An Improved Apriori Preserving Differential Privacy in the Framework of Spark

Zhiqiang Gao*, Longjun Zhang, Renyuan Hu, Qingpeng Li and Jihua Yang
Department of Information Engineering, University of CAPF Xi'an Shaanxi China
*Corresponding author

***Abstract*—Aimed at the problem that traditional methods fail to deal with malicious attacks under arbitrary background knowledge during the process of massive data analysis, an improved Apriori algorithm preserving differential privacy, combining with Laplace mechanism to mine the pattern of sensitive information in framework of Spark is proposed. Furthermore, it's theoretically proved to meet ε-differential privacy in spark. Finally, experimental results show that guaranteeing availability, our proposed algorithm has advantages over privacy protection and satisfaction in aspects of time as well as efficiency. Most importantly, our algorithm shows a good application prospect in the analysis of data pattern mining preserving privacy protection. Also, it has better ability of privacy protection and timeliness under the premise of ensuring availability.**

Keywords-spark; differential privacy; association analysis; pattern mining; association rule algorithm

I. INTRODUCTION AND BACKGROUND

With the explosive growth of data, the pattern mining analysis of massive data has become a new hot topic in cross disciplinary researches[1]. Under concrete open internet environment, the data mining process or the analysis result is likely to cause disclosure of user's privacy. Therefore, in the era of big data, it is very practical and of theoretical significance to promote the techniques that can deal massive data privacy protection pattern mining under arbitrary background knowledge attack[2].

Nowadays, pattern mining techniques are mainly faced with the following problems under big data environment: 1) Traditional machines are difficult to carry out effective pattern mining algorithm of TB level data in the limited time. 2) Though, providing valuable information, the mining process and count information of pattern mining are likely to leak user's privacy information or malicious attacker will intercept and tamper which may put a serious threat to data security.

In pattern mining, Apriori is the most influential in mining Boolean association rules frequent itemsets. Also, scholars home and abroad have done a lot of fruitful research work [3-7]. In [5], under the frame of Spark, classic Apriori is improved for parallel processing to mine association rules in large data which greatly improved the efficiency of mining massive data association rules. A paralleled Apriori is proposed in [6], which can effectively reduce the generation of key / value pairs with better performance. [7] proposed a paralleled association rules Apriori based on matrix, named MMR, which combines the idea of data partition to simplify the connection step, and simultaneously the affairs of compression so as to further improve the performance of pattern mining.

However, related researches considering both privacy preserving and data mining results are rarely involved. Therefore, in order to satisfy differential privacy, and reduce frequent I/O visits during recursive mining K frequent item-sets, an improved Apriori algorithm preserving differential privacy based on Spark is proposed this paper. During the analysis process of association rules, real support degree added with noise is implemented to reach the threshold of frequent candidate sequences. Then the sequence is divided into two groups: frequent and non-frequent according to the candidate frequent support so as to protect the privacy of user data and to achieve the purpose of being suitable for large data mining platform.

II. PRELIMINARIES

Differential privacy is a definition giving a strong privacy guarantee even in the presence of auxiliary information. Attackers can use a variety of ways to obtain the background knowledge of data privacy during the process of pattern mining or results publishing. Regardless of the size of query or analysis results, differential privacy (DP) [8] is able to add noise into data set to achieve privacy protection, which can overcome the shortcoming of traditional security models unable to meet the maximum background knowledge attacking. DP is defined as follows:

Definition 1. A randomized function A gives ε-differential privacy if for all datasets D_0 and D_1 differing on at most one row, and all S ⊆ Range(A),

$$P_r[\mathrm{A}(D_0) \subseteq S] \leq e^{\varepsilon} P_r[\mathrm{A}(D_1) \subseteq S] \quad (1)$$

where the probability space in each case is over the coin flips of A. The parameter ε is public, and its selection is a social question. We tend to think of ε as, say, 0.01, 0.1, or in some cases, ln2 or ln3[9].

The global sensitivity of the query function F is a measure of the impact of the change of a single record on the F output, which is defined as follows:

Definition 2. For $F : D \rightarrow R^d$, the L1 sensitivity of F is

$$\Delta F = \max_{D_0, D_1} \| F(D_0) - F(D_1) \|_1 \tag{2}$$

for all D_0, D_1 differing in at most one row.

The Laplace distribution with parameter b, denoted Lap(b), has density function

$$p(x) = \frac{1}{2\Delta F/\varepsilon} exp(\frac{|x|}{\Delta F/\varepsilon}) \tag{3}$$

This distribution has highest density at 0 (good for accuracy). Finally, the distribution gets flatter as ε decreases: smaller ε means better privacy, so the noise density should be less "peaked" at 0 and change more gradually as the magnitude of the noise increases.

