Concealing Sequential and Spatiotemporal Patterns using Polynomial Sanitization

A.Vijay
Dept. of IT
JNTUK UCEV
A.P., 535 003, India

MHM Krishna Prasad, PhD.
Associate Professor of CSE , JNTUK UCEK
A.P., 533 003, India

ABSTRACT

Earlier, Process of relevant pattern observation which is present in the database observed as a hurdle for database protection. Over the time, various approaches for hiding knowledge have emerged, mainly in the focus of Association rules and frequent item sets mining. This paper, have seen the problem in different view i.e., Knowledge hiding to the context where the data and extracted knowledge have a sequential structure. The concept of NP hardness is observed over the sequential pattern hiding. A polynomial sanitization algorithm was adopted and implemented over the spatiotemporal patterns extracted from moving objects databases. Disseminating datasets of this kind presents a considerable opportunity for knowledge patterns of interest. The developed model is kept under the attack, which exploits the knowledge of underlying road networks.

General Terms

Sequence Hiding, Data Sequences, HHA algorithm

Keywords

Length of sequence, Data sanitization, spatiotemporal patterns.

1. INTRODUCTION

Knowledge acquisition is regularly referred to as a bottleneck in the development of expert systems and other knowledge-based systems [4]. The support of numerous applications on sequential data are increased, where the primary interest rely on sequentiality of data. Web usage logs where the records of webpage accesses, mobility data captured by mobile devices at different points in mobility time are some of the driving examples in the day by day life. Broadly, sequential data extends a great deal of opportunities for discovery interesting behavioral patterns that can be beneficial to a various domains of people. Mining user mobility data can reveal interesting patterns that aid traffic engineers and environmentalists in their decision. The publishing of sequential data for data mining purpose may lead to severe violation of privacy. To address these concerns knowledge hiding methods [6] are necessary. Methods conceal sensitive patterns that can otherwise be mined from published data. The problem is to find all sequential with a user-specified minimum support, where the support of a sequential pattern is the percentage of data-sequences that contain the pattern. But the techniques applied over the sequential data by coarsening and the sanitization will yield the results in undefined time when the size or cost of the datasets is maximized. Sometimes it leads to the problem of NP-hardness. And the aspect of the security in the area of finding the behavior of the sequences has a great deal with these techniques. The information and quality of data is a serious issue just before the operation of before and after hiding. Here in this paper, proposed an efficient technique for finding the cost of the sequences which will be irrespective of the thresholds and produces entire patterns having the given input items present in the every pattern of the entire data set.

2. BACKGROUND

The level of security over database was sounding factor in general, set of controls. Most critical Security vulnerabilities in data applications are caused by inadequate manipulations of input strings. To secure data against knowledge discovery [4] Sequential pattern mining methods have been used to analyze this data and identify patterns. Such patterns have been used to implement efficient systems that can be recommended based on previously observed patterns, help in making predictions improve usability of system, detect events and in general help in making strategic product decisions [10]. Publishing the sequential data and spatiotemporal patterns may lead to the severe security find out by the data mining techniques [2]. In the area of the privacy preserving data mining has devoted much more effort to determine a trade-off between the right to privacy and need to knowledge Discovery which is a crucial in order to improve the decision making process [9]. Sequence search and retrieval techniques play an important role in interactive exploration of large sequential database .After applying the Sequence Hiding techniques over these data, the sensitive patterns irrespective of the threshold values get suppressed before publishing. Distortion is found as a problem and it can be overcome by allowing more useful data for sequential mining to be produced. The sequence hiding problem requires sanitizing the database D so that no sensitive sequence can be mined from D’ at a support threshold, no side effects are introduced by the hiding process D’. The least number of events in sequences supported in D is sanitized to derive D’, which implies that D’ should be kept as similar as possible to D. if symbols are marked with * when sanitized, then distance (D, D’) is equal to number of *(symbol) in D’. The problem is to discover all sequential patterns with a user-specified minimum support, where the support of a pattern with a user specified data-sequences [3]. The disturbances in the cost of the data will eventually distort the behavioral aspects [1] which lead to side effects. In real world scenarios, it may be the case that the sensitivity level of patterns differs. For these cases there is need to extend our framework to deal with multiple disclosure threshold. No ghost Sequences can be introduced by a choice adopted by related work in association rule hiding [8]. A straight forward way of implementing such an extension is to simply take the minimum of all thresholds. Though this approach is correct [2], it may easily result in over killing distortion especially when the disclosure thresholds vary significantly.

