Top-K High Utility itemset Mining Implemented on Medical Data

Akash Shinde
Department of Information Technology
Atharva College of Engineering
Mumbai, India

Komal Pareek
Department of Information Technology
Atharva College of Engineering
Mumbai, India

Ruchi Tiwari
Department of Information Technology
Atharva College of Engineering
Mumbai, India

Sachin Gavhane
Department of Information Technology
Atharva College of Engineering
Mumbai, India

***Abstract*— Top-k high utility itemset mining refers to the discovery of top-k patterns using a user-specified value k by considering the utility of items in a medical database. Existing top-k high utility itemset mining algorithms are based on the pattern-growth method which leads to the generation of many candidate keys and additional database scan for calculating exact utilities is unavoidable which leads to more memory usage. We propose a new algorithm, TKUL-Miner, to mine top-k high utility itemsets efficiently both in runtime and memory usage. It utilizes a new utility-list structure which stores necessary information at each node in the search tree for mining the itemsets. The proposed algorithm has a strategy using search order for a specific region to raise the border minimum utility threshold rapidly. Moreover, there are two additional strategies for calculating smaller overestimated utilities which are suggested to prune unpromising itemsets effectively which do not lead to the goal. Various experimental results have shown that the new proposed algorithm (TKUL-Miner) outperforms other top-k high utility algorithms both in runtime and memory efficiency.**

Keywords*— *High utility; itemset; top-k pattern mining; utility-list structure; data mining

I. INTRODUCTION

The health care domain has a lot of challenges and one of the main difficult challenges is in disease diagnosis. The data mining is the process of analysing huge data from a different perspective and summarizing it into useful information[1].
Clinical databases are elements of the domain where the procedure of data mining has developed into an inevitable aspect due to the gradual incline of medical and clinical research data. It is possible for the health care industries to gain the advantage of data mining by employing the same as an intelligent diagnostic tool. It is possible to acquire knowledge and information concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, the data mining has been developed into a vital domain in health care[2]. It is possible to predict the efficiency of medical treatments by building the data mining applications. Data mining can deliver an assessment of which courses of action prove effective by comparing and evaluating causes, symptoms, and courses of treatments [3].

The real-life data mining applications are attractive since they provide data miners with a varied set of problems, time and again. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process[4], [5].The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with the aid of data mining techniques. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. The proposed system is a system of searching through large amounts of data for patterns. The main goal of the proposed system is to extract important information from data that was not previously known. Data mining is commonly used to recognize certain patterns or trends. One important factor of data mining is that it will often be used to analyse information from a variety of different perspectives. The important information that is gained from data mining can be used to increase profits or lower costs. The goal of the person who uses data mining is, he/she should be able to predict certain behaviours or patterns. Once the user is able to predict the behaviour of something which he analysing, he will be able to make strategic decisions that can allow him to achieve certain goals. The main objective of this project is to create a fast, easy and an efficient mode for disease prediction, with less error rate and can apply with even large data sets and show reasonable patterns with dependent variables. For disease identification and prediction in data mining, appropriate algorithm should be used in order to maximize the accuracy rate

II. PROBLEM STATEMENT

Based on the above context, there is a need for a classification of data, which take care of all the different aspects in analysing the frequently occurring diseases. The proposed system devices a simple system, using various data mining algorithms to obtain better statistics from the available data. Healthcare industry today generates large amounts of complex data about patients, hospitals resources, diseases, diagnosis methods, electronic patient's records, etc. The data mining techniques are very useful to make medicinal decisions in curing diseases. The system planned to be developed will help in drawing effective conclusions. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. In the proposed system, we will develop a method to identify the frequency of diseases in particular geographical area at given time period with the help of data mining tools.

III. SCOPE OF PROPOSED SYSTEM

A large population needs a great demand for medical services. But their deficiency creates a problem and the console plays very important role to some extent. It facilitates the users to predict them even if they are at the remote location and it is very hard to reach doctors regularly. This process becomes cost and time saving if we integrate it to web portals. Data mining derives its name from the similarities between searching for valuable business information in a large database. For given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

A. Automated prediction of trends and behaviors

The proposed system automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted disease. Data mining uses data on past promotional entries to identify the targets most likely to maximize return in future predictions.

