EFFICIENT RECOMMENDATION OF DE-IDENTIFICATION POLICIES USING MAP REDUCE

[1]Kamatchi.K, [2]Krithica.J.S, [3]Induja.D, [4]Meenakshi.S
[1]B.Tech, [2]B.Tech, [3]B.Tech, [4]HOD
[1]Dept. Of Information Technology,
[1]Jeppiaar SRR Engineering College, Chennai, India

Abstract: Many data owners are required to release the data in a variety of real world application, since it is of vital importance to discovery valuable information stay behind the data. In existing re-identification attacks on the health care data sets. Such data publish directly to the world that may violate the individual privacy. It is urgent to solve the re-identification risks by using efficient recommendation of de-identification policies to guarantee both privacy and utility of the sensitive data. Using De-identification policies to achieve such requirements, at the same time number of de-identification policies is large due to the quasi-identifier attributes. To control the tradeoff between data utility and data privacy, to select policies skyline computation techniques is used but it is challenging for large number of policies. In this paper, we propose algorithm known Sky-Filter-MR. It is fully based on Map Reduce technique. De-identification policies are used to overcome the challenge of the skyline computing in large policies that are represented by in the form bit-strings. A novel skyline computation scheme was used to remove unqualified policies using the approximate relationship for further improve the performance. In our proposed SKY-FILTER-MR algorithm are Extensive experiments in both real life datasets are out perform approach by up to four times faster. It indicates good scalability in large policy datasets.

IndexTerms **- Re-identification, De-identification, Sky-Filter-MR, Quasi-identifier.**

I. INTRODUCTION

In the age of big data, it is important to exchange and share data among different parties. For example, all registered hospitals in California of US are required to submit specific demographic data on some patients which have been in good condition .However, publishing those data containing sensitive information could violate individual's privacy. Privacy-Preserving Data Publication (PPDP) is becoming an important to get sufficient protection while maintain high data utility. The High utility pattern mining was proposed to solve the drawbacks of traditional frequent pattern mining approach that cannot handle various features of real-world applications, many different techniques and algorithms for high utility pattern mining have been developed. Moreover, several advanced methods for incremental data processing have been proposed in recent years as the sizes of recent databases obtained in the real world become larger.

In this paper, we introduce the basic concept of incremental high utility pattern mining and analyze various relevant methods. In addition, we also conduct performance evaluation for the methods with famous benchmark datasets in order to determine their detailed characteristics. The evaluation shows that the less candidate patterns make algorithms faster. High utility item sets mining is relevant for business vendors. So that they can give more offers to high utility item sets. To understand the above sentence we need to know what high utility item sets is. High utility item sets are those ones that yield high profit when sold together or alone that meets a user-specified minimum utility threshold from a transactional database. This high utility item sets mining is not a new topic, but it is an emerging area. A high utility policy is useful for business vendors so they can give more offers to high utility policies. When sold together, high utility policies are the ones that yield high profit or alone that meets a user-specified minimum utility threshold from a transactional database. This high utility policies mining is an emerging area.

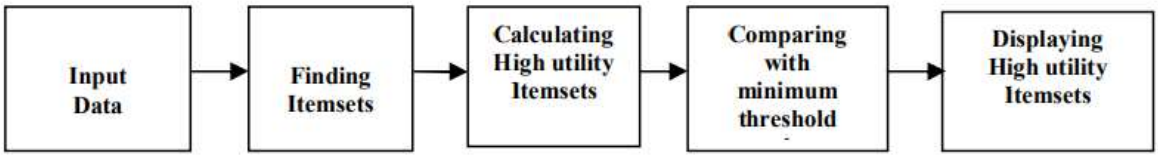


Figure 1 Processes for mining high utility item sets

II. AIM AND SCOPE

To improve the mining performance and to avoid scanning original database repeatedly, we use a compact map reduce to maintain the information of transactions and high utility item sets.

III. EXISTING SYSTEM

Big data is important to exchange and share data among different parties. However, publishing those data containing sensitive information could violate individual's privacy. In order to get sufficient protection while maintain high data utility, privacy-preserving data publication (PPDP) .In existing re-identification attacks on the health care datasets .

IV. PROBLEM IDENTIFIED

1. Publishing data set directly to the real world application sometimes it violate the individual privacy of the data.
2. There is no maintaining the privacy of the data and utility of the data.
3. Due to the quasi-identifier attributes, the de-identification policies are large.