3. RELATED WORK

Approaches for privacy preserving data sharing fall into two general categories. The first category of approaches attempt to protect the privacy of individuals, whose information is contained in the data, by preventing the disclosure of individual's identity or sensitive information. The second category, referred to as knowledge hiding, aims to prevent sensitive patterns from being mined from the data. Several hiding algorithms have been proposed with most of the research being conducted along the lines of protecting sensitive association and classification rules. In particular, association rule hiding evolved from efficient heuristic approaches, to border-based approaches and more recently, to exact hiding approaches that offer stronger quality guarantees at the expense of high computational cost [6]. The problem is concerned with efficiently locating subsequences in large archives of sequences or sometimes in single long sequences. Classification rule hiding , on the other hand, evolved around perturbation-based technique that reduce the confidence of sensitive rules, by modifying the values of attributes that support these rules , and reconstruction – based approaches that reconstruct the dataset by using only records supporting non sensitive rules . One important issue is that of what constitutes an interesting pattern in data. The notations of sequence patterns represent only the currently popular structures for patterns. The idea of patterns in sequential data describe how patterns are typically matched and retrieved from large sequential data achieve. The problem of sequential pattern hiding was recently investigated where the focus was on hiding the sensitive knowledge in a way that minimally affect sequences the support of the rest of the sequences in the database. The proposed HHA algorithm operates as follows, first for each sequences of original database, this algorithm computes the different ways called matching's in which sequence support any sensitive sequences [7].The search for sequential patterns begins with discovery of all possible item sets with sufficient support. Here support of an item set or events was defined as fraction of all sequences that contained item set. Then, the original sequences are sorted in ascending order with respect to the number of matching that they contribute to, and the top sequences are selected for sanitization , based on a user specified disclosure threshold. The sanitization operation eliminates all matches of the sensitive sequences in the sequences by marking selected events with a special symbol *. To sanitize the sequence, HHA finds, for each event e in sequence, the number of matching's to which e contributes and marks the events that contribute to most matches, until sequences no longer supports sensitive sequences. The approach has three limitations. First, the problem formulation adopted does not focus on side-effect s that may be introduced by the hiding large number of potentially interesting sequences. As a result, a large number of potentially interesting sequences may be lost in the sanitized dataset. Since the main reason behind enable the discovery of non sensitive frequent sequences, the problem formulation may lead to produce solutions of low data utility. Next, the global selection criterion, used by HHA to identify sequences for sanitization, often selects sequences that incur high distortion when sanitized. That is, it may example, sensitive sequences. At the end HHA needs to compute thresholds for each event in every sequence.

4. EXPERIMENTAL WORK

In typical data mining application like content based retrieval, it is approx matching that we are more interested. Sequence Hiding Algorithm eliminates all occurrences of input pattern with in a sequence by introducing the external symbol in chosen place. The problem of searching is concerned with efficiently locating subsequences often or sometimes in single long sequences. Query-based search have been extensively studied in language and automation theory. The problem of efficiently locating exact matches of substrings is well solved, the situation is approximate matches. Here the sequences are sanitized introducing the symbol. Under sanitization, symbols may be interpreted as missing values. The marking operations here don't create new subsequences. Thus there are no fake patterns, either sensitive or non sensitive. Though this approach is correct, it may easily result in over killing distortion. Polynomial Sanitization algorithm is defined for hiding a set of sensitive patterns from a database. The unreal trajectories in sanitization process avoidance are one of the requirements. Property of coarsening a trajectory is we must not worry about generating fake patterns or increasing the support of some sensitive patterns. The implemented Algorithm is applied over the road network data and the trajectory path of the entire sequences before and after sanitization is plotted. They are as shown

