Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT ALGORITHM FOR PRIVACY PRESERVING TEMPORAL PATTERN MINING

[1]S.S.ARUMUGAM, [2]Dr.V.PALANISAMY
[1]Assistant Professor, Sri Lakshmi Ammal Engineering College, Chennai
[2]Principal, Info Institute of Engineering, Coimbatore
E-mail:[1]ssarumugam.me@gmail.com, arumugam0879@gmail.com

ABSTRACT

Pattern mining increases extra awareness due to its practical relevance in many fields, such as, biology, medicine and so on. Recently, a few researches have been presented in the literature about the mining of privacy preserving patterns. There are more number of data mining techniques developed to find out important information, yet the way of mining the data from the database is not concerned about the privacy. In this paper, we have selected the medical database and created the privacy on the patient's medical database while the hospitals share their data to other organizations. In case of sharing the medical data, preserving the privacy of the data is also of importance as it may lead to violation of individual (patient) highly personal information. Based on the organizational need, the hospital can make the privacy on particular datum alone, say on type diseases. In that respective way, a choice is given to the hospital to where to make the privacy and this may vary depending on the organization. Here, we have made the privacy on four different types of diseases grouped as sensitive diseases, frequent diseases, seasonal diseases and geographical diseases. Privacy of the patient records are obtained by modifying the original database. Here, Prefix span algorithm was used for mining important diseases from the database. The result of mining was not worthy comparing the result with the original medical database. Finally the experimentation was made with the synthetic dataset and an analyzes on the difference among the four types of diseases while make the privacy in terms of the execution time, memory usage, was done.

Keywords: *Privacy Preserving, Medical Database, Prefix Span, Sensitive Diseases, Frequent Diseases, Seasonal Diseases And Geographical Diseases.*

1. INTRODUCTION

Based on numerous research works it has been found that, data mining concept generally deals with the extraction of potentially useful information from large collections of data with a variety of functional areas such as customer relationship management, market basket analysis, and bio-informatics [1]. Data mining can be used for predicting and analyzing the medical records of hospitals in a town, for example, potential outbreaks of infectious diseases, analysis of customer transactions for market research applications etc. The list of application areas for data mining is large and is bound to grow rapidly in the years [2]. Data mining can be a complex and difficult process. For example, data mining must be able to handle different types of data as data does not always exist in textual format. Secondly, data mining algorithms must be able to handle data in an efficient and scalable manner. Data mining algorithms must be able to produce an accurate representation of the data in the form of a model regardless of the size of the dataset. Thirdly, data mining must handle noisy or missing data within a dataset and still be able to produce an accurate representation of the data in the form of a model. Next, end users must be able to perform data mining tasks without having an extensive knowledge of data mining algorithms. Data quality is an important aspect of data mining. High quality data that has been prepared specifically for data mining tasks will result in useful data mining models and output. Data which is inaccurate, incomplete, insecure, ambiguous, or outdated may be considered as a low quality data [3].

Currently, for effective sharing of medical data it is essential to promote the collaboration within the health care community and with other parties such as research institutes, pharmaceutical and insurance companies, so as to enhance the quality and efficacy of health care provision. For example, a hospital may need to outsource clinical records in its independent databases to a research institute in an attempt to find out a new drug or evaluate a new

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

therapy. The widespread applications of medical data could also be used to satisfy legal requirements [4]. In recent years the data mining community has faced a new challenge which has shown how effective its tools are in revealing the knowledge locked within huge databases, it is now required to develop methods that restrain the power of these tools to protect the privacy of individuals [5]. The objective of data mining is to generalize across populations, rather than reveal information about individuals. The drawback of data mining is that it works by evaluating individual data that is subject to privacy concerns. So, the true problem is not data mining, but the way data mining is done. For receiving an appropriate medical care a patient may provide highly personal information to a health care provider.The hospitals need to maintain the privacy on the patient medical database because the patients give their own data to the hospital to get the good treatment from them. The direct release of medical data invariably violates individual privacy. In order to overcome this problem [6] and [7] proposed the concept of privacy preserving data mining (PPDM) aimed at alleviating the conflict between data mining and privacy. As a first introduction of PPDM, the idea of [8] is to agitate individual data values [8].

The need to share the data arises when many organizations outsource their management responsibility. The process of contracting an outside company for providing a service which is previously performed by a staff is known as outsourcing. Outsourcing involves a transfer of management responsibility for delivery of service and internal staffing patterns to an outside organization. The healthcare organization outsources a variety of activities and, second, that the major benefits of using outside services are improved performance, cost savings, and increased management time in core business [9]. Drug companies come up with new medicines and decisions by studying and analyzing the medical history of the various patients that is got from the hospitals.