B. Automated discovery of previously unknown patterns

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of symptoms data to identify seemingly unrelated symptoms that often occur together. Other pattern discovery problems include identifying anomalous data that could represent data entry keying errors
Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high-performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

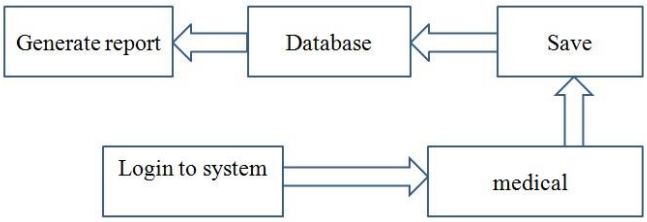

Fig.1 System Block Diagram

C. TKUL-Miner algorithm

The database with the user specified value k is given before the mining process starts. The border minimum utility threshold called minutil which is initially set to zero gradually increases during the mining process of the top-k HUI. The proposed algorithm starts with a pre-processing. It scans the database twice to construct the data structures including the utility-lists which are used throughout the whole mining process. After the pre-processing, the TKUL-Miner steps into the main top-k HUI mining. It extends searching from an item to each itemset by performing join operation of the utility-lists to find top-k high utility itemsets. Whenever an itemset contains no less utility than the minutil, it is added to the minimum heap. If the minimum heap is full and the smallest utility of the itemsets in the heap is bigger than the minutil, the minutil is updated to the smallest utility. This process ends when there are no more itemsets to generate. Moreover, the TKUL-Miner algorithm adopts several strategies called FSD, RUZ, and FCU to raise the minutil rapidly and prune the search space effectively, which contributes a lot to reduce execution time.

IV. ALGORITHM

TKUL Algorithm

1. Start
2. Select Excel File (Crude Database).
3. Import Data from Excel to SQL Server (Pre-Processing).
   Clean the Database if required.
4. Start the Algorithm.
   Scan Database Iteratively.
5. Searching for an element from each dataset to find Top-K high utility datasets.
6. Whenever a dataset contains no less utility than the minimum threshold, it is added to the minimum heap.
7. This process ends when there are no more datasets to generate.
8. Stop.

Fig.2 TKUL-Miner algorithm

V. CONCLUSION

The proposed algorithm is unique and different from the commonly used prediction algorithms in data mining. The proposed methods overcome the disadvantages of existing methods. The algorithm proved efficient in terms of time and space complexity and proved to be accurate when compared with other algorithms. This algorithm can be enhanced by considering and incorporating many more parameters. For disease identification and prediction on the basis of various parameters such as age, sex, geographical areas, the same algorithm can be applied. The proposed algorithm is accurate according to the given scenario. The user can enter the symptoms to check the disease which is likely to affect him and can take preventive measures accordingly.

The system can be used by researchers in order to predict future diseases. The graphical representation helps in better understanding of the available statistics. This algorithm can be enhanced by considering and incorporating many more parameters and creating a new hybrid algorithm which should be more feasible according to the given environment.

ACKNOWLEDGMENT

This research was supported by the department of information technology, Atharva College of Engineering, Mumbai.

REFERENCES

[1] Miss. A. A. Bhosale, S. V. Patil, Miss. P. M. Tare, Miss. P. S. Kadam, “High Utility Itemsets Mining on Incremental Transactions using UP-Growth and UP-Growth+ Algorithm” International Journal on Recent and Innovation Trends in Computing and Communication (ISSN: 2321-8169 Volume: 2 Issue: 11 3366 - 3368.)

[2] R.Nandhini, Dr.N.Suguna “Shrewd Technique for Mining High Utility Itemset via TKU and TKO Algorithm.” R.Nandhini et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 5261-5264 . (ISSN: 0975-9646)

[3] Noppol Thangsupachai, Phichayasini Kitwatthanathawon, Supachanun Wanapu, and Nittaya Kerdprasop “Clustering Large Datasets with Apriori-based Algorithm and Concurrent Processing, Proceedings of the Internatiinal MultiConference of Engineerings and computer scientist” 2011 Vol 1, IMECS 2011, March 16-18, 2011, Hong Kong.

[4] KDD '12 Proceedings of the 18th ACM SIGKDD “International conference on Knowledge discovery and data mining” Conference KDD '12 The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Beijing, China — August 12 - 16, 2012 ACM New York, NY, USA ©2012

[5] Elena Baralis, Tania Cerquitelli, Silvia Chiusano “A persistent HY-Tree to efficiently support itemset mining on large” datasets 2010-03-22 ACM New York, NY, USA ©2010 (ISBN: 978-1-60558-639)

[6] Karim K. Hirji “Discovering data mining: from concept to implementation” Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1998 (ISBN:0-13-743980-6)

[7] Daniel Kunkle, Donghui Zhang, Gene Cooperman “Efficient mining of max frequent patterns in a generalized environment” Pages 810-811 ACM New York, NY, USA ©2006 (ISBN:1-59593-433-2)

[8] Mengchi Liu, Junfeng QuMining “High utility itemsets without candidate generation” ACM New York, NY, USA ©2012 (ISBN: 978-1-4503-1156-4)

[9] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Ted Gueniche, Prashant Barhate “Efficient Incremental High Utility Itemset Mining” 2015-10-07 ACM New York, NY, USA ©2015 (ISBN: 978-1-4503-3735-9)

[10] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier Viger “Efficient Algorithms for Mining Top-K High Utility Itemsets” IEEE Transactions on knowledge and data engineering, VOL. 28, NO. 1, January 2016 (ISSN: 10414347)