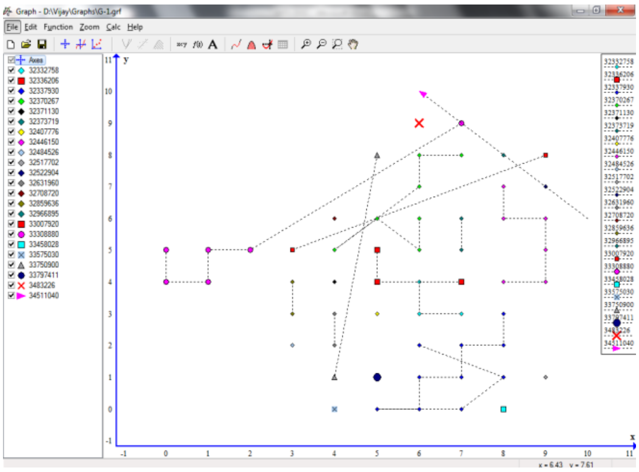

Fig.1 Representation of the Patterns before Hiding

And after sanitization by giving the sequences to be hidden are selected arbitrarily. The support of these sensitive patterns is important information since it strongly influence the distortion introduced by sanitization. Utmost repetitions are suppressed and unique events were taken for plotting. Fig1 represents the total number of the events suppressed and the next Fig2 represents the unique events got suppressed.

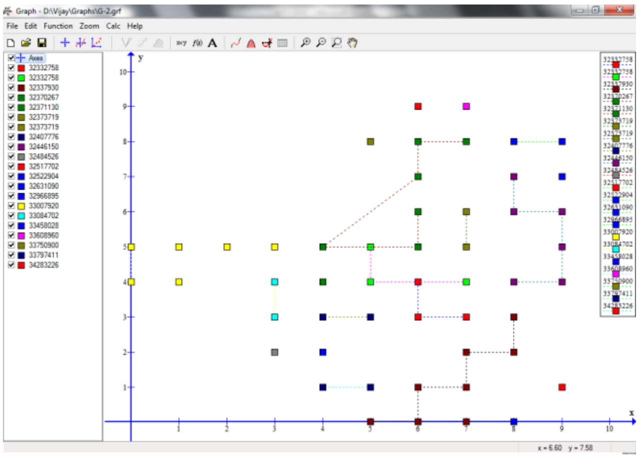

Fig.2 Representation of the Patterns after Hiding

Later the Sequence Hiding involves selecting a number of transactions to sanitize and then deciding which events need to be deleted to perform the sanitization. As our goal is to release a database in which all sensitive sequences are hidden, we first need to select the transactions from D that will be sanitized. Clearly, all transactions that do not support any sensitive sequences in S can safely be disclosed and will be part of D'. Among the transactions supporting sensitive sequences, a set of transactions can be chosen to be part of D' without being sanitized, as long as the support of each sensitive sequences and incur low distortion when sanitized. Approach show a significantly increase in speed over the other methods as the data size increases. Performance of mining process is highly associated with length of sequence and number of different symbols in sequence, the abstracted representation of our methods reduces the data size and prunes the search space in mining process. When it is applied on MSNBC dataset , randomly for first 11000 patterns. The results yielded will be cost with respect to time and it eliminated the costly function of find and replace. Initially the dataset will be read and each and every sequence will be re-arranged in incremental or decremented order, here the change of the behavior will not be considered due to the reason as fallow. The count of the events in the sequences and will not disturb behavior of the dataset. The matching of the input sequences given in the form of sensitive patterns will be done with each and every event in the sequence. Here those sequences having the sensitive patterns are gone under the operation of Distance based Sequence Hiding, eliminating the costly operation of comparisons in normal Sanitization and Hiding technique