This may pave way for fraudulent activities happening in the drug companies. In order to reduce the fraud activities of the drug company, the hospitals need to make the privacy on their database. The Health Information Portability and Accountability Act (HIPAA) of 1996 protect individually identifiable health information. The privacy and security of medical health information are protected by means of certain standards which are established by HIPAA. In case of the electronic exchange the act also requires security mechanisms which are to be used in the electronic exchange for individually identified health information.

In this paper, the privacy is made on the medical database, the data owner is the hospital. The hospitals may share their medical database to any other organizations like drug manufacturing companies and medical insurance companies or to any other hospitals. The hospitals need to maintain the privacy on the patient medical database because the patients give their own data to the hospital to get the good treatment from them.

The privacy pattern on the medical data of the patient are customized according to the organizational need. To provide the choice was given to data owner on which part to be anonymize. Here, the privacy was made on patient's name and disease's name by ASCII code and on the privacy on sensitive disease, frequent disease, seasonal disease and geographical diseases. After which the prefix span algorithm was used to mine the importance disease. The result of the mining did not affect the privacy of the patient.

2. RELATED WORKS

Keng-Pei Lin *et al.[2010]*,have proposed the privacy-preserving outsourcing support vector machines with random transformation [12]. In this paper, they presented a strategy for privacy-preserving outsourcing, the training of the SVM without exposing the actual content of the data to the service provider. In the proposed system, the data sent to the service provider was troubled by a random transformation, and the service provider trained the SVM for the data owner from the troubled data. The proposed plan was stronger in security as compared to the existing techniques, and obtained very few redundancies in communication and computation cost.

Jieh-Shan *et al.[2010]*, have discovered HHUIF and MSICF algorithms for privacy preserving utility mining [13]. Their study focused on privacy preserving utility mining (PPUM) and have also presented two algorithms, HHUIF and MSICF, attained the goal of hiding sensitive item sets so that the adversary cannot mine them from the modified database. The work also minimized the impact on the sanitized data base of hiding sensitive item sets. They showed that HHUIF attains lower miss costs than MSICF on two synthetic datasets. On the other hand, MSICF generally had a lower difference ratio than HHUIF between original and sanitized database.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Maryam Khan *et al.[2010]*, has presented medical tourism: outsourcing of healthcare [14]. Their study examined whether the growth in medical tourism will eventually have a result in the outsourcing of U.S. healthcare services. They showed that as long as people in developed countries lack affordable healthcare, medical tourism continues to grow. There were already an outsourcing of manufacturing, technology and service related jobs. The U.S. healthcare maintained its status quotient so that the healthcare services may also be outsourced.

Yonghong YU *et al.[2010]*, have proposed the integrated privacy protection and access control over outsourced database services [15]. Moreover, a solution to enforce data confidentiality has also been explained, data privacy, user privacy and access control over outsourced database services. They started from a flexible definition of privacy constraints on a relational outline, applied encryption on information in a parsimonious way and mostly rely on attribute partition to protect sensitive information. Their approximation algorithm for the minimal encryption attribute partitioned with quasi-identifier detection, they allowed storing the outsourced data on a single database server and minimized the amount of data represented in encrypted format. Here, they applied cryptographic technology on the auxiliary random server protocol that can solve the problem of private information retrieval to protect data privacy, user privacy and access control for outsourced database services. Their analysis showed that our new model can provide efficient data privacy protection and query processing, which is efficient in computational complexity without increasing the cost of communication complexity of user privacy protection and access control.

Ken Barker *et al.[2009]*, gave the idea of data privacy taxonomy [16]. They offered an explicit definition of data privacy which was suitable for ongoing work in data deposits such as, a DBMS or data mining. Their work contributed by briefly providing the larger context for the way privacy was defined legally and legislatively but primarily providing taxonomy that is capable of thinking of data privacy technologically. They demonstrated the taxonomy's utility by illustrating how this perspective makes it possible to understand the important contribution made by researchers to the issue of privacy. They declared privacy was indeed multifaceted so no single current research effort adequately addressed the true breadth of the issues necessary for fully understanding the scope of this important issue.

Jingquan Li *et al.[2012]*,provided a technique for safe guarding the privacy of electronic medical records [17]. They developed a formal privacy policy for safeguarding the privacy of EMRs. They described the impact of EMRs and HIPAA on patient privacy. They proposed access control and audit logs policies to protect patient privacy. To illustrate the best practice in the healthcare industry, they presented the case of the University of Texas M. D. Anderson Cancer Center. Where it has been demonstrated that it was critical for a healthcare organization to a formal privacy policy in place.

