Frequent Itemsets Mining in Transactional Databases using Utility Pattern Algorithms

S. Shantha kumar[1] V. Anusuya[2]
[1] PG Student[2] Assistant Professor (Sl. Gr)
[1, 2]Computer Science Engineering
[1, 2] P.S.R Engineering College, Sivakasi, India

Abstract---Recently, high utility pattern (HUP) mining is one of the most important research issues in data mining due to its ability to consider the non binary frequency values of items in transactions and different profit values for every item. On the other hand, incremental and interactive data mining provide the ability to use previous data structures and mining results in order to reduce unnecessary calculations when a database is updated, or when the minimum threshold is changed. This paper proposes three novel tree structures to efficiently perform incremental and interactive HUP mining. This propose two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database. The first tree structure, Incremental UP Lexicographic Tree (IUP_L-Tree), is arranged according to an item's lexicographic order. It can capture the incremental data without any restructuring operation. The second tree structure is the IHUP Transaction frequency Tree (IUP_{TF}-Tree), which obtains a compact size by arranging items according to their transaction frequency (descending order). To reduce the mining time, the third tree, IHUP-Transaction-Weighted Utilization Tree (IUP_{TWU}-Tree) is designed based on the TWU value of items in descending order. Extensive performance analyses show that our tree structures are very efficient and scalable for incremental and interactive HUP mining.

Keywords: frequent itemset, high utility itemset, utility mining, data mining

I. INTRODUCTION

A. *Overview of Data Mining*

Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices such as marketing, surveillance, fraud detection, and scientific discovery.

Data Mining, also popularly known as Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

1) Types of information

Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

a) Business transactions

Every transaction in the business industry is memorized for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock

b) Scientific data

Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed

c) Medical and personal data

From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele.

d) Text reports and memos

Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

e) The World Wide Web repositories

Since the inception of the World Wide Web in 1993, documents of all sorts of formats, content and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers. Many believe that the World Wide Web will become the compilation of human knowledge.

2) Data Preprocessing

Data cleaning ,which is also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection. Data integration at this stage, multiple data sources, often heterogeneous, may be combined in a common source. Data selection at this step, the data relevant to the analysis is decided on and retrieved from the data collection. Data transformation also known as data

consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining, which is the crucial step in which clever techniques are the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

3) Kind of knowledge mined

Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time series databases and textual databases, and even flat files. Here are some examples in more detail.

Flat files are actually the most common data source for data mining algorithms especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

A data warehouse as a storehouse, is a repository of data collected from multiple data sources and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof. Let us suppose that Our VideoStore becomes a franchise in North America. Many video stores belonging to Our VideoStore company may have different databases and different structures. If the executive of the company wants to access the data from all stores for strategic decision-making, future direction, marketing, it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores would be loaded, cleaned, transformed and integrated together. To facilitate decision making and multidimensional views, data warehouses are usually modeled by a multi-dimensional data structure.

A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items.

Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms. Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time. The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications.

4) Data mining systems

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive. Data mining systems can be categorized according to various criteria among other classification are the following

This classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web. This classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse transactional. This classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, and clustering.

Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, and visualization

5) Issues in Data Mining

Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the inform .The knowledge discovered by data mining tools is useful as long as it is interesting and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual

presentation. The major issues related to user interfaces and visualization are screen real-estate, information rendering, and interaction.

These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs, the assessment of the knowledge discovered, are all examples that can dictate mining methodology choices. Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data.

II. RELATED WORKS

As Reviwed in [2], Traditional association rules mining techniques treat all items in the database equally by taking into consideration only the presence of an item within a transaction. The necessity to develop methods for discovering association patterns to increase a utility within a given application has long been recognized in the data mining community. So far, this has been addressed by modeling specific association patterns that are either statistically (based on support and confidence), or semantically (based on objective utility) related to an objective defined by the user.

Traditional association rules mining cannot meet the demands arising from some real applications. By considering the different values of individual items as utilities, utility mining focuses on identifying the item sets with high utilities. This paper present a Two-Phase algorithm to efficiently prune down the number of candidates and precisely obtain the complete set of high utility itemsets. It performs very efficiently in terms of speed and memory cost both on synthetic and real databases, even on large databases that are difficult for existing algorithms to handle.

Traditional Association rules mining model treat all the items in the database equally by only considering if an item is present in a transaction or not. Frequent itemsets identified by ARM may only contribute a small portion of the overall profit, whereas nonfrequent itemsets may contribute a large portion of the profit. In reality, a retail business may be interested in identifying its most valuable customers (customers who contribute a major fraction of the profits to the company). These are the customers, who may buy full priced items, high margin items, or gourmet items, which may be absent from a large number of transactions because most customers do not buy these items. In a traditional frequency oriented ARM, these transactions representing highly profitable customers may be left out. Utility mining is likely to be useful in a wide range of practical applications. Utility is a measure of how useful an itemset is.