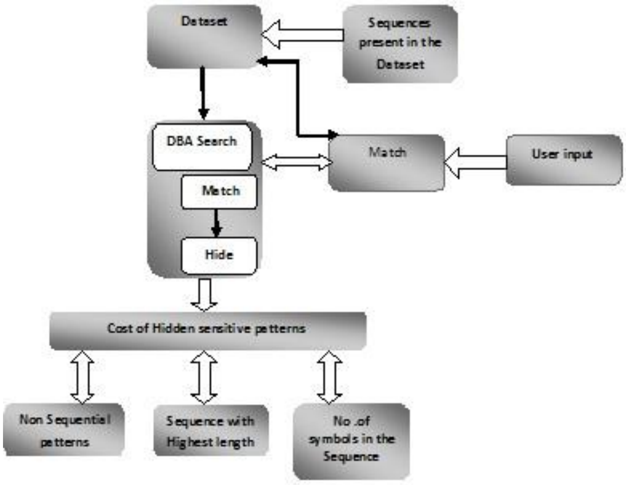
Dataset
Sequences present in the Dataset
DBA Search
Match
Hide
Match
User input
Cost of Hidden sensitive patterns
Non Sequential patterns
Sequence with Highest length
No .of symbols in the Sequence

Fig.3 Procedural representation of hiding using DBSH

Here the patterns of the dataset is taken for process which finds match with the input sensitive sequences and there by yield the cost in terms of the time. The data sets which are in the large size consisting of continuous events and items will be sorted with the ascending or descending order, for which the events will be counted and matching can be done easily. The goal can be restated as minimize the number of non sensitive frequent sequences that are lost. It incur low distortion favors fewer side-effects instead of a smaller distances. It doesn't disturb infrequent events present in sequences.

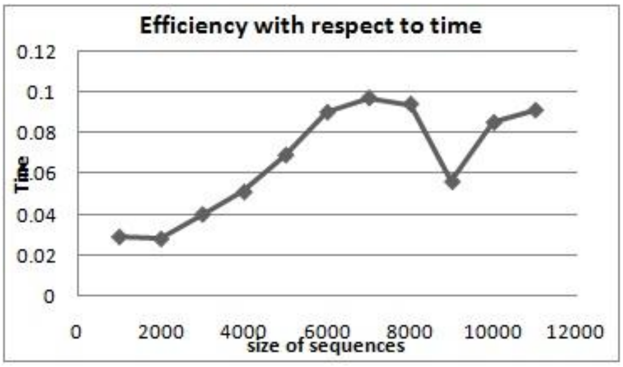

Fig.4 Graphical representation of patterns over the time

The performed evaluations are done over the 11000 patterns and it was observed the time taken of the matching with increase of matching will not vary. When the Distance based Sequence Hiding is applied it begins by computing the number of deletions required to sanitize each transaction. Distortion is eliminated by finding the exact distance between the events.

5. CONCLUSION AND FUTURE WORK

In this paper, authors presented a solution for the problem of sequential pattern hiding. By utilizing the coarsening approach that consists in reducing information contained in some sensitive trajectories by suppressing required spatial points. From our experimental observations, it is clear the technique is efficient in selection of items which must be selected for Hiding. And observed the graph based representation can be used to solve other privacy problem related to sequential and trajectory data. The cost function will eventually decrease the distance between the events. Experience with different applications would give rise to other useful notations and the problem of defining structures for interesting patterns would be a problem that deserves attention. Other algorithmic solution like autocorrelation of sequence is worth of further investigation.

6. REFERENCES

[1]. Revisiting Sequential Pattern Hiding to Enhance Utility, Aris Gkoulalas-Divanis, Grigorios Loukides .

[2]. O.Abul, F.Bonchi , and F.Giannotti .Hiding sequential and spatiotemporal patterns. IEEE KDE, 22(12):1709–1723, 2010.

[3]. Mining Sequential Patters, Rakesh Agarwal and Ramakrishnan Srikanth IBM Almaden Research Centre.

[4]. Knowledge Discovery as a treat to Database Security by Daniel E.O'Leary.

[5]. Computer and intractability by Micheal R. Garey , David S. Johnson.

[6]. C.C Agarwal and P.S.Yu . Privacy Preserving data mining : Models and algorithms , Springer,2008.

[7]. O.Abul, F.Bonchi, and F.Giannotti. Hiding sequences . In ICDM workshop.

[8]. A.Gkoulalas-Divanis and V.S.Verykios. Association Rule Hiding for DataMining.Springer, 2010.

[9]. A Framework for Evaluating Privacy Preserving Data mining. Springer 2009.

[10]. Approaches for pattern Discovery using sequential Data Mining. Manish Gupta, JiaweiHan