A K-means Approach for Map-Reduce Model and Social Network Privacy Protection

Shou-Lin Yin¹ and Jie Liu 1,2*

¹Software College Shenyang Normal University No.253, HuangHe Bei Street, HuangGu District, Shenyang, P.C 110034 - China 352720214@qq.com

 ²Department of Information Engineering Harbin Institute of Technology
 92 West Straight Street In Nangang District of Harbin Harbin,150001-China *Corresponding author:nan127@sohu.com

Received April, 2016; revised July, 2016

ABSTRACT. As we all know, traditional privacy protection methods are not always reliable. In this paper we propose a new social network privacy protection based on a new Map-Reduce model with a k-means approach. Main task controls k-means to start iterative execution. Mapper sub-task independently computes the distance between each record and cluster center, then tags them. Reducer sub-task calculates the sum of the record number in the same cluster and attributes vector. And we use the noise disturbance generated by Laplace. Through above processes, we can realize privacy protection. The experimental results show that the new method provides privacy and timeliness, and ensures the good usability.

Keywords: Privacy protection, Map-Reduce model, K-means, Noise disturbance, Laplace.

1. Introduction. Data mining[1-2] is an important way to obtain information. Useful information can be collected from big data. Clustering analysis[3-4] is a typical nondirective learning method in data mining, its main idea is that it divides data into several types. There is a minimum differentiation among all the data in every cluster. However, the differentiation between clusters is the biggest. Therefore, clustering method is used widely, such as in network intrusion detection[5], large-scale location[6] and market segmentation[7] et al,. Under the background of big data, there are two problems in cluster technology: one is that data volume is increasing, a single machine is difficult to effectively make data cluster analysis in an acceptable time [8]. So it needs to use parallel distributed computing resources for quick clustering analysis. Another is that the results of data cluster analysis can provide valuable information, but it may leak a single record information in data set, which poses a threat to data privacy. In the big data time, attackers own more knowledge that helps them easily steal data privacy. So it requires to study various data privacy protection methods under any environment.

Social network [9-10] has attracted more attention as an emerging Internet application. The data in social network plays an important role in predicting economy and public opinion analysis. But there are much personal privacy information in data, when we directly release data, which may cause the leaking of user' privacy. Hence, we must effectively protect user data privacy when publishing data information. Data privacy protection includes many aspects, protection methods, parallelism and dynamics [11]. Privacy protection methods contain clustering, data perturbation, generalization, randomization and inference control etc. Parallelism reflects that massive social network data analysis and processing needs efficient parallel algorithms to implement [12]. Meanwhile, most existing privacy protection methods are only suitable for a single data releasing, namely it can not be changed after releasing. The development of social network determines dynamics of the social network. Static privacy protection methods cannot guarantee the dynamic social network privacy.

1.1. Related work. Anchalia [13] proposed an improved and efficient method to implement the k-means clustering technique using the Map-Reduce paradigm, whose main idea was to introduce a combiner in the mapper function to decrease the amount of data to be written by the mapper and the amount of data to be read by the reducer which had considerably reduced the redundant Map-Reduce called that had resulted in a significant reduction in the time required for clustering as it had decreased the read/write operations to a large extent. Wang[14] presented an improved Map-Reduce model, named Map-Check-Reduce (MCR) which could terminate the map process when the requirements of imprecise applications were satisfied. Compared to Map-Reduce, a Check mechanism and a set of extended programming interfaces were added to MCR. The Check mechanism could receive and analyze messages submitted by mappers and then determined whether to terminate the map phase. The programming interfaces also could be used by the programmers of imprecise applications to define the termination conditions of the map phase. Viswanath[15] analyzed the dynamic of social network and proposed that anonymous methods should meet the feature of the network data changing over time. Chen[16] used node identifier to deal with multiple releasing of dynamic network, but he didnt give the method to delete or add node.