3. PROBLEM DESCRIPTION

The medical database(DB) consist of patient' s name, disease's name D_i, time duration (t_s, t_e) of the disease like starting time and ending time of the disease D_i. To create the needed segregation of the different diseases we need such data. The sensitive diseases has long time duration, the maximum duration of the disease is shown as $S_n t$ as a sensitive threshold got from the user. The difference between the starting time and ending time of the disease is denoted as $diff\ of\ D_i$ if this value is greater than $diff\ of\ D_i > S_n t$ threshold then it should be subtracted($S_n t$ values) from the $diff\ of\ D_i$,this process repeats until the value of $diff\ of\ D_i$ less than sensitive threshold value and make the changes in the original database named it as sanitized database. This sanitized database is sent by the data owner to another organization from this sanitized database they can't mine the sensitive data. In order to find the frequent disease of each disease was counted from the database $D_i(CNT)$ and the maximum count value was got from the user denoted as $C_n t$. If the value of the $D_i(CNT) > C_n t$ was removed, the corresponds disease was sort from the database. This process repeats until the above condition was not satisfied. After removing the frequent disease from the original database then it was named as sanitized database. The seasonal diseases were found with the help of the sliding window method, the window size is normally 30, for every 30 days the diseases were mined from the database $D_i(CNT)$ and it moves to the window table WT. Ratio between counts of the disease on the window

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

table $D_i(CNT)_{WT}$ to the size of the window table was $S(WT)$. From this result the seasonal disease was arrived by set of the threshold value. Finally the prefix span output was checked with the original database on how much was achieved from the privacy on such diseases.

4. PROPOSED APPROACH OF EFFICIENT ALGORITHMS FOR PRIVACY PRESERVING TEMPORAL PATTERN MINING

The second type of problem that is likely to be faced while sharing data could arise, while sharing two different data for two different outsourcing group, say, drug company and insurance group. There are so many drug manufacturing companies presented, they need the medical data to research on the existing medical condition of the patients since the privacy of the patient become affected and so also the medical insurance companies withdraw their offers for some people who are affected by particular diseases thus privacy of the patient gets affected. A patient offers extremely personal information to the hospital to make certain they get a suitable medical diagnosis and treatments. Hospitals need to make the privacy on the patient diseases while sharing the database to any organizations.

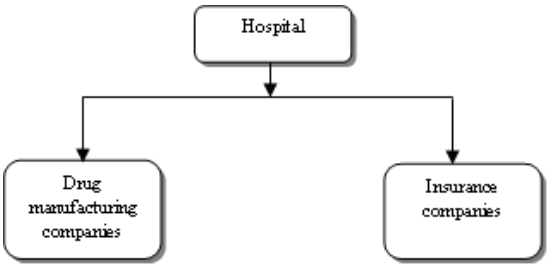

Figure: 1 Shows That The Sharing Of Hospital Data To Other Organizations

The data owner may share their data to any organization, according to the organization, the data owner may make the privacy on the database. For example the drug manufacturing companies need to mine the frequent diseases from the shared database and the insurance companies need to mine the sensitive diseases from the database since the privacy of the each patient has different base on the organizations. By considering the above motivations, in this paper, the choice was given to data owner where to make the privacy of the database. In this paper, the diseases were separated into four types like

1. Sensitive diseases
2. Frequent diseases
3. Seasonal diseases
4. Geographical diseases

Based on the organization, the hospital (data owner) anonymizes and makes the privacy on those diseases. The data owner may choose any kind of the diseases as mentioned above and they can make the privacy on that disease by reducing the sensitivity or the count of the diseases. The following Table:1 shows that the medical data of the data owner. The database DB consists of the patient's name, disease's name and time duration as starting time t_s and ending time t_e of the disease. The diseases are classified by the time duration and the count of the diseases. By changing the time duration and the count of the disease we can make the privacy on that. In order to make the privacy on patient's name and disease's name of the database, ASCII code is used to convert the patient's name and disease's name into the numeric format.

Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table: 1 Describes The Database Of The Medical Data Of The Patients

Patient name	Disease name	Place	Duration	Patient name	Disease name	Place	Duration
David	*Biopsy*	North	05-05-09 -- 09-05-09	*David*	*Autism*	North	25-08-09 – 30-08-09
David	*Cheilosis*	North	13-01-09 – 21-02-09	*Suman*	*Epilepsy*	East	02-08-09 – 9-08-09
David	*Dysphagia*	North	12-05-09 – 16-05-09	*Suman*	*Autism*	East	04-08-09 - 09-08-09
David	*Autism*	North	03-08-09 -- 09-08-09	*Suman*	*Dysphagia*	East	18-02-09 – 25-02-09
David	*Epilepsy*	North	02-02-09 – 10-02-09	*Suman*	*Biopsy*	East	06-05-09 - 09-05-09
David	*Glaucoma*	North	18-08-09 – 27-08-09	*Suman*	*Dysphagia*	East	02-08-09 – 12-08-09
David	*Felon*	North	06-06-09 –14-06-09	*Suman*	*Hernia*	East	30-07-09 - 08-09-09
Patient name	**Disease name**	**Place**	**Duration**	**Patient name**	**Disease name**	**Place**	**Duration**
Peter	*Cheilosis*	South	21-09-09 – 25-10-09	*Peter*	*Epilepsy*	South	04-06-09 – 13-06-09
Peter	*Biopsy*	South	03-05-09 -- 08-05-09	*Lee*	*Biopsy*	West	04-05-09 - 09-05-09
Peter	*Autism*	South	05-08-09 -- 09-08-09	*Lee*	*Epilepsy*	West	13-03-09 – 21-03-09
Peter	*Felon*	South	21-12-09 – 30-12-09	*Lee*	*Dysphagia*	West	21-11-09 – 30-11-09
Peter	*Hernia*	South	21-03-09 -- 08-05-09	*Lee*	*Autism*	West	05-08-09 -- 09-08-09
Peter	*Glaucoma*	South	09-09-09 – 17-09-09	*Lee*	*Epilepsy*	West	19-11-09 – 26-11-09
Peter	*Epilepsy*	South	05-04-09 – 11-04-09	*Lee*	*Felon*	West	04-07-09 – 11-07-09

