ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Privacy Protection Using Sensitive Data Protection Algorithm In Frequent Itemset Mining Of Medical Datasets.

R Dheepa*, and D Usha Nandini.

Faculty of Computing, Sathyabama University, Chennai, Tamil Nadu, India.

ABSTRACT

Frequent Itemset Mining (FIM) is one of the most eminent techniques in the Data mining systems. The exploration of Frequent Itemset Mining distills the recurring knowledge from the incessant data. Explosion of Frequent Itemset Mining in the field of Data Analysis and Data Mining becomes an inescapable one. The paper focuses on "searching the accurate records of efficient database queries without compromising the breach of trust using Sensitive Data Protection Algorithm". This algorithm divides the database into several partitions. Each partition holds the frequent items of the database in a ranked manner. In data protection perspective, the data gets luxated rather than adding blowing of noise. Next, a user-defined threshold is located to retrieve the necessitate records from the datasets which reduces the time consumed for scanning the whole database. The above process is executed for only authorized user, if any violation, an alert is activated and forwarded to the specified user. In experimental view, it is checked on "Doctor online" dataset that widely used for analysis of medical database queries. Performance metrics such as Precision and Recall is analyzed. Experimental results prove that the medical records are easily retrieved and protected with an improved sensitive data protection.

Keywords: Data Mining systems, Frequent Itemset Mining, Sensitive Data Protection (SDP) , Medical database queries.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Frequent itemsets mining is the most eminent and crucial part in the data analysis and its applications. It is one of the best researched approaches for finding efficient and effective relationships among the variables [1]. Each item in transaction supports to predict the accurate model for frequent itemsets. The latest innovation in the technologies effects to the generation of massive amounts of data storage. Investigation on the large databases stimulates a great demand for both business and academics. As the name "frequent" suggests, it extract information from the large databases in accord to the frequently occurring events of a user specified least frequency threshold. Nevertheless, the tracking of real data is more prone to noise and measurement error [2]. Several techniques developed to extract information from databases based on frequent events. Still, there is a lack of algorithms that not suited for large databases. Nowadays, the issue of noise on the conventional frequent itemsets mining techniques is limited to the improvement of noise- resistant algorithms. And also, several recent approaches controlled the outcomes and its runtime by increasing the minimum frequency threshold that leads to lessen the number of candidate sets and frequent itemsets. Therefore, yet a clear solution required to deal with low-frequency thresholds in mining the large databases.

An inclined growth of the information makes to delve in sensitive data investigation. These sensitive data should well protect for the business development. The original information is transferred to its specified database [3]. The sensitive data analyzed and based on their results; privacy protection scheme is developed with the aid of data mining systems. The data include purchasing details, credit information details, criminal details etc. Data plays a crucial role in the business organizations and governments for decision-making processes.

On the one hand, this data paves a way to new threat of misusing of user's privacy and added efforts on the computational power. Several companies need to share their data with the colleagues to cut the data burden. In that case, the sensitive information requires the application of mechanisms to make sure about the data privacy.

When the data shared, some attributes are set as in private or protected mode to prevent from misuse behaviors in such a way that the quality of the original goal of data mining is not affected. Several masking techniques are available for protecting the sensitive data attributes.

Many business applications expect the factors such as less cost, more availability, agility development and risk management towards the cloud computing. Cloud is a way of delivering IT services. Rapid development of data in the field of real-time business applications intends to cloud storage. Now, there is a possibility to store all our data in the internet. IT managers are grabbed by cloud storage with its low-cost and the ability of adjustments with the other cloud servers [4].

Though it offers reduced capital investment cost, clients face some technical and security at various levels. Data security is the biggest issue in cloud storage. This paper intends to supply the privacy protection using Sensitive Data Protection mechanism in Frequent Itemset Mining. It works on extracting the sensitive information from the often visited events to make a cleared decision for the occurred events.

Literature Survey

Fatih Altiparmak et al, 2006 suggested Information Mining over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases. They proposed a novel methodology for data mining that includes two noteworthy steps: applying an information mining estimation over homogeneous subsets of information, and distinguishing basic or distinct pattern over the data assembled in the initial step [1]. Our methodology is actualized particularly for heterogeneous and high dimensional time arrangement clinical trials information. Their methodology was executed particularly for heterogeneous and high dimensional time oriented clinical trials information. Utilizing this system, they proposed another method for using continuous itemset mining, and in addition grouping declustering systems with novel separation measurements for measuring closeness between time series information. By clustering the information, they discovered gatherings of analytes (substances in blood) that are most emphatically correlated. Most of these connections definitely known are checked by the clinical boards and recognized novel bunches that need for further biomedical investigation.