V. PROPOSED SYSTEM

In this paper, we propose algorithm known Sky-Filter-MR. It is fully based on Map Reduce technique. A de-identification policy is used to overcome the challenge of the skyline computing in large policies that are represented by in the form bit-strings. Using De-identification policies to achieve such requirements, at the same time number of de-identification policies is large due to the quasi-identifier attributes. To protect personal sensitive information, various laws require that personal data which can be used to link one record in one table to another table containing explicit identifiers (e.g., name) based on QI attributes (e.g., Age, Gender, ZIP), should be de-identified. There are two approaches defined by the Health Insurance Portability and Accountability Act based on Privacy Rule to achieve de-identification. Safe Harbor and Expert Determination are known as the defined approaches.

The advantages of the proposed system are

1. Extensive experiments in both real life datasets are out perform approach by up to four times faster.
2. It indicates good scalability in large policy datasets.
3. To improve the performance, a novel approximate skyline computation scheme was proposed.
4. Privacy protection
5. Decrease the cost of computation over large number of policies.
6. Strengthened the policy space generation by power of filtering

The research methodologies are as follows:

1. Extract the load Dataset
2. De-identification
3. Sky-filter extraction using map reduce
4. Quasi-identifier assumptions
5. Statistical visualization

VI. PROPOSED MODEL

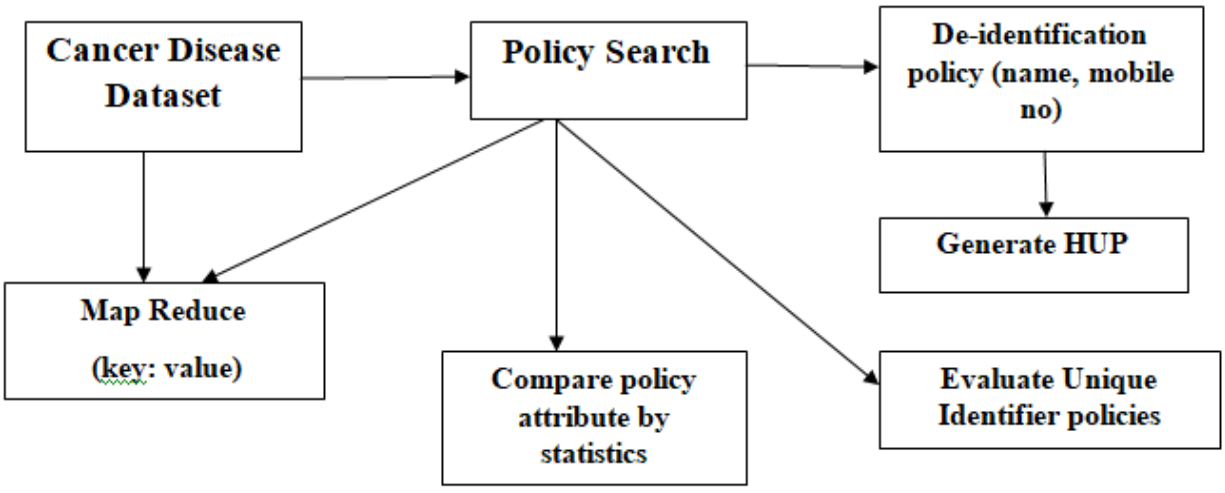


Figure 1 Architecture diagram

VII.BLOCK DIAGRAM EXPLANATION

The overall architecture which describes the policy searches and de-identification. Before the policy search and de-identification to be done, dataset have to be loaded which is of .xls format. There are different kinds of policy searches namely Map Reduce, Compare policy attribute by statistics, evaluate the unique identifier policies. Map Reduce uses the Key value pair. Policy attributes are compared based on different strategies and statistics are made. Dataset may have quasi-identifier attributes in order to make it unique, it is combined with another quasi-identifier attributes and evaluated. De-identification technique is applied to prevent individual's sensitive information's.

VIII. RESEARCH METHODOLOGY

8.1 EXTRACT THE LOAD DATASET

When upload the excel file is to determine an appropriate sampling parameter to divide the whole dataset effectively and fairly, which can ensure that each participating node has balanced storage and computation workload.

8.1.1 APACHE POI

A 100% open source library by Apache Software Foundation is the Apache POI. Most of the small and medium scale application developers rely a lot on Apache POI (HSSF + XSSF). It maintains all the fundamental features of Excel libraries; however, executing and text extraction are its main features.