Therefore, we propose a new Map-Reduce model with k-means clustering of differential privacy protection, which adopts distributed computing functions provided by Map-Reduce model to improve clustering analysis efficiency. And we also add random noisy, which can make the results of clustering meet differential privacy. At last, the experiments show that our new method has higher security and can provide serious protection for data privacy. The following are the structures of this paper. In section2, we detailed introduce the new model and we also explain differential privacy. Section3 is the experiments part, which is used for demonstrating our method. There is a conclusion is section4.

2. K-means approach for Map-Reduce model.

2.1. Differential privacy protection. Differential privacy protection [17-18] is a kind of privacy protection technology based on the part of information hiding, it implements information interface through random response and random noise add, at the same time, output information after interface can keep its original statistical properties to a certain extent. So the availability of the data mining results can be kept in an acceptable range.

Differential privacy technology gives a strict and provable privacy protection definition. It guarantees that variable quantity of query results are very small when changing any record in data set. Attackers still cannot analyze any information of target record with all the extra information except target record. Therefore, this method can deal with malicious analysis under any background knowledge. Basal principle of differential privacy protection is that operation of users extracted from data set D is defined as query F, algorithm A randomly processes the output of query F. And the results can meet the condition of differential privacy protection [19].

Theorem 2.1. Supposing that data set D and D' are the same or only have one different record. Random(A) is domain of one random algorithm A. Pr[X] is probability of event X occur. For any $S \in Random(A)$:

$$Pr[A(D) \subseteq R_A \le e^{\varepsilon} Pr[A(D') \subseteq R_A]] \tag{1}$$

Random algorithm A provides ε -differential privacy protection. Where ε is privacy protection budget. Global sensitivity is an important inherent attribute for the query function, it reflects the impact of a single record change on query function output. Global sensitivity is defined as :

$$\Delta F = max_{D,D'} ||F(D) - F(D')||.$$
(2)

Where $|| \cdot ||$ denotes sum of absolute values of each vector element. Implement mechanism of differential privacy protection includes Laplace and Index which adopts randomly noise adding and randomly response to implement privacy protection respectively. Laplace method is suitable for protecting numerical results [20], its main idea is as follows.

Theorem 2.2. For query F, data set D, if query output is F(D), global sensitivity of F is ΔF and noise Y obeys Laplace distribution whose scale is $\frac{\Delta F}{\varepsilon}Lp$, then algorithm A(D) = F(D) + Y meets ε -differential privacy protection. Probability density function of random noise $Lap(\frac{\Delta F}{\varepsilon})$ is :

$$p(x) = \frac{1}{2(\Delta F/\varepsilon)} e^{-|x|/\frac{\Delta F}{\varepsilon}}$$
(3)

In addition, differential privacy protection has two combined characteristics including sequence combination and parallel combination which plays an important role in proving that whether the algorithm meets differential privacy or not.

Definition 2.1. Supposing that *m* random algorithms A_1, \dots, A_m . $A_i (1 \le i \le m)$ provide ε -differential privacy protection. For the same data set *D*, sequence combination of A_1, \dots, A_m in *D* provides ε -differential privacy protection and $\varepsilon = \sum_{i=1}^m \varepsilon_i$.

Definition 2.2. Supposing that algorithm A and data set D, D is divided into disjoint subset D_1, \dots, D_n . If A provides ε -differential privacy protection, then sequence combination of A in D provides ε -differential privacy protection.

2.2. Map-Reduce model with k-means clustering. Under the Map-Reduce distribute environment, the new method can guarantee that when any record is changed in data set, the change of every clustering center and record number cannot leak privacy information. In other words, malicious analyzer cannot use the data similar to original data to get the initial single record privacy information through mining. Attack model is as figure 1.

The basic idea of our method is that we use Map task in distributed computing nodes to judge the clustering type of each record and then adopt Reduce task to calculate sum between record number in clustering and corresponding attributes. Finally, we add moderate Laplace noise, which makes clustering analysis meet ε -differential privacy protection.