Ehud Gudes et al, 2006 framed Discovering Frequent Graph Patterns Using Disjoint Paths. They concentrated on the issue of finding typical patterns of graph data. An assignment made troublesome in view of the multifaceted nature of required subtasks, particularly subgraph isomorphism. They proposed another Apriori-based calculation for mining graph information, where the fundamental building blocks are moderately substantial in a disjoint ways. In scanning for incessant patterns, applicants are built utilizing continuous ways. The plan proposed here can be stretched out in a few courses, for example, partially labeled patterns [2].

Zhaonian Zou et al, 2010 framed Mining Frequent Subgraph Patterns from Uncertain Graph Data. Novel model of uncertain graphs is exhibited, and the continuous subgraph example mining issue is formalized by presenting another measure, called expected support. This issue was turned out to be NP-hard [3]. An inexact mining calculation was proposed to locate an arrangement by allowing an error tolerance on expected supports of discovered subgraph patterns. The algorithm utilized productive systems to figure out if a subgraph example can be yield or not and another pruning strategy to diminish the complexity of finding subgraph designs. The computational complexity of this problem has been formally proved.

Avrilia Floratou et al, 2011 framed an Efficient and Accurate Discovery of Patterns in Sequence Data Sets. They exhibited another estimation called Flexible and Accurate Motif Detector (FLAME). FLAME [4] is an adaptable suffix tree-based estimation that can be utilized to discover incessant designs with an assortment of meanings of theme (example) models. It is additionally accurate, as it generally discovers the patterns that it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics. In addition, based on FLAME, we also address a more general problem, named extended structured motif extraction, which allows mining frequent combinations of motifs under relaxed constraints.

Asier Aztiria et al, 2013 framed learning Frequent Behaviors of the Users in Intelligent Environments. This paper displayed a framework, Learning Frequent Patterns of User Behavior System (LFPUBS), which finds clients frequent patterns into the particular components of IEs. The center of LFPUBS [5] is the learning layer, which, not at all like some different segments, is autonomous of the specific environment in which the framework is being connected. On one hand, it incorporated a dialect that permits the representation of found practices in a reasonable and unambiguous way. Then again, coupled with the dialect, a calculation that finds incessant practices has been outlined and actualized. However, the learning layer, which implements all of the algorithms that discover users' frequent behaviors, is free of any influence of particular environments.

Faraz Rasheed and Reda Alhajj, 2014 proposed A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences [6]. They exhibited a strong and time productive suffix tree-based estimation equipped for distinguishing the periodicity of exception pattern in a period arrangement by giving more noteworthy to less incessant patterns. A postfix tree for a string signified its suffixes. It contains a recognized way from the root for each of the suffixes of the string. The most essential part of the suffix tree identified with their work was that it effectively catches and highlights the reiterations of substrings inside of a string. They also proved a better Suffix Tree-based Noise Resilient (STNR) in terms of time and space efficient.

Yaling Xun et al, 2015 framed a parallel mining of frequent itemsets using Mapreduce. They outlined a parallel continuous itemsets mining algorithms called FiDoop utilizing the MapReduce programming model [7]. To enhance FiDoop's execution, they add workload parity metric to gauge load parity over the cluster processing hubs. They created FiDoop-HD, an expansion of FiDoop, to accelerate the digging execution for high-dimensional information analysis.

Sen Su et al, 2015 framed differentially private FIM via transaction splitting. They developed the algorithm based on FP-growth, named as PFP- growth that comprised of preprocessing phase and mining phase [8]. The novel splitting method is used for the database transformation in preprocessing phase. In mining phase, the loss of information produced by transaction splitting is done. Noise is added to the itemsets in database to enhance the privacy level. Still, it lacks to add privacy to the data.

ISSN: 0975-8585

Intensified Sensitive Data Protection Algorithm

Problem Definition

Nowadays, the development of uncertain data is increasing. The mining of the information over uncertain databases has grabbed much attention. The traditional transaction databases extract the data based on the frequent itemsets i.e content of each transaction where item is precised. The real-life applications such as medical databases, location-based services etc make an idea about the need of mining of uncertain data. From various existing studies, the problems identified are listed as:

- Transaction splitting of a database is a time-absorbing one. Several databases are truncated into several forms that lead to frequency loss of information. As a result of high frequency loss, the inaccurate results are retrieved.
- Noise is mingled with the databases to increase its intensity. Several researchers added noise to the datasets, so as to increase the privacy level. Anyhow, it is deficient one.
- In case of authorized user access, there is no precised algorithm that detects the actual itemset even noise is added to it.
- No notification is sent to the authorized user, if any privacy violation occurs.
- The execution time of algorithm in preprocessing phase and mining phase found in higher rate in terms of precision and recall.