4.1 Making Privacy on Sensitive Disease

In order to make the privacy on the sensitive database, the sensitive data were first sorted out then a modification on the sensitivity (time duration) of those diseases to change it into the non-sensitive disease was done. Each disease has significant time duration, the time duration here refers to the starting time and ending time of the disease. The sensitive diseases were obtained by the indication of long time duration of disease identified by the user defined sensitive threshold value, the threshold value represents the maximum time duration of the disease, the modification of the sensitivity of the diseases was repeated until the sensitivity of the diseases were below or equal to the threshold value. At the end of this process, the database had extracted the data on the sensitive disease.

Pseudo code
Input: original database DB, sensitive threshold S_nt
Output: sanitized database SDB

1. Read DB
2. calculate $diff\ of\ D_i = (t_e - t_s)D_i$
3. Get the value of S_nt
4. If $diff\ of\ D_i > S_nt$
5. $diff\ of\ D_i - S_nt$
6. Go to step 4
7. Else
8. No change in duration
9. Update SDB
10. Return SDB

Example: To find the value of the sensitive disease from the database, first the system gets the sensitive threshold value S_nt from the user, after getting the value of threshold from the user, the system fetches the diseases which has the time duration greater than sensitive threshold value S_nt from the database. For example consider the database as given in the above table 1 and the threshold value is 15, now the system finds the sensitive diseases as tabulated below

Table2: Shows That The Count Value Of The Sensitive Diseases.

Patient ID	Disease ID	Count value
6897118105100	*671041011051081111115105115*	39
8311710997110	*7210111411010597*	40
80101116101114	*671041011051081111115105115*	34
80101116101114	*7210111411010597*	48

The privacy of the sensitive disease is obtained through the modification of the count value of the disease; the modification process is repeated until the count value of the disease less than the value of sensitive threshold value. The modification is done by subtracting the sensitivity (count value) of the disease from the sensitive threshold value, the subtraction process is repeated until the sensitivity value of the disease is less than the sensitivity count of the disease. After that the modifications, the count values are updated and stored in different database called sanitized database. From this sanitized database the other organizations feel very difficult to find or mine the sensitive diseases from that since this sanitized database is shared to other organizations by the hospital management.

Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2 Making Privacy on Frequent Disease

In order to make the privacy on the frequent disease in the database, the frequent disease were first sorted out. Here, the frequent diseases were arrived at identifying the repetitive diseases i.e. the diseases that have more number of counts in the database. The frequent diseases were filtered from the database by getting the count threshold value C_nt from the user. After finding the frequent diseases they were anonymized by modifying the count of those diseases to become normal ones. Then for the modification process, the count value of frequent diseases was subtracted from the count threshold value C_nt. Thus reducing process repeats until the count value of the frequent diseases were below than the count threshold C_nt value.

Example: To find the value of the frequent diseases from the database, first the system gets the count threshold C_nt value from the user, after that, the system fetches the diseases that have the count value greater than the threshold value $D_i(CNT) > C_nt$ from the database. Consider the database as given in the above table 1 and the count threshold value C_nt is 4 now the disease id 69112105108101112115121 has the count value $D_i(CNT)$= 6. The disease 69112105108101112115121 is a frequent disease from the database since the only disease that has the count value more than the count threshold value. In order to anonymize frequent disease, the count value of the disease $D_i(CNT) = 6$ is subtracted by the count threshold $C_nt = 4$. Remove the disease D_i up to C_nt times after that check the count value, if it is greater than the count threshold value then remove the disease again C_nt else move the diseases into the sanitized database SDB.

