

## Examination Mining High Utility Patterns

Satyajeet Bankar<sup>1</sup>, Krushna Barve<sup>2</sup>, Rohit Dusane<sup>3</sup>, Tejas Desale<sup>4</sup>

<sup>1</sup> Student , Computer Engineering , LOGMIEER, Maharashtra, India

<sup>2</sup> Student , Computer Engineering , LOGMIEER, Maharashtra, India

<sup>3</sup> Student , Computer Engineering , LOGMIEER, Maharashtra, India

<sup>4</sup> Student , Computer Engineering , LOGMIEER, Maharashtra, India

\*\*\*

**Abstract** - In utility mining is a new development of data mining technology. Among utility mining problems, utility mining with the item set share framework is a difficult one as there is order reversing property that holds interestingness measures. In prior we works on this concepts all employ a two-phase, candidate generation strategy with one exception that is how-ever not sufficient and not scalable with large databases. The two-phase approach suffers from scalability issue due to the huge number of candidates. This project uses a novel algorithm that depends high utility patterns in a single phase without generating candidates. The new innovations lie in a high utility pattern growth approach, a look-ahead strategy, and a linear data structure. In our pattern growth approach is to search a reverse set enumeration tree and to prune search space by utility upper bounding.

**Key Words:** *HUP (High Utility Patterns), XUT (External utility Table), d2HUP (Direct Discovery Of High Utility Patterns).*

### 1. INTRODUCTION

Utility mining has raised recently to address the limitation of frequent pattern mining by seeing the user's prospector objective as well as the raw data. Utility mining with the itemset share structure for example, discovering combination so products with high profit or revenues, is much harder than other categories of utility mining problems, for instance, weighted itemset mining and objective-oriented utility-based association mining.

Data mining is defined broadly as a process to extract implicit, people do not know in advance, but is potentially useful information and knowledge from a lot of noisy, uncertain, and stored in various forms or incomplete large data sets.

The experiment is the number of candidates can be huge, which is the scalability and proficiency hold up. Although a lot of effort has been made to reduce the number of candidates generated in the first phase, the test still persists when the raw data contains many long transactions or the least utility threshold is small.

### 2. Related Work

High utility pattern mining problem is thoroughly related to frequent pattern mining, with constraint-based mining. In this section, we briefly review prior works both on frequent pattern mining and on utility mining, and discuss how our work connects to and differs from the prior works.

### 2.1 Comparative Analysis

Paper Name	Author	Domain	Issues	Solution	Algorithm	Implementation	Conclusion
Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-Mining-As-a-Service Paradigm	1)Boxing Dong 2) Hui Wang	Data Mining	how can the client of weak computational power verify that the server returned correct mining result?	Before sending the dataset D to the server, the client constructs an authenticated data structure.	Bilinear algorithm	We implemented a prototype of our probabilistic approach in Java,	In this paper, we present two integrity verification approaches for outsourced frequent itemset mining.
A Study on Effective Chart Visualization and Interaction Techniques for Time-Series Data	1)James Walker 2) Rita Borgo	Data Mining	Pilot study results were positive and revealed trends in the data supporting our initial hypothesis,	Regular basis (every couple of weeks). They provide us with a constant source of feedback.	Layout algorithm	We have implemented all of the methods in our software in a consistent way.	In this paper, we comparatively evaluate existing methods for exploring time-series data.
A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining	1)LelXu 2) Yan chen	Data Mining	A wider perspective than data mining, we investigated a range of approaches across the entire KDD process.	To protect sensitive information in mined data, researchers need a variety of ongoing work.	Sanitation algorithm	To find that agreement, we developed a game model, based heavily on earlier <i>k</i> -anonymity algorithm to the gathered data.	In this paper, we comparatively work has broaden the horizons for early assessment and Predicon.
Mining Disease Sequential Risk Patterns from Nationwide Clinical Databases for Early Assessment of Chronic Obstructive Pulmonary Disease	1)Yi-Ting Cheng	Data Mining	To our best knowledge, this is the first work that addresses the important issue of early assessment on COPD through mining	propose a novel approach for early assessment on COPD by mining COPD-related sequential risk patterns from diagnostic records.	CBS (classify by sequence) algorithm	We can effectively classify sequential risk patterns for prediction. To support our work of COPD assessment and analysis.	The approach and corresponding techniques presented in this work has broaden the horizons for early assessment and prediction on the disease
Review on State of Art Data Mining and Machine Learning Techniques for Intelligent Airport Systems	1)Jain Bo 2) Chamta Aryawana	Data Mining	The challenge faced by airports today is the complexity of players and processes.	The airport ecosystem must become more 'intelligent' to optimize its supply chain.	FP-TREE algorithm	The starving management bodies of airports have always tried to implement isolated solutions	A system with high recalls but low precision returns many results, but most of its predicted labels are incorrect