8.1.2 HSSFSHEET

HSSFSheet is a class under the org.apache.poi.hssf.usermodel package. It creates excel spreadsheets and allows to format the sheet styles and sheet data.

Table 1 Constructor & Description

S.No.	Constructor & Description
1	**HSSFSheet(HSSFWorkbook workbook)** Creates new HSSFSheet called by HSSFWorkbook to create a sheet from scratch.
2	**HSSFSheet(HSSFWorkbook workbook, InternalSheet sheet)** Creates an HSSFSheet representing the given sheet object.

8.2 DE-IDENTIFICATION

De-identify the data record; the QI values of each tuple have two choices:

i) Remain the same

ii) Be recoded to the generalization state.

Meanwhile, we choose the full-sub tree generalization model, which means all QI values are in an ordered domain, and they are required to map into a list of non-overlapping intervals. De-identify the sensitive data AES(Advanced Encryption Standard).AES is an iterative rather than Feistel cipher. It is based on 'substitution–permutation network'. It comprises of a series of linked operations, some of which involve replacing inputs by specific outputs(substitutions) and others involve shuffling bits around (permutations).Interestingly, AES performs all its computations on bytes rather than bits. Hence, AES treats the 128 bits of a plaintext block as 16 bytes. For processing as a matrix, these 16 bytes are arranged in four columns and four rows .The number of rounds in AES is variable and depends on the length of the key, unlike DES. For 128-bit keys AES uses 10 rounds, for 192-bit keys 12rounds and for 256-bit keys 14 rounds. In each of these rounds a different 128-bit round key is used and is calculated from the original AES key.

8.3 SKY-FILTER EXTRACTION USING MAP REDUCE

We give the formal definition of recommendation over de-identification policies and denote the problem as RIDP. We propose algorithms using Map Reduce to speed up the parallel computation efficiency and obtain high scalability. Through analyzing the characteristics of data distribution, we give the formal definition of independent property, which can be used to generate new policies effectively. To reduce the sort overhead of skyline and decrease the number of alternative policy set in Map phase of the first round, a new scheme was introduced for recommendation of de-identification policies. We demonstrate the superiority of our methods

through extensive experiments, and the results show that our approach can preserve the privacy substantially with high data utility and query efficiency.

8.3.1 MAP REDUCE:

The Map Reduce algorithm has two important tasks, Map and Reduce .By means of Mapper Class the map task is done and by means of Reducer Class, the reduce task is done. Mapper class takes the input, substitutes any sensitive data, maps and sorts it. The output of Mapper class is used as input by Reducer class, which in turn searches matching pairs and reduces them. Map Reduce implements various mathematical algorithms to separate a task into small parts and assign them to multiple systems. Map Reduce algorithm helps in sending the Map & Reduce tasks to appropriate servers in a cluster, in technical terms.

8.4 QUASI-IDENTIFIER ASSUMPTIONS

De-identification policy stands for the set of domain partitions for each QI attribute. Risk cost stands for the re-identification risk, and utility cost stands for the information loss. If the domain values of QI attributes are provided more specific, the risk cost is becoming larger and utility cost is getting smaller. That is, the smaller is the information loss, the larger is the re-identification risk. Generally speaking, our goal is to minimize the risk and utility cost.

1. **Risk Cost.** The risk cost of a policy stands for the re-identification risk of the data table which is generalized by the policy.
2. **Utility Cost.** The utility cost of a policy stands for the information loss of the data table which is generalized by the policy. And it is measured by the KL-divergence by the generalized data table with respect to its original table.

8.5 STATISTICAL VISUALIZATION

In this process, we used statistical visualization techniques the health care data related to a policy generated is filtered from their list out of main policy will be visualized. When a particular data is selected by policy related that healthcare.