Pseudo code
Input: original database DB, disease count threshold C_nt
Output: sanitized database SDB

1. Read DB s and calculate the count of each disease $D_i(CNT)$
2. Get the value of C_nt
3. If $D_i(CNT) > C_nt$
4. $D_i(CNT) - C_nt$
5. Remove D_i up to C_nt times
6. Go to step 3
7. Else
8. No change in disease count
9. Update SDB
10. Return SDB

4.3 Making Privacy on Seasonal Disease

In order to make the privacy on the seasonal disease from the database, the seasonal disease were identified first. The seasonal diseases have more number of counts in a particular time period in the database. Sliding window method was used to identify seasonal diseases. Normally the window size is 30 for every 30 days the diseases were sorted out and the count value was calculated also. Then the count threshold C_nt value was got from the user to find the frequent diseases from that. For every window the frequent diseases were selected with the help of count threshold value and plotted it into window table. Each disease in the window table was counted and divided by the size of the window table. By setting the threshold value the seasonal diseases were got. Threshold value was given by the user. After finding the seasonal diseases the privacy was made by reducing the count of seasonal disease to become the normal disease. In order to reduce the count of that disease it was removed from the database. The removing of the disease process was repeated until the count value was below the count threshold value.

Example: To find the value of the seasonal diseases from the database, first the system finds the sensitive disease from which the sliding window method finds the seasonal diseases. To elaborate, every 30 days sliding window calculates the disease and its counts are $D_i(CNT)$, based on value the count threshold value C_nt the selected diseases are moved to window table. From the window table the seasonal diseases can find out by the ratio of the count of the disease in the window table to the size of the window table. Get the seasonal threshold S_nt value from the user, if the seasonal diseases are greater than the S_nt then remove the diseases from the database and make the sanitized database. This sanitized database does not have any seasonal database.

Pseudo code
Input: original database DB, size of the window W and the count of the disease C_nt seasonal count S_nt

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Output: sanitized database SDB

1. Read DB for every W and calculate the count of each disease $D_i(CNT)$
2. Get the value of C_nt
3. If $D_i(CNT) > C_nt$
4. D_i Go to the window table
5. calculate the disease count in the window table $D_i(CNT)_{WT}$
6. calculate the size of window table $S(W_T)$
7. find $D_i(WT) = \frac{D_i(CNT)_{WT}}{S(W_T)}$
8. if $D_i(WT) > S_nt$
9. $D_i(CNT) - C_nt$
10. Remove D_i up to C_nt times
11. Go to step 8
12. Return the SDB

Sanitized database of sensitive disease

4.4 Making Privacy On Geographical Diseases

With the aim of making the privacy on the geographical disease, we are calculating the following things, initially we find out the geographical diseases from the medical database and make the privacy on the geographical disease. In order to find the geographical diseases, first we find the frequent diseases from total area (whole database) subsequently we discover the frequent disease in the particular area. With the help of the two frequent diseases, we can calculate the geographical disease. After getting the geographical disease, in order to make the privacy on geographical disease we reduce the count of the geographical disease with the help of user defined threshold value to become a normal disease.

Example: Consider there are n number of areas a is presented in the medical database and each area has some number of diseases. To find the frequent disease, first we get the threshold C_nt value from the user after that we calculate count for the area disease from all area $D_i(CNT)$. If the count of a disease is greater than threshold value $D_i(CNT) > C_nt$ then the corresponding

Disease has considered as frequent disease the result of this condition we get set of frequent diseases like $(FD)\forall a = \{D_1, D_2, .. D_n\}$. The next process is to find the frequent disease in the particular area. Each of the area has consists set of frequent diseases $(FD)a_i = \{D_1, D_2, .. D_n\}$. The diseases in the unique area are said to be a geographical disease whenever the following condition satisfied $(GD)a_i = (FD)a_i \setminus (FD)\forall a$ .This calculation leads to get the set of geographical diseases $(GD)a_i = \{D_1, D_2, .. D_n\}$ from which we are reducing the privacy by reduce the count of the geographical disease with the help of user defined count value $GD_i(CNT) - (C_nt)a_i$. This process is repeated until the count value of the geographical disease becomes less than a user defined threshold value; this leads to become the geographical disease into normal disease.

Pseudo code
Input: original database DB, count value of all area $(C_nt)\forall a$, count value of particular area $(C_nt)a_i$
Output: sanitized database SDB

1. Read DB and calculate the count of each disease $D_i(CNT)\forall a$
2. Get the value of $(C_nt)\forall a$
3. If $D_i(CNT)\forall a > (C_nt)\forall a$
4. D_i moves to $(FD)\forall a = \{D_1, D_2, .. D_n\}$
5. Get the value of $(C_nt)a_i$
6. If $D_i(CNT)a_i > (C_nt)a_i$
7. D_i moves to $(FD)\ a_i = \{D_1, D_2, .. D_n\}$
8. Calculate $(GD)a_i = (FD)a_i \setminus (FD)\forall a$
9. The geographical diseases are $(GD)a_i = \{D_1, D_2, .. D_n\}$
10. If $GD_i(CNT) > (C_nt)a_i$
11. $GD_i(CNT) - (C_nt)a_i$
12. Remove GD_i up to $(C_nt)a_i$ times from area a_i
13. Go to step 10
14. Else
15. No change in disease count
16. Update SDB
17. Return SDB

Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. MINING SEQUENTIAL PATTERNS BY PREFIX PROJECTIONS

In this section, pattern growth technique was used for mining sequential patterns prefixspan algorithm to mine the important patterns from the databases. The hospitals only shares the sanitized database to the organization, the organization wants to mine the importance diseases from the sanitized database.