Accelerating Graph Mining Algorithms Via Uniform Random Edge Sampling	1) Ruohan Gao 2) Pili Hu	Data Mining	The random subgraphs resulted from URE sampling can often be used as substitutes for the original graphs.	These graph mining tasks can yield well-approximated solutions for the original graph with substantial reduction.	Apriori algorithm	We demonstrate how to make use of URE sampling to accelerate common graph mining	This work is an attempt to use Uniform Random Edge (URE) sampling to accelerate graph mining algorithms.
An Approach Based On Association Rules Mining To Improve Road Safety In Morocco	1) Garnat hi Phatim a 2) R. Agrawal	Data Mining	we find three problems like, sorting, selection, and arrangement. In this context of an important number of extracted rules.	resulting from an accident using real data obtained from the Ministry of Equipment and Transport of morocco.	Machine learning algorithm	The implementation of the case provides a preference relationship between rules and profiles.	In this paper, we have discussed the usefulness and relevance problem issued from a KDD process.
Use of Reality Mining Dataset for Human Behavior Analysis – A Survey	1) Sunita K v n 2) Rama prasad	Cloud Computing	This system addressed various issues like who should receive an incoming notification	The sensor data collected from smart phones is experimented with survey data	Significant Subgraph algorithm	Implement numerous applications and systems, such as a mobile app that helps smokers quit smoking.	Smart phones of current and future generations will play a vital role in everyone's life.
Mining the Enriched Subgraphs for Specific Vertices in a Biological Graph	1) Pieter Meysman 2) Kris Laukens	Data Mining	A common problem is the discovery of frequent subgraphs in a graph data set.	Detect personalized and abnormal behaviors of internet users, We propose Sequential topic patterns.	DP-based algorithm	This implementation allows the user to set any of the introduced interestingness.	In this paper, we have presented a novel type of subgraph mining approach for subgroup discovery subgraph.
Mining User - Aware Rare Sequential Topic Pattern in Document Stream.	1) Yunkun Hu 2) Gurumit Baggo	Cloud Computing	In this paper we formally define and a group of algorithm and design to combine systematically solves this problem.	Its algorithm performance directly affects the efficiency of data mining and the integrity, effectiveness of ultimate data mining results.	Apriori algorithm	its algorithm performance directly affects the efficiency of data mining and the integrity, effectiveness of ultimate data mining results.	Mining URSTP in publish document stream on the internet is the significant and challenging problem.

## 2.2 Algorithm

d2HUP(D, XUT, minU til)

1 create the root of prefix extension tree

2 build T S caul({}) with  $\Omega$  in descending order of uBitem

3 N  $\leftarrow$  the root

4 W  $\leftarrow$  {i | i < pat(N)  $\wedge$  uBitem (i, pat(N))  $\geq$  minU til}

5 if  $\forall i \in W, s(\text{pat}(N)) = s(\{i\} \cup \text{pat}(N)) \wedge$

$u(\{i\} \cup \text{pat}(N)) \geq \text{minU til}$  then output each non-empty

subset of  $W \cup \text{pat}(N)$  as an HUP, goto step 12

6  $\Delta \leftarrow$  minimum of  $u(\{j\} \cup \text{pat}(N)) - u(\text{pat}(N)), \forall j \in W$

7 if  $\forall i \in W, s(\text{pat}(N)) = s(\{i\} \cup \text{pat}(N)) \wedge$

$\text{minU til} \leq u(W \cup \text{pat}(N)) < \text{minU til} + \Delta$

then output  $W \cup \text{pat}(N)$  as an HUP, goto step 12

8 for each item  $i \in W$  do

9 if  $u(\{i\} \cup \text{pat}(N)) \geq \text{minU til}$  then output  $\{i\} \cup \text{pat}(N)$

10 if  $u(B \text{ fpe}(\{i\} \cup \text{pat}(N))) \geq \text{minU til}$

then create a child node of N for i

11 end for each

12 while N is not null and has no child do

13 P  $\leftarrow$  parent(N), delete N, N  $\leftarrow$  P

14 end while

15 if N is null then stop

else P  $\leftarrow$  N, N  $\leftarrow$  firstChild(N)

16 build T S caul(pat(N)) by projection, goto Step 4.