8.5.1 MINING FREQUENT ITEM SETS USING THE APRIORI ALGORITHM

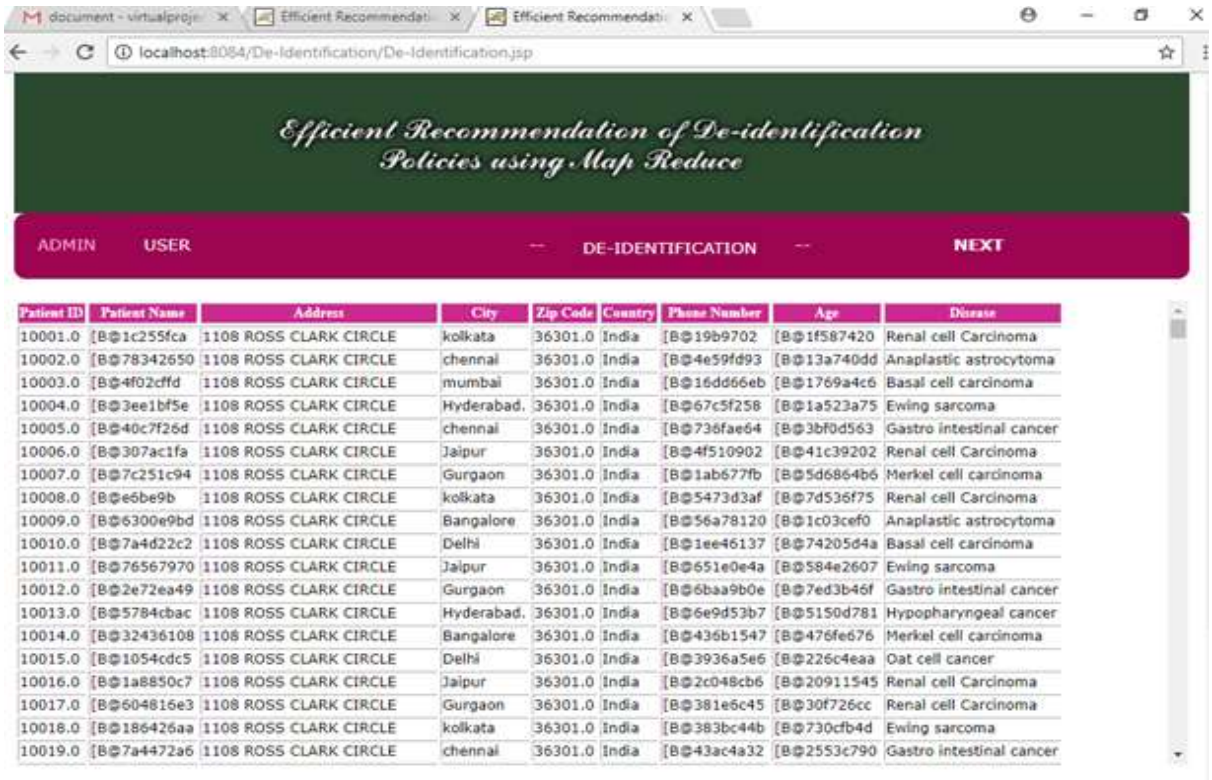
Apriori is an algorithm for discovering frequent item sets in transaction databases. The input is a transaction database (a binary context) and a threshold named min up (a value between 0 and 100 %). A transaction database is a set of transactions. Each transaction is a set of items. For example, consider the following transaction database. It contains 5 transactions (t1, t2... t5) and 5 items (1, 2, 3, 4, 5). For example, the first transaction represents the set of items 1, 3 and 4. This database is provided as the SSSfile contextPasquier99.txt in the SPMF distribution. It is important to note that an item is not allowed appearing twice in the same transaction and those items are assumed to be sorted by lexicographical order in a transaction.

8.5.2 JFREECHART-BARCHART

A bar chart uses different orientation (horizontal or vertical) bars to show comparisons in various categories. One axis (domain axis) of the chart shows the specific domain being compared, and the other axis (range axis) represents discrete values

