inc.	108.86	1.3757	6.45
86	Apple Inc.	108.86	1.3757	6.45
103	Apple Inc.	108.86	1.3757	6.45
121	Apple Inc.	108.86	1.3757	6.45
141	Apple Inc.	108.86	1.3757	6.45
161	Apple Inc.	109.01	1.8457	6.45
181	Apple Inc.	109.01	1.8457	6.45
201	Apple Inc.	109.01	1.8457	6.45
221	Apple Inc.	109.01	1.8457	6.45
247	Apple Inc.	109.01	1.8457	6.45
280	Apple Inc.	109.01	1.8457	6.45
313	Apple Inc.	109.01	1.8457	6.45
346	Apple Inc.	109.01	1.8457	6.45
379	Apple Inc.	109.01	1.8457	6.45
412	Apple Inc.	109.01	1.8457	6.45
445	Apple Inc.	109.01	1.8457	6.45
478	Apple Inc.	109.03		7.68

name	ChangeFromYearHigh	TWU
Sirius XM Hol...	4	4
iShares Silver Tr	15	15
Riverbed Tec...	24	24
Intel Corporation	34	34
Facebook, Inc.	76	152
Apple Inc.	109	327
SPDR Gold Tr...	113	113
Baidu, Inc.	240	240

Figure 4.2: High Utility Item-set (HUI)

5. Conclusion & Future Scope

Uncertainty will be extremely habitual in every kind of databases. To tackle the problem of insecurity this paper has examined calculation associated to it. This paper has centered on calculations which mines successive design from unverifiable database. Past work utilizes help consider a premise to take care of the issue. PrefixSpan is mainly broadly utilized calculation to tackle the issue. This project will think about the issue of mining probabilistically incessant successive examples (p-Fsps) in dubious databases. The rules implemented are able to improve the mining efficiency. This paper has discussed multiple algorithms like sequence level U-PrefixSpan, p-FSE, Element level UPrefixSpan etc. all this algorithm is especially helpful to mine sequential pattern from uncertain database. The future work can be comprehensive in influential the probability with user requirement that authenticates the outcome with guarantee about excellence and appreciation.

References

[1] M. Muzammal and R. Raman, "Mining sequential patterns from probabilistic databases", in Proc. 15th PAKDD, Shenzhen, China, 2011

[2] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining", in Proc. 13th ACM SIGKDD, San Jose, CA, USA, 2007

[3] D. Tanasa, J. A. Lpez, and B. Trousse, "Extracting sequential patterns for gene regulatory expressions proles", in Proc. KELSI, Milan, Italy, 2004.

[4] J. Pei et al., "PrexSpan: Mining sequential patterns efciently by prexprojected pattern growth", in Proc. 17th ICDE, Berlin, Germany, 2001

[5] R. Agrawal and R. Srikant, "Mining sequential patterns", in Proc. 11th ICDE, Taipei, Taiwan, 1995

[6] M.J.Zaki, "SPADE: An efficient algorithm for mining frequent sequences", Mach. Learn., vol. 42, no. 12, pp. 3160, 2001.

[7] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. C. Hsu, "FreeSpan: Frequent pattern-projected sequential pattern mining", in Proc. 6th SIGKDD, New York, NY, USA, 2000.

[8] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements", in Proc. 5th Int. Conf. EDBT, Avignon, France, 1996

[9] Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in uncertain databases", in Proc 15th Int. Conf. EDBT, New York, NY, USA, 2012

[10] C. Gao and J. Wang, "Direct mining of discriminative patterns for classifying uncertain data", in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.

[11] C. C. Aggarwal, Y. Li, J.Wang, and J.Wang, "Frequent pattern mining with uncertain data", in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[12] Q. Zhang, F. Li, and K. Yi, "Finding frequent items in probabilistic data", in Proc. ACM SIGMOD, Vancouver, BC, Canada, 2008

[13] Nikos Pelekis, Ioannis Kopanakis, Evangelos E. Kotsifakos, Elias Frentzos "Clustering uncertain trajectories", 2010

[14] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng, "Mining uncertain data with probabilistic guarantees," in *Proc. 16th ACM SIGKDD*, Washington, DC, 2010.

[15] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in *Proc. 15th ACM SIGKDD*, Paris, France, 2009.

[16] P. Agrawal *et al.*, "Trio: A system for data, uncertainty, and lineage," in *Proc. VLDB*, Seoul, Korea, 2006.

[17] X. Lian and L. Chen, "Set similarity join on probabilistic data," in *Proc. VLDB*, Singapore, 2010

[18] J. Jestes, F. Li, Z. Yan, and K. Yi, "Probabilistic string similarity joins," in *Proc. ACM SIGMOD*, Indianapolis, IN, USA, 2010

[19] Z. Zou, J. Li, and H. Gao, "Discovering frequent sub-graphs over uncertain graph databases under probabilistic semantics," in *Proc. 16th ACM SIGKDD*, Washington, DC, USA, 2010

[20] D. Cheung, J. Han, V. Ng, and C. Wong, "Maintenance of Discovered Association Rules in Large Databases:An Incremental Updating Technique," Proc. 12th Int'lConf. Data Eng. (ICDE), 1996