### 3. Performance Analysis of Association Rule Mining Concept

At present, there are many frequent item sets generation algorithms, and when generating frequent k- item sets these algorithms will scan each transaction of the database for statistics of the support of k- item sets, and according to the determined minimum support find all frequent k- item sets at the first k iteration . However, because that the size of the databases usually very large, so the above method consumes too much time.

### 4. Frequent Pattern Mining

Frequent pattern mining was first projected by Agarwal, which is to discover all patterns whose supports are no less than a user-define minimum support threshold. Frequent pattern mining employs the anti-monotonicity property. The provision of a superset of a pattern is no more than the provision of the pattern. Algorithms for mining frequent patterns as well as algorithms for mining high utility patterns tumble into three classes, breadth-first search, depth-first search, and hybrid search.

### 5. Constraint-Based Mining

Constraint-based mining is a landmark in evolving from frequent pattern mining to utility mining. Works on this are mainly focus on how to push constraints into frequent pat-tern mining algorithms. Peietal discussed constraints that are related to (normalized) weighted ropes.

L.DeRaedt inspected how standard constraint programming techniques can be applied to constraint-based mining difficulties with constraints that are monotone, anti-monotone, and adaptable.

### 6. Conclusion and Future Scope

This paper proposes a new algorithm, d2HUP, for utility mining with the item sets are frame work, which finds high utility patterns without candidate generation. Our influences contain: A linear data structure, is proposed, which marks the root cause of the two-phase, candidate generation approach adopted by prior algorithms, that is, their data structures cannot keep the unique utility information.

### Acknowledgment

In the future, we will work on high utility sequential pattern mining, parallel and distributed algorithms, and their application in big data analytics. With the development of information technology, data mining technology has obtained widespread concern, which prompted the scientific community more in-depth study of this technology. Data mining technology has many research fields, in which association rule mining is one of the important research directions, the in-depth study on it not only has important theoretical significance, but also has a very important application value.

### REFERENCES

- [1]R.Agarwal, C.Agarwal, and V.Prasad, "Depth first generation of long patterns," in Proc. ACMSIGKDDInt.Conf.Knowl.DiscoveryDataMining,2000,pp.108–118.
- [2]R.Agrawal, T. Imielinski, and A.Swami, "Mining association rules between sets of items in large databases," in Proc.ACMSIG-MODInt.Conf.Manage.Data,1993,pp.207–216.
- [3]R.Agrawal and R.Srikant,"Fast algorithms for mining association rules," in Proc. 20thInt.Conf.VeryLargeDatabases,1994,pp.487–499.
- [4] L.DeRaedt, T.Guns, andS.Nijssen, "Constraint programming for itemset mining," in Proc.ACMSIGKDD, 2008, pp.204–212.

[5] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan. Frequent Pattern Mining: Current Status and Future Directions[J], Data Mining and Knowledge Discovery, 2007, volume(15): 55-86.

[6] HAN J, PEI J, YIN Y. Mining Frequent Patterns Without Candidate Generation[C], Proc of the 2000 ACM SIGMOD Internal Conference on Management of Data. Dallas, Texas: ACM Press, 2000: 1-12.

[7] Dickson, 2014. A Survey of Predictive Analytics in Data Mining with Big Data I Dickson Lam.

[8] Eagle N., Greene K., "Reality Mining – Using Big Data to Engineer a Better World", Association of American Publishers, MIT Press.

[9] C.K. Chui and B. Kao, "A decremental approach for mining frequent item sets from uncertain data," in Proc. PAKDD'08, 2008, pp. 64-75.

[10] M. J. Zaki, "Efficient enumeration of frequent sequences", Journal of Machine learning, vol. 42, no. 1-2, pp. 31-60, 2001

[11] C. T. Li, Emerging Digital Forensics Applications for Crime Detection, Prevention and Security, IGI-Global, 2013.