Table: 3 Shows That The Sanitized Database Of The Sensitive Disease

Patient ID	Disease	place	Duration	Patient ID	Disease	Place	Duration
6897118105100	*66105111112115121*	110111114116104	05-05-09 -- 09-05-09	*6897118105100*	*65117116105115109*	11011111411610 4	25-08-09 – 30-08-09
6897118105100	*67104101105108111115105115*	110111114116104	13-01-09 – 22-01-09	*8311710997110*	*69112105108101112115121*	10197115116	02-08-09 – 9-08-09
6897118105100	*68121115112104 9710310597*	110111114116104	12-05-09 – 16-05-09	*8311710997110*	*65117116105115109*	10197115116	04-08-09 - 09-08-09
6897118105100	*65117116105115109*	110111114116104	03-08-09 -- 09-08-09	*8311710997110*	*68121115112104 9710310597*	10197115116	18-02-09 – 25-02-09
6897118105100	*69112105108101112115121*	110111114116104	02-02-09 – 10-02-09	*8311710997110*	*66105111112115121*	10197115116	06-05-09 - 09-05-09
6897118105100	*71108971179911110997*	110111114116104	18-08-09 – 27-08-09	*8311710997110*	*68121115112104 9710310597*	10197115116	02-08-09 – 12-08-09
6897118105100	*70101108111110*	110111114116104	06-06-09 –14-06-09	*8311710997110*	*72101114110105 97*	10197115116	30-07-09 - 09-08-09
Patient ID	Disease	Place	Duration	Patient ID	Disease	Place	Duration
8010111610111 4	*67104101105108111115105115*	115111117116104	21-09-09 – 25-09-09	*8010111610111 4*	*69112105108101112115121*	11511111711610 4	04-06-09 – 13-06-09
8010111610111 4	*66105111112115121*	115111117116104	03-05-09 -- 08-05-09	*76101101*	*66105111112115121*	119101115116	04-05-09 - 09-05-09
8010111610111 4	*65117116105115109*	115111117116104	05-08-09 -- 09-08-09	*76101101*	*69112105108101112115121*	119101115116	13-03-09 – 21-03-09
8010111610111 4	*70101108111110*	115111117116104	21-12-09 – 30-12-09	*76101101*	*68121115112104 9710310597*	119101115116	21- 11-09 – 30-11-09
8010111610111 4	*72101114110105 97*	115111117116104	21-03-09 -- 24-03-09	*76101101*	*65117116105115109*	119101115116	05-08-09 -- 09-08-09
8010111610111 4	*71108971179911110997*	115111117116104	09-09-09 – 17-09-09	*76101101*	*69112105108101112115121*	119101115116	19-11-09 – 26-11-09
8010111610111 4	*69112105108101112115121*	115111117116104	05-04-09 – 11-04-09	*76101101*	*70101108111110*	119101115116	04-07-09 - 11-07-09

5.1 Make the Sequential Disease Based On Time

Each of the patients has different diseases in different ways with difference since the sequential of the diseases are differs from patients to patients. The diseases are in ascending order based on the time of occurrence. Consider that the above table consists of patient id and disease id and duration of the disease, based on that the diseases were ordered sequentially. For example the patient *6897118105100* has the sequential disease based on the time *671040101105108111115105115* → *69112105108101112115121* → *66105111112115121* → *68121115112104 9710310597* → *70101108111110* → *65117116105115109* → *71108971179911110997* → *65117116105115109.*

5.2 Projected Database

The projected database *PDB* consists of set of sequential diseases δ. Consider the diseases β is a disease from the sequential disease δ then the projected database *PDB* consists of postfix of the sequential disease with respect to the disease β. Consider the sequential disease of patient *6897118105100* and let the value of β be *68121115112104 9710310597* then the postfix of the sequential diseases are *70101108111110* → *65117116105115109* → *71108971179911110997* → *6511711610511510* . These are the diseases placed in the projected database *PDB* from that we can find the frequent sequential diseases with the help of the various levels of support counts.