Objectives of the System

The research aims to solve the challenges exists in the previous works. The target of the system is listed as follows:

- To analyze the database without itemset splitting.
- To protect the data using Elliptical Curve Cryptography (ECC) algorithm. The data gets encrypted and stored in database in a jumbled manner.
- To eliminate the intensity of the databases i.e Reduction of adding noise to the datasets.
- To create a notification in case of misuse of data.
- To decrypts and rejumbled the data for authorized access.
- To well-utilize the sensitive attributes for Frequent Itemset Mining.
- To analyze the performance metrics such as Precision and Recall.

Improved Elliptical Curve Cryptography

The advent of cryptography is to guarantee the security of the information over the public channel. The two objectives of the cryptography are the authorization and authentication. Elliptic Curve Cryptography (ECC) belongs to the class of public-key cryptography. It works on the basis of algebraic structure of elliptic curves over finite fields. The novel Sensitive Data Protection (SDP) is patterned from the base of Elliptical Curve Cryptography (ECC) with additional effort of choosing infinite Galois points. The improved Sensitive Data Protection algorithm works as follows: a) Data Insertion and Encryption b) Data decryption and retrieval

Data Insertion and Encryption

Data Encryption and Insertion is the first step in an improved Sensitive Data Protection algorithm. Each unique ID is selected as the public key. Doc_Id and Pat_Id are chosen as the finite points. The sensitive attributes are Doc_id, Pat_dis, pat_dis_date, treatment, suggestion and results are treated as the private key. Since communication is done in public channel, the property of sensitivity is achieved. In an encryption process, two ciphertexts are generated. There are two tables namely Jumbled and Keytab.

The table 'Jumbled' is treated as the Ciphertext 1 and the table 'Keytab' is treated as the Ciphertext 2. The Ciphertext 1 consists of Pat_Id, Doc_Id and Sensitive attributes. The Ciphertext 2 consists of Sensitive attributes, Pat_Id and generating Max (id). These are encrypted and stored in a jumbled manner. The working of Data insertion and encryption is given as:

Pseudo code for Data Insertion and Encryption phase:

```
Begin
{
tables-jumbled, keytab;
input-' Sensitive attributes' for jumbled table.
   n- number of attributes.
  id- an integer used in random num generation.
if id=1
   insert attributes in 'jumbled' table;
   for (all columns)
   {   Set 1 in 'j'th column of keytab;   }
else   {
   Insert attributes in 'jumbled' table;
   get Max (id) of 'jumbled';
   insert max (id) at first column of keytab;
      for (except first column)
    {    rn = randomnumber (max (id));
    set rn at jth column (keytab);
    set max (id) at rn'th row (keytab);
    swap (value at jumbled (max (id), j) , value at jumbled (rn, j) )
    }}} end
```

In the below manner, the original data gets encrypted and stored to the database in the jumbled manner.

The Ciphertext 1 and Ciphertext 2 is encrypted and stored to the database in combined manner. The sample output is shown as follows:

4200AECE1CE90000000467414D410000B18F0BFC6105000000097048597300000EC300000EC301C7

Data Decryption and Retrieval

Data Decryption and retrieval is the second step in the intensified Sensitive Data Protection algorithm. This process is executed only for the authorized users.

The original data should retrieve only for the authorized users. It takes unique ID as the identifier to retrieve the stored records of authorized users. The original data is obtained with the help of private key.

The original data $M= C_2 - Id * C_1$

Pseudo code for Data Decryption and Retrieval

```
Begin {
Tables-jumbled, keytab;
input-id attribute
for (except first column)
{search id at keytab;
  while (id found)
  { get row number at keytab;
  display value at row number of jumbled;
  }}} End
```