IX. RESULT AND DISCUSSION

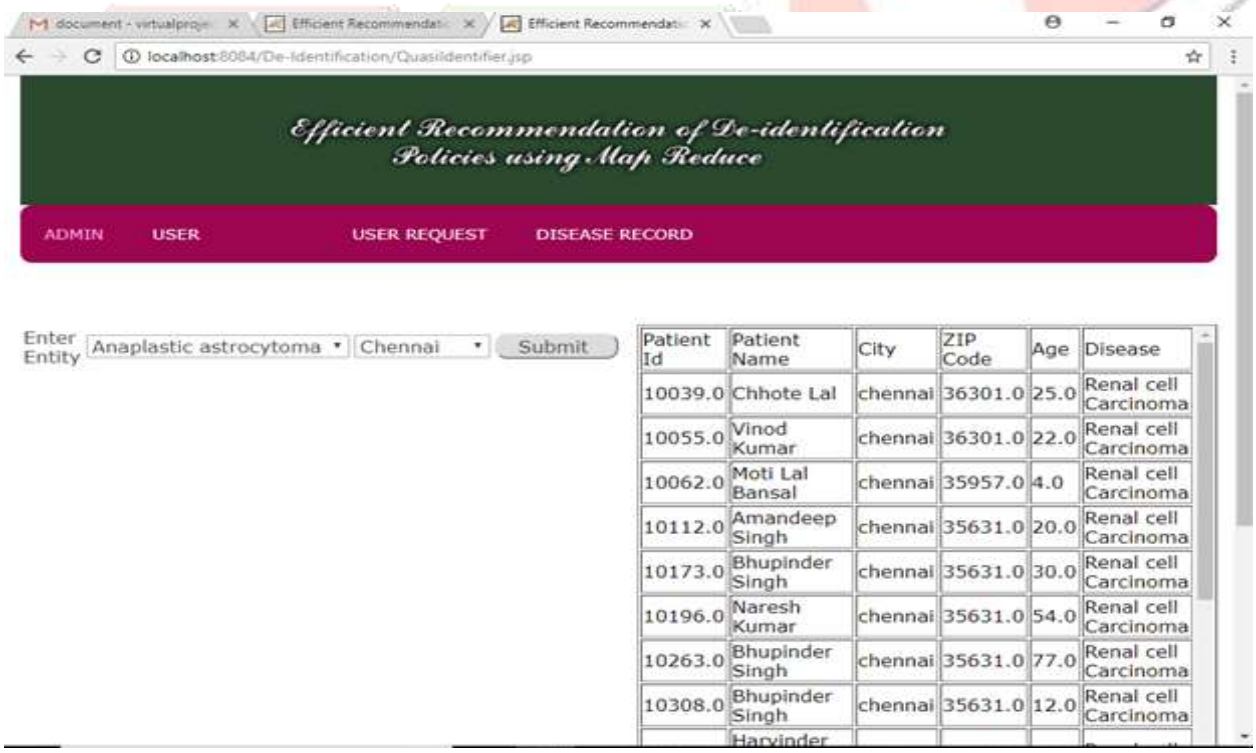
Choose the file and Upload the dataset.
Extract the data from appropriate dataset location.
De-identify the data set using AES techniques.



Patient ID	Patient Name	Address	City	Zip Code	Country	Phone Number	Age	Disease
10001.0	[B@1c255fca	1108 ROSS CLARK CIRCLE	kolkata	36301.0	India	[B@19b9702	[B@1f587420	Renal cell Carcinoma
10002.0	[B@78342650	1108 ROSS CLARK CIRCLE	chennai	36301.0	India	[B@4e59fd93	[B@13a740dd	Anaplastic astrocytoma
10003.0	[B@4f02cffd	1108 ROSS CLARK CIRCLE	mumbai	36301.0	India	[B@16dd66eb	[B@1769a4c6	Basal cell carcinoma
10004.0	[B@3ee1bf5e	1108 ROSS CLARK CIRCLE	Hyderabad.	36301.0	India	[B@67c5f258	[B@1a523a75	Ewing sarcoma
10005.0	[B@40c7f26d	1108 ROSS CLARK CIRCLE	chennai	36301.0	India	[B@736fae64	[B@3bf0d563	Gastro intestinal cancer
10006.0	[B@307ac1fa	1108 ROSS CLARK CIRCLE	Jaipur	36301.0	India	[B@4f510902	[B@41c39202	Renal cell Carcinoma
10007.0	[B@7c251c94	1108 ROSS CLARK CIRCLE	Gurgaon	36301.0	India	[B@1ab677fb	[B@5d6864b6	Merkel cell carcinoma
10008.0	[B@e6be9b	1108 ROSS CLARK CIRCLE	kolkata	36301.0	India	[B@5473d3af	[B@7d536f75	Renal cell Carcinoma
10009.0	[B@6300e9bd	1108 ROSS CLARK CIRCLE	Bangalore	36301.0	India	[B@56a78120	[B@1c03cef0	Anaplastic astrocytoma
10010.0	[B@7a4d22c2	1108 ROSS CLARK CIRCLE	Delhi	36301.0	India	[B@1ee46137	[B@74205d4a	Basal cell carcinoma
10011.0	[B@76567970	1108 ROSS CLARK CIRCLE	Jaipur	36301.0	India	[B@651e0e4a	[B@584e2607	Ewing sarcoma
10012.0	[B@2e72ea49	1108 ROSS CLARK CIRCLE	Gurgaon	36301.0	India	[B@6baa9b0e	[B@7ed3b46f	Gastro intestinal cancer
10013.0	[B@5784cbac	1108 ROSS CLARK CIRCLE	Hyderabad.	36301.0	India	[B@6e9d53b7	[B@5150d781	Hypopharyngeal cancer
10014.0	[B@32436108	1108 ROSS CLARK CIRCLE	Bangalore	36301.0	India	[B@436b1547	[B@476fe676	Merkel cell carcinoma
10015.0	[B@1054cdc5	1108 ROSS CLARK CIRCLE	Delhi	36301.0	India	[B@3936a5e6	[B@226c4eaa	Oat cell cancer
10016.0	[B@1a8850c7	1108 ROSS CLARK CIRCLE	Jaipur	36301.0	India	[B@2c048cb6	[B@20911545	Renal cell Carcinoma
10017.0	[B@604816e3	1108 ROSS CLARK CIRCLE	Gurgaon	36301.0	India	[B@381e6c45	[B@30f726cc	Renal cell Carcinoma
10018.0	[B@186426aa	1108 ROSS CLARK CIRCLE	kolkata	36301.0	India	[B@383bc44b	[B@730cfb4d	Ewing sarcoma
10019.0	[B@7a4472a6	1108 ROSS CLARK CIRCLE	chennai	36301.0	India	[B@43ac4a32	[B@2553c790	Gastro intestinal cancer