In this paper, we can better maintain the original data statistical characteristics of the Laplace mechanism, the principle is as follows:

Theorem 1. [7] For $F : D \rightarrow R^d$, the mechanism K that adds independently generated noise with distribution Lap(Δf/ε) to each of the d output terms enjoys ε-differential privacy.

In addition, differential privacy preserving has sequence composition and parallel combination [11].Based on the properties, this paper proves that our improved algorithm can satisfy the differential privacy and guide the allocation process of the privacy budget.

III. APRIORI DP ALGORITHM IN SPARK

A malicious attacker can intrude into the pattern mining at any stage of the analysis, therefore, this paper put forward a improved Apriori to satisfy the differential privacy protection. through adding Laplace noise to real support degree to protect the frequent item sets under the background knowledge of any privacy preserving association analysis.

A. Apriori DP Algorithm Design

As an association rule analysis, Apriori can be divided into two steps: mining frequent itemsets and generating association rules. The key part is calculating the support of the frequent item sets. In order to overcome the shortcomings of the Apriori, which requires multiple iterative computation and lack of privacy protection, a Apriori DP algorithm which is suitable for the Spark framework is proposed in this paper. Flow chart of Apriori DP under the Spark algorithm is shown in Figure I.

The main steps are as follows:

Step 1 converts the initial data set obtaining from HDFS into the RDD data set;

Step 2 seek out the one dimensional frequent itemsets from the original data set RDD and then code;

Step 3 encode and group the one dimensional frequent item set for tree-building operations;

Step 4 in order to achieve the purpose of privacy protection, add the Laplace noise disturbance to the count of the support of the item set for each FP tree for frequent item set mining;

Step 5 integration of each FP tree in a new final RDD data sets, then save the final frequent item set by SaveAsTextFile method;

Step 6 output.

In step 1, 2 and 3, RDD data sets are grouped and each node in spark cluster corresponds to a FP Tree; in step 4, add differential privacy to the frequent item set support count in order to improve protection efficiency and mining accuracy; in step 5, memory computing mode of spark is adopted to avoid frequent I/O communication in HDFS and improve implementation efficiency.

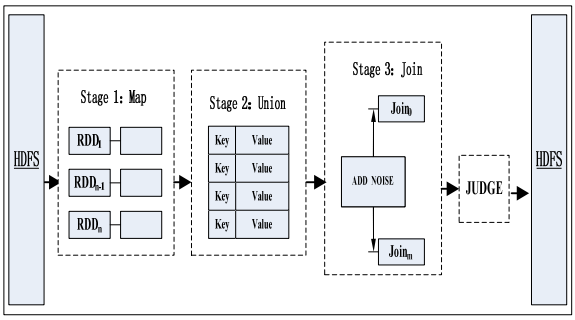

FIGURE I. FLOW CHART OF APRIORI DP IN SPARK

B. Privacy Analysis

From flow chart of Apriori DP under the Spark, data are formed in flexible distributed data set RDD. Apriori DP algorithm adds frequent item set support counts with Laplace noise disturbance in FP Tree to achieve privacy and data protection.

Because each iteration during Apriori DP based on memory Calculation meets privacy series combination properties, so epsilon privacy budget properties are satisfied. In addition, by using the method in literature [10], privacy budget allocation scheme in each iteration is $\varepsilon_t = \varepsilon/2^t$.

Analysis shows that [10], global sensitivity of frequent item count support degree is d+1, where D is the dimension of the data set. Therefore, in each iteration, DP Apriori under the Spark can satisfy ε-differential privacy by adding Lap(d+1)$2^{t+1}/\varepsilon$ Laplace noise. In this paper, adding noise disturbance to the frequent item count support, increase security and improve the iterative efficiency, reduce the impact of excessive noise on the analysis of the results of association rules.

IV. EXPERIMENT AND ANALYSIS

Experimental platform is comprised of four distributed nodes. configuration is as follows: operating system CentOS 6.4, CPU 3.30GHz and memory 16GB and 320GB hard disk. 1000Mb/s network Ethernet communication, spark on yarn: Hadoop 2.6.4 and spark 2.0 [11]. Experiment data is randomly

generated by the IBM database generator. Compared Apriori runs on MapReduce platform.

