ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Secure Mining of Association Rules in Cloud Distributed Databases

Khalid Amjed Mohammed
Master of Science (Information System),
Nizam College (Autonomous),O.U,
Basheer Bagh, Hyderabad.

T. Ramdas Naik
Assistant Professor, Computer Science (PG)
Nizam College (Autonomous),O.U,
Basheer Bagh, Hyderabad.

ABSTRACT:

We propose a protocol for secure mining of association rules in cloud distributed databases. The current leading protocol is that of Kantarcioglu and Clifton. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al. which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost. With the existence of many large transaction databases, the huge amounts of data, the high scalability of distributed systems, and the easy partition and distribution of a centralized database, it is important to investigate efficient methods for distributed mining of association rules. This study discloses some interesting relationships between locally large and globally large item sets and proposes an interesting distributed association rule mining algorithm, FDM (Fast Distributed Mining of association rules). And with Using of CDBMS (Cloud Database Management System) is a distributed database that delivers computing as a service. It is sharing of web infrastructure for resources, software and information over a network. The cloud is used as a storage location and database can be accessed and computed from anywhere. The large number of web application makes the use of distributed storage solution in order to scale up. It enables user to outsource the resource and services to the third party server.

INTRODUCTION:

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.

- Web graph, from links between pages, people and other data.

- Web content, for the data found on Web pages and inside of documents.

At Scale Unlimited we focus on the last one – extracting value from web pages and other documents found on the web.Note that there's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results. A partial list includes:

- Business intelligence

- Competitive intelligence

- Pricing analysis

- Events

- Product data

- Popularity
- Reputation

Four Steps in Content Web Mining:- When extracting Web content information using web mining, there are four typical steps.

1.Collect – fetch the content from the Web

2.Parse – extract usable data from formatted data (HTML, PDF, etc)

3.Analyze – tokenize, rate, classify, cluster, filter, sort, etc.

4.Produce – turn the results of analysis into something useful (report, search index, etc)

WEB MINING VERSUS DATA MINING:

When comparing web mining with traditional data mining, there are three main differences to consider:

1.Scale – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.

2.Access – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler (a) promises not to overload the site, and (b) has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful/polite during the crawling process, to avoid causing any problems for the webmaster.

3.Structure – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

MINING OF DATA RECORDS :

Data Records are regular structures of information found on the Web. It is useful to mine such data records in order to obtain useful information from the corresponding web pages. There are different existing techniques for extracting them. One of the important strategies or techniques is based on two observations about web pages. In web pages the data records that describe some similar data items are found in a particular region known as Data Region. From the tag tree, i.e , the nested structure of html tags of the web page, these data regions can be easily identified as they will be under one single parent node as shown. Thus the technique is to build the tag tree of the page first and then the data region itself is mined from the tree. Then the data records within the region can be easily obtained. Another technique for mining data records from web pages is based on tag path clustering. The data records in an HTML web page are usually visually aligned and have a similarity in their appearance. The method identifies the tag path that appears repeatedly in the tag tree of the web document. In this approach, a pair of tag path occurrence patterns known as visual signals is compared to determine whether they represent the same list of objects.

MINING OF WEB TABLES :

Web is mostly found as huge corpora of unstructured documents, but it also have data organized in structured formats such as web tables that represent relational data. The corpus of such web tables is larger and many researchers have focused on the problem of identifying useful and relevant web tables rich in information and extracting knowledge from them. One of such systems is called Web Tables system that provides an efficient technique for performing keyword search over a corpus of tables. The corpus of html tables is large as most of them are used for content or page layout. Finding meaningful or related relations from them is important as they are rich in valuable information. The Web Tables system extracts such related relations from the corpus and ranks the relations by relevance using a keyword query as input in a search-engine-style.

Many ranking functions such as NaiveRank , FilterRank are used for performing the ranking of relations. Another technique of mining html tables is independent of the tree-based representation of web pages and focuses on the domain-independent information extraction. The approach makes use of the visual box model used by web browsers to display information for extracting information. The html documents when they are rendered by a browser are represented by rectangular boxes and it is called CSS (Cascading Style Sheet) box model or Visual box model. The web tables topologically form a frame in the visual box model. Firstly, the task involves identifying or locating these frames representing the tables in web documents. Then the relative spatial relationship between the logical cells of a table are recognized is identified. Meaningful information contained in all available visual relationship can then be extracted from by interpreting the obtained topological structure of the table.

MINING OF TOP-K WEB PAGES :

Top-k web pages are those web pages that describe k items of a particular topic of interest. They are rich and valuable source of information. Many recent researches have focused on extracting information from these top-k web pages in order to enrich knowledge bases for supporting many applications like fact or search answering. There is an approach for extracting information called top-k lists from these web pages. Compared to other structured information, these top-k lists are more useful as they are cleaner, larger and richer with high quality and precise information. In all the above mentioned data mining approaches, the focus was on structured data such as web tables or data records. But this one focuses on less structured or almost free-text information in top-k web pages and guides its extraction.

The system first recognizes top-k web pages by using a title classifier. The classifier is trained using a labeled dataset of positive and negative titles. Once it is recognized as a top-k web page, the system extracts all potential top-k lists. They are initially considered as candidate lists. Then obtained candidate lists are scored or ranked by using ranking functions or algorithms and the best one is chosen as top-k list. The extracted list is further enriched with attribute values and only the essential information is obtained from the top-k web page in the form of a top-k list.