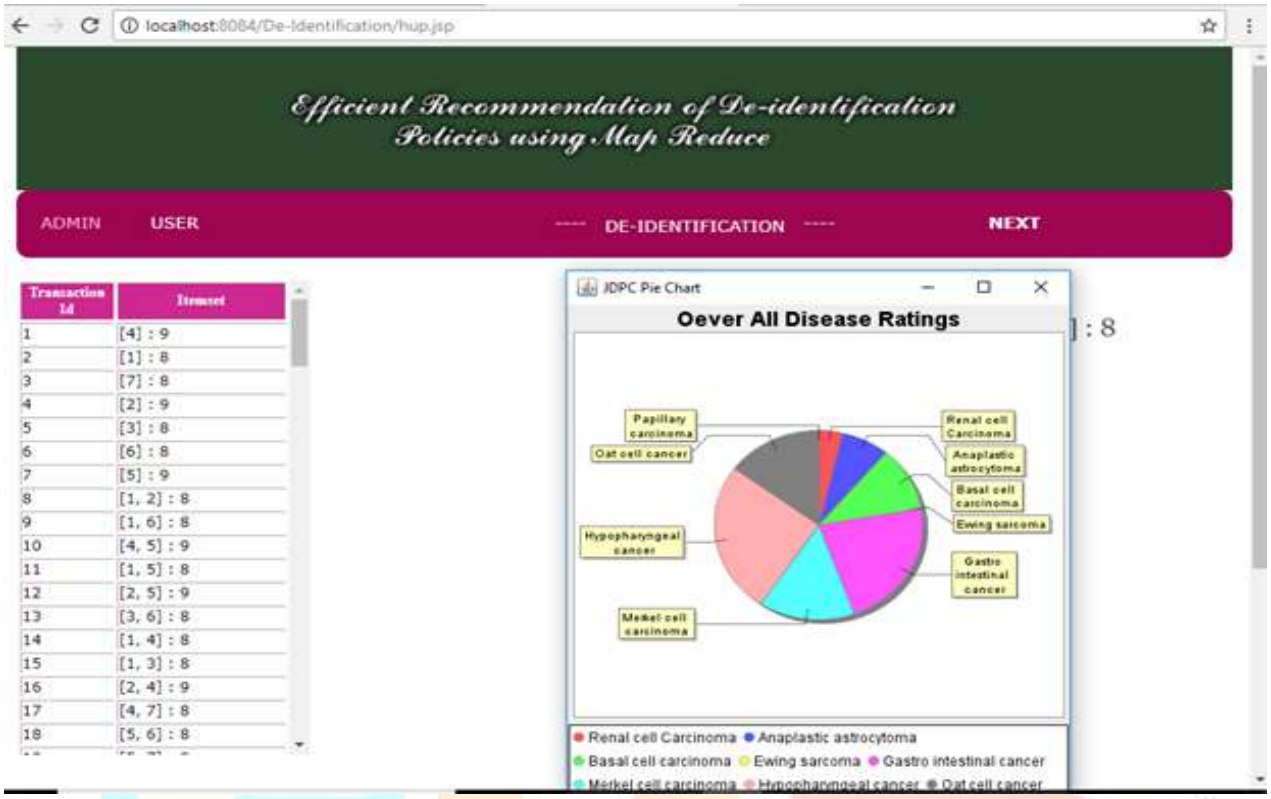
Quasi-identifier assumptions.

Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.



Patient Id	Patient Name	City	ZIP Code	Age	Disease
10039.0	Chhote Lal	chennai	36301.0	25.0	Renal cell Carcinoma
10055.0	Vinod Kumar	chennai	36301.0	22.0	Renal cell Carcinoma
10062.0	Moti Lal Bansal	chennai	35957.0	4.0	Renal cell Carcinoma
10112.0	Amandeep Singh	chennai	35631.0	20.0	Renal cell Carcinoma
10173.0	Bhupinder Singh	chennai	35631.0	30.0	Renal cell Carcinoma
10196.0	Naresh Kumar	chennai	35631.0	54.0	Renal cell Carcinoma
10263.0	Bhupinder Singh	chennai	35631.0	77.0	Renal cell Carcinoma
10308.0	Bhupinder Singh	chennai	35631.0	12.0	Renal cell Carcinoma
	Harvinder				

statistical visualization using bar chart.



X. CONCLUSION

Development of data mining technology is only by using high utility pattern. According to the user's expectation there is a limitation in frequent pattern mining. A large number of algorithms available for utility pattern mining which considers mining of frequent item set based on utility considerations. These research papers propose an overview of various existing high utility item set mining algorithms. The reviewed algorithms effectively mining high utility item sets based on the various data structure and constraint techniques. However to discover patterns for large transactional datasets an effective high utility pattern mining algorithm is required for improving the performance and search space of the item sets.

REFERENCES

[1] Ahmed C. F., Tanbeer S. K., Jeong B.-S., and Lee Y. -K., "Efficient tree structures for high utility pattern mining in incremental databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, pp. 1708– 1721, 2009.

[2] Agrawal R., Imielinski T., and Swami A., "Mining association rules between sets of items in large databases," In Special Interest Group on Knowledge Discovery in Data. Association for Computing Machinery, pp. 207–216, 1993.

[3] Anusmitha A., Renjana Ramachandran M., "Utility pattern mining: a concise and lossless representation using up growth", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, No. 7, pp. 451– 457, 2015.

[4] Chun-Wei Lin J., Wensheng Gan., Fournier-Viger P., and Yang L., Liu Q., Frnda J., Sevcik L., Voznak M., " High utility item set-mining and privacy-preserving utility mining," Vol. 7, No. 11, pp. 74–80, 2016.

[5] Dawar S., Goya V. l., "UP - Hist tree: An efficient data structure for mining high utility patterns from transaction databases," In Proceedings of the 19th International Database Engineering & Applications Symposium. Association for Computing Machinery, pp. 56–61, 2015.

[6] De Bie T., "Maximum entropy models and subjective interestingness: an application to tiles in binary databases," Data Mining and Knowledge Discovery, Vol. 23, No. 3, pp. 407–446, 2011.

[7] Erwin A., Gopalan R. P and. Achuthan N. R., "Efficient mining of high utility item sets from large datasets," In Proceeding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 554–561, 2008.

[8] Fournier-Viger P., Wu C.-W., Zida S., and Tseng V.S., "Fhm: Faster high-utility item set mining using estimated utility Cooccurrence pruning," In Proceedings of the 21th International Symposium on Methodologies for Intelligent Systems. Springer, pp.83-92, 2014.

[9] Geng L., Hamilton H.J, "Interestingness measures for data mining: A survey," Association for Computing Machinery. Vol. 38, No. 3, pp.1–9, 2006.

[10] Han J., Pei J., Yin Y., Mao R., "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," Data Mining Knowledge Discovery in Data.Vol. 8, No. 1, pp. 53–87, 2004.