Frequent Itemsets Mining with Differential Privacy over Large-scale Data

Abstract

Frequent itemsets mining with differential privacy refers to the problem of mining all frequent itemsets whosesupports are above a given threshold in a given transactional dataset, with the constraint that the mined results should not break the privacy of any single transaction. Current solutions for this problem cannot well balance efficiency, privacy and data utility over large scaled data. Toward this end, we propose an efficient, differential private frequent itemsets mining algorithm over large scale data. Based on the ideas of sampling and transaction truncation using length constraints, our algorithm reduces the computation intensity, reduces mining sensitivity, and thus improves data utility given a fixed privacy budget. Experimental results show that our algorithm achieves betterperformance than prior approaches on multiple datasets.

Existing System

Explosive growth of data and the rapid development of information technology, various industries have accumulated large amounts of data through various channels. To discover useful knowledge from largeamounts of data for upper-layer applications (e.g. business decisions, potential customer analysis, etc.), data mining has been developed rapidly. It has produced a positive impact in many areas such as business and medical care. Along with the great benefits of these advances, the largeamount of data also contains privacy sensitive information, which may be leaked if not well managed. For instance. Medical records are also storing potential relationshipsbetween diseases and a variety of data. Mining on user location data or medical record data both provide invaluable information; however, they may also leak user privacy. The company would like to make the dataset public and therefore allow the public to execute frequent itemsets mining for getting cooperation or profits. But due to privacy considerations, the company cannot provide the original dataset directly. Therefore, privacy mechanisms are needed to process the data.

Proposed System

We propose a novel differential private frequent itemsets mining algorithm for big data by merging theideas, which has better performance due to the new sampling and better truncation techniques. We build our algorithm on FP-Tree for frequent itemsets mining. In order to solve the problem of building FP-Tree with large-scale data, we first use the sampling idea to obtain representative data to mine potential closed frequent itemsets, which are later used to find the final frequent items in the large-scale data.

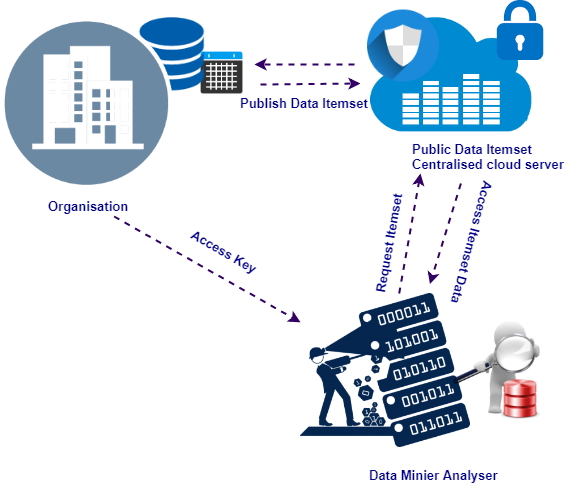
Future Work

DPFIM Data Partition Frequent Itemset Mining, which merges the ideas of, but employs a different (better) truncationscheme and boosts computation efficiency using both sampling and truncation. Compared with previous work using random truncation, our new string similarity- matching-based truncation mechanism has better performance than previous work, which is because string-similarity-matching-based truncation preserves more useful frequent itemset candidates.

Algorithm

Newly proposed algorithm, called DPFIM, which merges the ideas of, but employs adifferent(better) truncation scheme and boosts computation efficiency using both sampling and truncation. Compared with previous work using random truncation, our new stringsimilarity- matching-based truncation mechanism has better performance than previous work, which is because string-similarity-matching-based truncation preserves more useful frequent itemset candidates. The experimental results also confirm the better performance. The algorithm is differentially private; it takes a threshold value and outputs the frequent itemsets with support at least. The basic idea is as follows: first, compute a noisy support for the threshold,then truncate the original database noisily, finally construct a noisy FP-Tree for mining frequent itemsets.

Architecture



Modules

Data Itemset Publisher

A company has a large-scale dataset. The company would like to make the data Itemset as publiccentralized and therefore allow the public/Data Analyzer to execute frequent itemsets mining for getting cooperation or profits. But due to privacy considerations, the company cannot provide the original dataset directly. Therefore, privacy mechanisms are added to process the data.

Data analyzer

Request the Dataset and get access key to view and analyze the datasetTo discover useful knowledge from large amounts of data for upper-layer applications (e.g. businessdecisions, potential customer analysis, etc.), data mining has been developed rapidly. It has produced a positive impact in many areas such as business and medical care.

Frequent Itemsets Mining

Frequent itemsets mining, be a transactional dataset consisting ofN transactions, be a set of different items, and X be a subset of I such that X U I. If X is containedin a transaction and X has k items, X is called a k-itemset. The support of an itemset is defined as the total number of transactions that contains the itemset. frequent itemsets mining is to find all itemsetsthat have support greater than a given threshold. Frequent itemsets is employed for finding association rules for a group of data items. Given the large-scale dataset, we first sample the dataset and then compute the closed frequent itemsets in the smaller sample using a traditional frequent itemsets mining algorithm. We later estimate the length distribution of the sampled dataset and obtain the maximum length constraint, which is later used to shrink the dataset. Some elements out of the closed frequent itemsets are removed from the source dataset if their supports are below the support threshold. We then employ string matching ideas to cut off the transactions in the dataset; in this step, the purpose of converting the dataset is to shrink the data size and simultaneously retain the potential frequent items.

FIM Phases

We first sample the dataset to have a roughestimation of the dataset using the central limit theorem. Wefirst compute the sample size and then use SAS data analysissoftware for random sampling. The samples can reduce thecomputational intensity of the constructed FP-Tree and findthe potential frequent itemsets of the source dataset. We obtain a maximum length constraint lmax to shrinkthe transactions in the dataset.We deduce the sample size now. Fix an item modelled asa binomial distribution with occurring probability p.After thepreprocessing phase, we get the shrinking dataset which hassmaller number of transactions and smaller dimension to builda noisy FP-Tree. Because computing support directly destroysthe privacy.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

Conclusion

We propose a novel differentially private algorithm for frequent itemsets mining. The algorithm featuresbetter data utility and better computation efficiency. Various experimental evaluations validate that the proposed algorithm has high F-Score and low relative error. That fine-tuned parameters lead to better differentially private frequent itemsets mining algorithms with regard to data utility.