5.3 Support Counts in Projected Database

Support counts *SC* are used to find the subset of sequence δ_{sub} from the projected database *PDB*. For each support count *SC* the subset sequences δ_{sub} are different. The projected database consists of postfix diseases of the selected prefix disease. On counting the diseases of the projected database if the diseases had the count value greater than the support count then that diseases were moved to subset sequences. The subset sequences

Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

were combined with the prefix disease for next projection. This process was repeated until the count of the subset sequence became single. The result of the prefixspan algorithm is *691121051081011121 12115121s3→691121051081011 12115121s4→ 681211151121049710310597s2→ 70101108111110s1→65117116105115109s1→661 05111112115121s1→71108971179911110997s1→ 681211151121049710310597s3→65117116105115 109s2→7210111411010597s1*

Pseudo code for prefixspan
Input: sanitized database
Output: frequent sequential database
1. Make the $D_{seq} \forall P_i$
2. Select $P_{re}D_i$ from the (D_{seq})
3. Select $P_{st}D_i$
4. Calculate $C(P_{st}D_i)$
5. Get the value of SC
6. If $C(P_{st}D_i) > SC$
7. $D_i \rightarrow S_{seq}$
8. $P_{re}D_i + S_{seq}$
9. If $C(S_{seq}) > 1$
10. Else
11. Go to step 2

6. RESULTS AND DISCUSSION

The experimental results of the proposed efficient algorithm for privacy preserving temporal pattern mining isdescribed in this section. The comparative analysis of the privacy preserving with seasonal diseases, sensitive diseases, frequent diseases and geographical diseases is presented for synthetic datasets.

6.1 Experimental Design

The proposed approach for making the privacy on different diseases is programmed using Java (jdk 1.6). The experimentation has been carried out using the synthetic datasets as well as the real datasets with dual core processor PC machine with 2 GB main memory running a 32-bit version of Windows XP. We have generated the synthetic data that comprises of three attributes like patient's name, disease's name and duration of the disease.

6.2 Evaluation Metrics

The performance of the proposed event efficient algorithms for privacy preserving temporal pattern mining was evaluated by means of three evaluation measures. They are: 1) "Changing of original database"- the original database converts into sanitized database for sharing other organizations. How many changes are made for each type of diseases.2) Running time- the time taken to execute the algorithm and it typically grows with the input size and 3) Memory usage- the memory utilized by the algorithm to convert the original database into sanitized database.

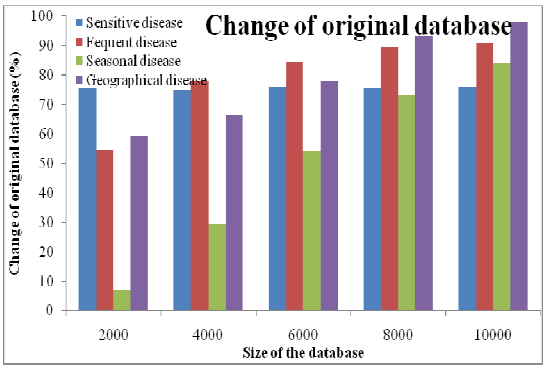

Figure: 2 Shows That The Percentage Of Changing The Original Database

The medical database consists details of sensitive diseases, seasonal diseases, frequent diseases and geographical diseases. In order to make the privacy on those diseases, some modifications were needed. Based on the modification the privacy of the data becomes stronger. Here, the original database with the sanitized database were compared, according to that there were three types of diseases modified here to become normal diseases. While comparing the three diseases the frequent disease has projected to be of a very high level because the frequent disease has repeated many times since the modification of each frequent disease was very high. The sensitive disease were less than the frequent diseases since the modification process became less than the frequent diseases. Finally the seasonal diseases that were of very less count in the original database were also very less in the modification of that disease.

Journal of Theoretical and Applied Information Technology
20th December 2013. Vol. 58 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

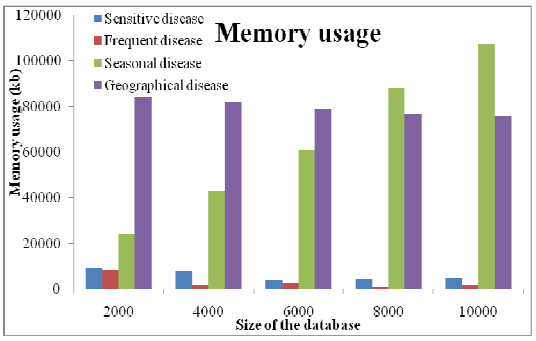

Figure: 3 Shows That The Memory Usage Of The Changing The Original Database

The sensitive diseases take memory usage since it was calculated with the difference of the starting time and ending time and stored for future use. The sensitive diseases were found out by setting the threshold values. After changing the sensitive diseases, format was made different from the normal format since the sensitive disease takes more memory space. Normally the seasonal diseases take more memory usage but here the seasonal diseases count was very less since it took very less time. The frequent diseases had the moderate memory usage when compared with the other type of diseases.