PROBLEM STATEMENT:

Kantarcioglu and Clifton studied that problems and devised a protocol for its solution. The main part of the protocol is a sub-protocol for the secure computation of the union of private subsets that are held by the different players. The private subset of a given player, as we explain below, includes the item sets that are s-frequent in his partial database. That is the most costly part of the protocol and its implementation relies upon cryptographic primitives such as commutative encryption, oblivious transfer, and hash functions.

This is also the only part in the protocol in which the players may extract from their view of the protocol information on other databases, beyond what is implied by the final output and their own input. While such leakage of information renders the protocol not perfectly secure, the perimeter of the excess information is explicitly bounded and it is argued there that such information leakage is innocuous, whence acceptable from a practical point of view.

Drawbacks:

- Insufficient security, simplicity and efficiency are not well in the databases, not sure in privacy in an existing system.
- While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players.
- Our protocol may leak is less sensitive than the excess information leaked by the protocol.

PROBLEM DEFINITION:

The protocol that we propose here computes a parameterized family of functions, which we call threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact general-purpose protocols that can be used in other contexts as well. Another problem of secure multiparty computation that we solve here as part of our discussion is the set inclusion problem; namely, the problem where Alice holds a private subset of some ground set, and Bob holds an element in the ground set, and they wish to determine whether Bob's element is within Alice's subset, without revealing to either of them information about the other party's input beyond the above described inclusion.

ADVANTAGES :

•We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency.

•The main ingredient in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players holds.

IMPLEMENTATION:

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

PRIVACY PRESERVING DATA MINING :

One, in which the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold. In the first setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonym zing the data prior to its release. The main approach in this context is to apply data perturbation. The idea is that. Computation and communication costs versus the number of transactions N the perturbed data can be used to infer general trends in the data, without revealing original record information. In the second setting, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners. This is a problem of secure multiparty computation. The usual approach here is cryptographic rather than probabilistic.

DISTRIBUTED COMPUTATION:

We compared the performance of two secure implementations of the FDM algorithm Section In the first implementation (denoted FDM-KC), we executed the unification step using Protocol UNIFI-KC, where the commutative cipher was 1024-bit RSA in the second implementation (denoted FDM) we used our Protocol UNIFI, where the keyed-hash function was HMAC. In both implementations, we implemented Step 5 of the FDM algorithm in the secure manner that was described in later. We tested the two implementations with respect to three measures:

1) Total computation time of the complete protocols (FDMKC and FDM) over all players. That measure includes the Apriori computation time, and the time to identify the globally s-frequent item sets, as described in later.

2) Total computation time of the unification protocols only (UNIFI-KC and UNIFI) over all players. 3) Total message size. We ran three experiment sets, where each set tested the dependence of the above measures on a different parameter: • N — the number of transactions in the unified database,

FREQUENT ITEMSETS:

We describe here the solution that was proposed by Kantarcioglu and Clifton. They onsidered two possible settings. If the required output includes all globally s-frequent item sets, as well as the sizes of their supports, then the values of Δ(x) can be revealed for all. In such a case, those values may be computed using a secure summation protocol, where the private addend of Pm is suppm(x) – sNm. The more interesting setting, however, is the one where the support sizes are not part of the required output. We proceed to discuss it.

ASSOCIATIO RULES:

Once the set Fs of all s-frequent item sets is found, we may proceed to look for all (s, c)-association rules (rules with support at least sN and confidence at least c). In order to derive from Fs all (s, c)-association rules in an efficient manner we rely upon the straightforward lemma.

ACKNOWLEDGEMENT:

I would like to thank the Mr. Mohammed Shia' Al-Sudani, Minister Of Labour And Social Affairs & Mr. WISAM KADHIM SALMAN DIRECTOR OF THE DEPARTMENT OF MONITORING UNHCR FOR HUMAN RIGHTS & Mr. Salim Juma Abdalsahb Director General in Minister Of Labour And Social Affairs & Miss. Jacqueline S Eleia Director General Of The Social Protection Fund & Mr. Saad Abdullah Hameed Programmer . In IT who gave me enough strength and health to complete this task without any interruption and My Parents , wife , brother and my Sisters who gave me all the support during the project work.

CONCLUSION:

We proposed a protocol for secure mining of association rules in cloud distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency. One of the main ingredients in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players hold. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

BIBILOGRAPHY:

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB, pages 487–499, 1994.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In SIGMOD Conference, pages 439–450, 2000.

[3] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In STOC, pages 503–513, 1990.

[4] M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication. In Crypto, pages 1–15, 1996.

o[5] A. Ben-David, N. Nisan, and B. Pinkas. FairplayMP - A system for secure multi-party computation. In CCS, pages 257–266, 2008.

[6] J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secret secret. In Crypto, pages 251–260, 1986.

[7] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In ASIACRYPT, pages 236–252, 2005.

[8] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In PDIS, pages 31– 42, 1996.

[9] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. IEEE Trans. Knowl. Data Eng., 8(6):911–922, 1996.

[10] T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, 31:469–472, 1985.

[11] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In KDD, pages 217–228, 2002. [12] R. Fagin, M. Naor, and P. Winkler. Comparing Information Without Leaking It. Communications of the ACM, 39:77–85, 1996.

AUTHORS BIOGRAPHY:

Khalid Amjed Mohammed pursuing his Master of Science in Information System, from Nizam College (Autonomous),O.U,Basheer Bagh, Hyderabad,India.

T. Ramdas Naik, Assistant Professor Dept, Computer Science (PG), Qualifications : B.E, MCA,M.Tech,(Ph.D) Nizam College (Autonomous),O.U, Basheer Bagh, Hyderabad, India.