ISSN: 0975-8585

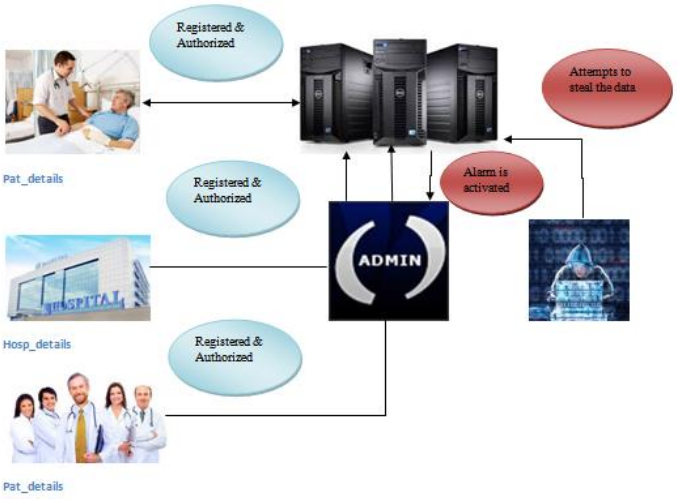

Figure 1: Working flow of proposed approach

As the data gets jumbled and stored in the server, the data size is lessened. The lessened data size elongates the storage space of the data. This is one of the core regions of our study and it is achieved using Sensitive Data Protection algorithm.

RESULTS AND DISCUSSION

The hospital details are obtained from the public dataset "Doctors Online". The study is restricted only to Tamilnadu location.

Performance Metrics and Validation

The core area of this research is the study of mining the frequent medical queries. This research approaches to extract the frequent information from the known information in order to retrieve the records in user-friendly manner. Performance metrics such as Precision and Recall is analyzed. The sample medical dataset is given below:

Table 1: List of Hospital details

Name	Hos_ Id	Password	Email_id	Address	Founder	Experience
Apollo specialty hospitals	Hos_ 1	Universe	Dheepam123@gmail.com	320, Anna Salai, Teynampet, Chennai – 600035.	Dr.Velayutham	12
Asian Hospitals	Hos_ 2	Asiana	Asianhosp123@yahoo.co.in	E-147\A, 2nd Avenue, Besant Nagar, Chennai -600090.	Dr.Manjula	20
A. G. Hospital	Hos_ 3	aghosp	aghospitals@gmail.com	8-C, Kakkan Street, Chennai - 600045	Dr.Gurusamy	12
A.G. M Hospital	Hos_ 4	Agmhosp	agmhosp@yahoo.co.in	923,Poonamallee High Road, Chennai - 600084	Dr.Muthaiah	14
A.V.M.Medicals ENT Research Foundation Pvt Ltd	Hos_ 5	avmmedic	avmmedicalsent@gmail.com	No-3, P.S.Sivasamy Salai, Chennai - 600004	Dr.Sharaon	14
Aashiana Hospital	Hos_ 6	ashianahos	ashianahosp@gmail.com	154, P.H.Road, Chennai- 600010	Dr.Agarwal	17
Agarwalls Eye Institute Limited	Hos_ 7	Agarwaleye	agarwaleye@yahoo.co.in	13, C.D.L.Road, Chennai- 600086	Dr.Amarnath	12

ISSN: 0975-8585

Table 2: List of doctor's suggestions for patients

Doc_Id	Pat_Id	Treat	Suggestions	Result
Doc-1	P1	Dialysis	Less Water Intake	Critical
Doc-2	P2	Byepass	Normal Bp	Fine
Doc-3	P3	Surgery	Lnormal Bp	Emergency
Doc-4	P4	Surgery	Bed Rest	Intensive Care
Doc-5	P5	Physiotherapy	Bed Rest	Critical
Doc-6	P6	Medication	Diet	Fine
Doc-7	P7	Injection	Less Sugar	Emergency
Doc-8	P8	Injection	Precaution	Intensive Care
Doc-9	P9	Injection	Precaution	Emergency
Doc-10	P10	Injection	Precaution	Intensive Care
Doc-11	P11	Surgery	Normal Bp	Critical
Doc-12	P12	Dialysis	Less Water Intake	Fine
Doc-13	P13	Injection	Precaution	Emergency
Doc-14	P14	Injection	Normal Temperature	Intensive Care
Doc-15	P15	Injection	Precaution	Emergency

Table 3: List of Patient's details

Hos_Id	Name	Pat_Id	Gender	DOB	addr	P_Dis	P_DisF	P-Temp	P-bp	P-Pul	Doc_Name
Hos_1	Karen	Pat_1	M	3/4/1975	Chennai	Cardiac	31/10/15	98	120/80	73	Nadine
Hos_2	Charlotte	Pat_2	F	5/6/1987	Mumbai	Liver	10/04/15	97	110/70	72	Dana
Hos_3	Chandler	Pat_3	M	3/8/1979	Delhi	Heart	20/10/15	98	120/80	69	Kai
Hos_4	Travis	Pat_4	M	7/9/1999	Chennai	Spinal	19/07/15	99	100/80	81	Carol
Hos_5	Odessa	Pat_5	M	1/4/1978	Delhi	Dengue	11/09/15	100	120/80	73	Deborah
Hos_6	Karen	Pat_6	M	1/5/1980	Mumbai	Swine Flu	15/11/15	102	110/90	76	Caleb