The goal of utility mining is to identify high utility itemsets that drive a large portion of the total utility. Traditional ARM problem is a special case of utility mining, where the utility of each item is always 1 and the sales quantity is either 0 or 1.

As Reviwed in [1], We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. This present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. This also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period was available on the computer.

A. Discovering large itemsets

Given a set of items I, an itemset X + Y of items in I is said to be an extension of the itemset X if X ∩Y =Φ ;. The parameter db_{size} is the total number of tuples in the database.

Initially the frontier set consists of only one element, which is an empty set. At the end of a pass, the support for a candidate itemset is compared with minsupport to determine if it is a large itemset. At the same time, it is determined if this itemset should be added to the frontier set for the next pass. The algorithm terminates when the frontier set becomes empty.

III. SYSTEM DESIGN

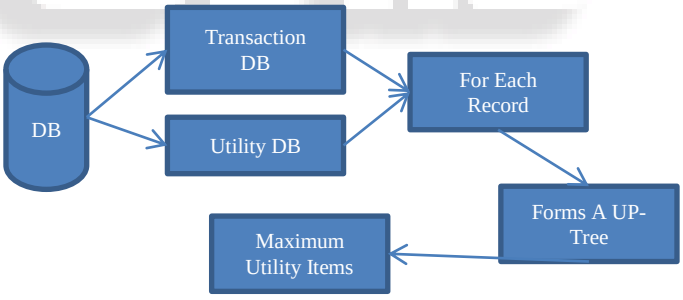

Fig. 1: Architecture

IV. RESULTS

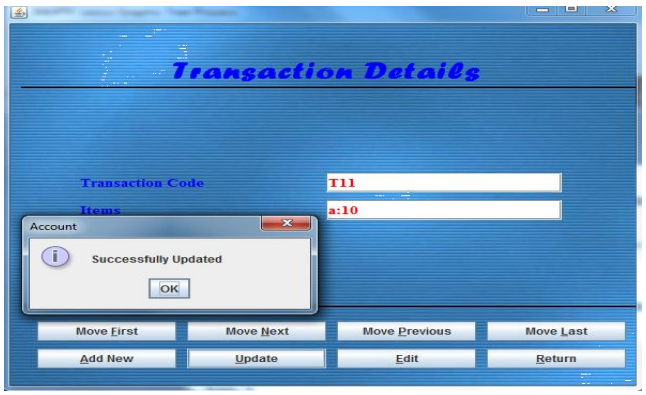

Fig. 2: Transaction details (insertion items for high utility itemsets and high transaction frequency)

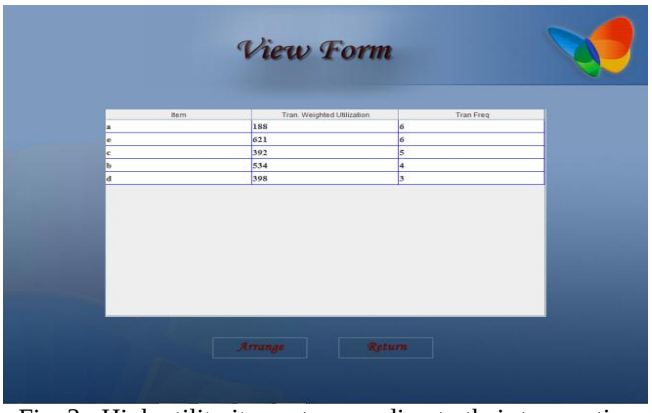

Fig. 3: High utility itemsets according to their transaction frequency

Fig. 4: High utility itemsets according to their transaction weighted utilization.

V. CONCLUSION

In this paper, we have proposed two efficient algorithms named UP-Growth and UP-Growth+ for mining high utility itemsets from transaction databases. A data structure named UP-Tree was proposed for maintaining the information of high utility itemsets. PHUIs can be efficiently generated from UP-Tree with only two database scans. Moreover, we developed several strategies to decrease overestimated utility and enhance the performance of utility mining. In the experiments, both real and synthetic data sets were used to perform a thorough performance evaluation. Results show that the strategies considerably improved performance by reducing both the search space and the number of candidates. Moreover, the proposed algorithms, especially UP-Growth+, outperform the state of-the-art algorithms substantially especially when databases contain lots of long transactions or a low minimum utility threshold is used.

REFERENCES

[1] Efficient Algorithms For Mining High Utility Itemsets From Transactional Databases Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, And Philip S. Yu, Fellow, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 8, AUGUST 2013

[2] Erwin, R.P. Gopalan, And N.R. Achuthan, "CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using The Pattern Growth Approach," Proc. Seventh IEEE Int'l Conf. Computer And Information Technology (CIT '07), Pp. 71-76, 2007.

[3] Mining Weighted Association Rules Without Preassigned Weights Ke Sun And Fengshan Bai. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 4, APRIL 2008.

[4] Efficient Tree Structures For High Utility Pattern Mining In Incremental Databases Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer,Byeong-Soo Jeong, And Young-Koo Lee, Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 12, DECEMBER 2009.