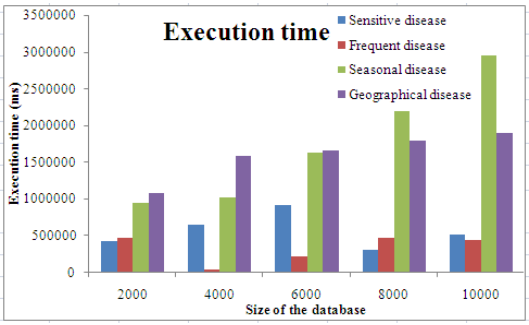

Figure: 4 Shows That The Execution Time Of The Changing The Original Database

The frequent diseases can be easily found out from the database as it had been modified to become normal diseases also it took very less time while comparing with the other types of diseases. Usually, the finding of sensitive diseases take some more time than the frequent diseases as the running time of the frequent diseases is high. In the present study, when comparing the other two diseases, the diseases took more time since the discovering of the seasonal disease took more time as the sliding window method was used, which takes more time to find the seasonal disease.

7. CONCLUSION

The present study has presented an efficient algorithm for privacy preserving temporal pattern mining. Hospitals sharing their medical data to different organization for different purposes may violate the privacy of the patient while doing so. In order to prevent that, the privacy on four different types of diseases of the patients like sensitive diseases, frequent diseases, seasonal diseases and geographical diseases have been made. Hospitals can choose any of the above four diseases and make the privacy according to the organizations with whom their data is shared. Since every organization needs some particular data, according to that the hospitals can make the privacy on such data. Here, the privacy is achieved by changing the original database into sanitized database, this sanitized database is used for sharing from hospital to any organizations. Here, the prefixspan algorithm was used to mine the important data from the original database. Finally the experimentation was made with the synthetic dataset and analyze for the differences among the four types of diseases while making the privacy in terms of the execution time, memory usage.

REFERENCES

[1] Ali İnan, Selim V. Kaya, YücelSaygin, ErkaySavas , Ayca A. Hintoglu, Albert Levi, "Privacy Preserving Clustering On Horizontally Partitioned Data", Data and knowledge engineering , vol. 63, no. 3, pp. 646-666, 2007.

[2] S. Laxman and P. S. Sastry, " A survey of temporal data mining", SADHANA, Academy Proceedings in Engineering Sciences, vol. 31, no. 2 pp. 173-198, 2006.

[3] Richard Huebner, "Barriers to Adopting Privacy-preserving Data Mining", In proceedings of academic and business research institute international conference, 2012.

[4] Elisa Bertino, Beng Chin Ooi, Yanjiang Yang, Robert H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data"Proceedings of the 21st International Conference on Data Engineering, pp. 521-532, 2005.

[5].Arik Friedman, Ran Wolff and Assaf Schuster, "Providing k-anonymity in data mining" , The International Journal on Very Large Data Bases, vol. 17, no.4, 2008.

[6] R. Agrawal and R. Srikant, "Privacy-preserving data mining", In Proceedings of the 2000 ACM

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SIGMOD international conference on Management of data, pp. 439-450, 2000.

[7] Y. Lindell and B. Pinkas, "Privacy preserving data mining", Advances in Cryptology, pp. 36-54, 2000.

[8]. Jinlong Wang, CongfuXu, YunhePan,"An Incremental Algorithm for Mining Privacy-Preserving Frequent Itemsets, " proceedings of international conference on machine learning and cybernetics , pp.1132-1137, 2006.

[9] Socrates J. Moschuris and Michael N. Kondylis, "Outsourcing in public hospitals: a Greek perspective", Journal of Health Organization and Management, vol. 20, no. 1, pp. 4-14, 2006.

[10]. Marcia Angell, "Drug Companies & Doctors: A Story of Corruption", The New York review of books, vol. 56, no.1, 2009.

[11]. Aaron S. Kesselheim, M.D., J.D.,,, David M. Studdert, and Michelle M. Mello, J.D.,"Whistle-Blowers' Experiences in Fraud Litigation against Pharmaceutical Companies" the new England journal of medicine, vol.362, no.19, pp. 1832-1839, 2010.

[12] Keng-Pei Lin and Ming-Syan Chen, "Privacy-Preserving Outsourcing Support Vector Machines with Random Transformation", In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 363-372, 2010.

[13] Jieh-Shan Yeh and Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining", Expert Systems with Applications: Elsevier Publication, vol. 37, no. 7, pp. 4779-4786, 2010.

[14] Maryam Khan, "Medical Tourism: Outsourcing of Healthcare", International CHRIE Conference-Refereed Track, pp. 23, 2010.

[15] Yonghong YU and Wenyang BAI, "Integrated Privacy Protection and Access Control over Outsourced Database Services", Journal of Computational Information Systems, vol. 6, no. 8, pp. 2767-2777, 2010.

[16] Ken Barker, Mina Askari, Mishtu Banerjee, KambizGhazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams, "A Data Privacy Taxonomy", In Proceedings of the 26th British National Conference on Databases, pp. 42-54, 2009.

[17] Jingquan Li and Michael J. Shaw, "Safeguarding the Privacy of Electronic Medical Records", Cyber Crime: Concepts, Methodologies, Tools and Applications, vol. 2, no. 1, pp. 232-249, 2012.