Table 1-3 describes the sample details of No. of hospitals, No.of patients and Suggestions for diseases. Let us consider, No.of Hospitals =10, No. of Patients = 100 and diseases such as Cardiac, Dengue, Swineflu, Spinal and Liver. The limitation in this study is the location. We analyzed our study in TamilNadu location. According to our study, the Frequent Itemset Mining is designed as:

Input: Get the City name (Cname); Disease name (Dname)
Steps:
Begin
get Cname;
get Dname;
for each dname in the node n.
set count(dname) for every dname.cname.n
Display count (max (dname))
End
Output: max (dname) is the most frequently searched disease of particular location.

Classification Matrix (CM) is a significant tool in data mining. Classification matrix arranges the events into groups, by fixing whether the observed value equates with the actual value. CM is given as:

Observed Value

Actual value

	Relevant	Irrelevant
Retrieved	TP (40)	FP (15)
Not Retrieved	FN (25)	TN (20)

Suppose let us try to get the information about the disease 'Swineflu' in Tamilnadu location.

Precision

Precision is defined as the percentage of retrieved disease that is relevant to the location. It is given by:

$$\text{Precision} = TP/(TP+FP) = (40)/ (40+15)$$
$$= 0.727 * 100\% = 72\%$$

Recall

Recall is defined as the percentage of relevant disease that is retrieved. It is given by:

$$\text{Recall} = Tp/(TP+FN) = (40)/ (40+ 25)$$
$$= 0.615 * 100\% = 61\%$$

In reference to data mining rule, precision and recall are inversely related to each other. As precision value increases, recall value decreases and vice versa. From the above results, it is concluded that the disease and its suggestions in particular location is precisely retrieved.

CONCLUSION

Frequent Itemset Mining is the core area in Data Mining. It intends to extract the information from the incessant patterns. In this paper, the intention of our study is "searching the accurate records of efficient database queries without compromising the breach of trust using Sensitive Data Protection Algorithm". It works in two phases. Firstly, the sensitive data are protected using improved Sensitive Data Protection. Before storing into the databases, the records are encrypted and arranged in a jumbled manner. Whenever any authorized doctor or patient searches for any medical queries, it gets decrypted and rejumbled to its original form. If any violation occurs, a notification mail is sent to the administrator.

Secondly, it is concentrated on easily retrieving the frequently visited medical queries using the idea of Frequent Itemset Mining. Based on the name of a city and disease, the topmost visited queries get recorded.

This recorded pattern will display to the users. In experimental perspectives, the performance metrics such as precision and recall are studied. From the study, it is inferred that the disease and its suggestions in particular location is precisely retrieved. As a future work, Frequent Itemset Mining will be extended to other location. And, some masking techniques will also used for privacy protection.

REFERENCES

[1] Fatih Altiparmak et al. IEEE Transactions On Information Technology In Biomedicine 2006; 10: 255-263.

[2] Ehud Gudes et al, Discovering frequent graph patterns using Disjoint paths, IEEE Transactions on Knowledge and Data Engineering, November 2006; 18.
[3] Zhaonian Zou et al, Mining frequent subgraph patterns from uncertain graph data, IEEE Transactions on Knowledge and Data Engineering , September 2010; 22.
[4] Avrilia Floratou et al, Efficient and accurate discovery of patterns in sequence data sets, IEEE Transactions on Knowledge and Data Engineering , August 2011; 23: 1154-1168.
[5] Asier Aztiria et al, Learning frequent behaviors of the users in intelligent environment, IEEE Transactions on Systems, Man, And Cybernetics: Systems , November 2013; 43: 1265-1278.
[6] Faraz Rasheed and Reda Alhajj, A framework for periodic outlier pattern detection in Time-series sequences, IEEE Transactions On Cybernetics , May 2014; 44: 569-582.
[7] Yaling Xun et al, FiDoop: Parallel Mining of frequent itemsets using Mapreduce, IEEE Transactions on Systems, Man, And Cybernetics: Systems, 2015.
[8] Sen Su et al, Differentially private frequent itemset mining via Transaction Splitting, IEEE Transactions On Knowledge And Data Engineering , July 2015; 27: 1875-1891.