In this paper, the efficiency is verified by the Speedup acceleration ratio [11], as shown in the formula (4), where Ts is the running time on a single machine, and Tc is time consuming is performed in the cluster.

$$Speedup=T_s/T_c \quad (4)$$

The result of speedup is shown in the Table I.

TABLE I. EXPERIMENTAL RESULTS

Algorithms	Speedup	
	3 nodes	*4 nodes*
A1	3.54	5.63
A2	1.82	2.79

(A1: our algorithm, A2: compared algorithm)

Table I shows, our algorithm in this paper accelerates ratio significantly and outperforms the compared algorithms. with the cluster node number increases, verify the spark memory computing mode in the iterative computation efficiency is better than the MapReduce framework and can greatly improve parallelism. Moreover, the Spark memory computing framework can reduce the frequent I/O communication with HDFS with excellent operating speed. In addition, in order to ensure the availability of the mining results, integrated into the privacy protection mechanism, the efficiency of the algorithm is not significantly affected.

As shown in Figure II, when level of privacy protection budget ε is low, association rules mining is relatively available. when ε reaches a certain degree, results tend to be stable. It is also verified that ε can balance effect of privacy and usability.

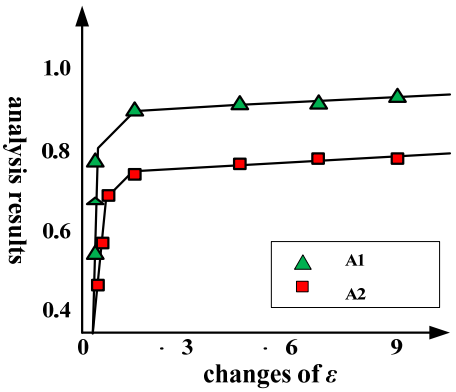

FIGURE II. ASSOCIATION RULES ANALYSIS RESULTS WITH THE CHANGES OF E

V. CONCLUSIONS

The spark memory parallel computing technology to achieve ε-differential privacy, the improved Apriori algorithm, combined with the Laplace mechanism of frequent item counting process, the most crucial privacy information of noise disturbance. At the same time, the performance of improved algorithm is obviously better than that of the traditional association rule analysis algorithm based on MapReduce framework. In the future, the research work will mainly focus on the premise of ensuring the level of privacy protection, through more efficient protection mechanism to reduce the noise and reduce the complexity of the algorithm.

ACKNOWLEDGMENT

Supported by the National Natural Science Foundation of China (61309008); the natural science foundation of Shaanxi province (2014JQ8049)

REFERENCES

[1] Feng Dengguo, Aman Chang, Li Hao, big data security and privacy protection [J]. Journal of computers, 2014,37 (1): 246-258.

[2] Fang, Jia Yan, Li Aiping, et al. Large data privacy protection technology of [J]. data, 2016 (1).

[3] Chanchalyadav, Shuliang Wang, An approach to improve Apriori algorithm based on association rule mining.[C] 2013 4th Inernational Conference Computing working technology, USA.

[4] Mohammadhossein Barkhordari, Mahdinia Manesh. An effective MapReduce- based association rule mining method [C]. 2014 the sixteenth International Conference on Electronic Commerce, Philadelphia.

[5] Niu Hailing, Lu Huimin, Liu Zhenjie, an improved Spark algorithm based on Apriori [J]. Journal of Northeast Normal University (NATURAL SCIENCE EDITION): 2016, 48 (1): 84-89.

[6] Huang Liqin, Liu Yanhuang, MapReduce based parallel Apriori algorithm improvement research [J]. Journal of Fuzhou University (NATURAL SCIENCE EDITION) 2011, 39 (5): 34-39.

[7] Xie Zhiming, Wang Peng, a parallel matrix Apriori algorithm based on Reduce Map architecture [J], computer application research, 34 (1): 17-21.

[8] DWORK C. A Firm Foundation for Private Data Analysis[J]. Communications of the ACM, 2011, 54(1):86-95.

[9] Xiong Ping, Zhu Tianqing, Wang Xiaofeng. Differential privacy protection and its application [J]. Journal of computer science, 2014, 37 (1): 101-122.

[10] Li Hongcheng, Wu Xiaoping, Chen Yan. Reduce clustering method for differential privacy protection under the framework of Map [J]. K-means Journal of communication, 2016, 37 (2): 124-130.

[11] Wang Baoyi, Wang Dongyang, Zhang Shaomin. Short term distributed power load forecasting algorithm based on Spark and [J]. IPPSO_LSSVM electric power automation equipment, 2016,36 (1): 117-122.