Traditional differential privacy protection k-means algorithm has the low accuracy in that randomly selecting initial center results in a low convergence speed. The increasing of iteration number, however, causes the adding of noise. What's more, initial center with noise often deviates from the original center far away. So in this paper, we not only make Map-Reduce parallel design for differential privacy protection k-means algorithm, but adopt improved initial center selection and add noise method.

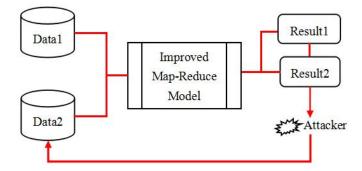


FIGURE 1. Attack model

2.3. Detailed algorithm process. Supposing that total record number is N in data set, each record is $a_i(1 \le i \le N)$, dimension of every record is d. These records are divided into M data slices (written as $D_j(1 \le i \le M)$). Clustering number is K, the center of each clustering is $u_k(1 \le i \le K)$.

- 1. Step 1. Main task D normalizes each record into space $[0, 1]^d$. N records a_1, \dots, a_N can be divided into K subsets C_1, \dots, C_K , record number $|C_k| \leq ceil(N/K)$. Where $ceil(\cdot)$ denotes round up the value function. Calculating sum sum_k^0 between number of records num_k^0 and attribute vector of each record in C_K , and we add random noisy into num_k^0 and $sum_{k'}^0$, get $num_{k'}^0$ and $sum_{k'}^0$ respectively. We calculate $u_{k'}^0 = \frac{sum_{k'}^0}{num_{k'}^0}$. So $u_{k'}^0$ is the initial clustering center.
- 2. Step 2. Main task divides all the data record into M data slices, it designates M sub-tasks to execute Map operation and designates K sub-tasks to execute Reduce operation.
- 3. Step 3. Mapper sub-task receives data slices with N/M records and runs Map function. It will calculate the distance from each record to cluster center and select cluster center with the smallest distance.
- 4. Step 4. Reducer sub-task receives all $\langle key, value \rangle$ belonging one cluster center and runs Reduce function. It calculates the number of cluster sum *sum* between number of records *num* and attribute vector of each record and adds random noise into *sum* and *num*. At last, it calculates cluster center u' with noise.
- 5. Step 5. Main task receives output u' of each Reduce node. It respectively calculates the current and previous the distance of K cluster centers. If the distance of center attribute vector difference is less than threshold, then algorithm will stop and output each cluster center and number of record in cluster. Otherwise, repeat step3 to step5.

2.4. Analysis of privacy. Privacy of new Map-Reduce model is implemented by adding Laplace noise into num_k and sum_k in Reduce operation. Each iteration of k-means algorithm is equivalent to sequence combination of random algorithm. So according to definition 1, privacy protection budget is:

$$\varepsilon = \sum_{t=1}^{T} \varepsilon_t. \tag{4}$$

Where T is total number of iteration, ε_t is the t - th privacy protection budget. For budget, each iteration costs half of the rest of privacy budget. So at t - th iteration, the privacy budget is $\varepsilon_t = \frac{\varepsilon}{2^t}$.

In each iteration, K Reducer sub-nodes execute operation independently. Results of each iteration are equivalent to parallel combination of Reduce operation. Therefore, according to definition 2, whether t - th iteration meets ε_t -differential privacy depending on that operation of each Reducer sub-task meets ε_t -differential privacy under distributed environment.

Therefore, according to Theorem2, we add random noise $Lap(d+1)2/\varepsilon$ into $num_{k'}^0$ and $sum_{k'}^0$ respectively in the process of calculating initial center, and add $Lap(d+1)2^{t+1}/\varepsilon$ into num_k and sum_k in t-th iteration. That can guarantee new k-means method under Map-Reduce model satisfying ε -differential privacy protection. Compared to original differential privacy protection k-means algorithm, improved calculating initial center can reduce iteration number at the same privacy protection budget, further it reduces adding of random noise.

3. Experiments and analysis. In order to verify the effectiveness of differential privacy protection with new Map-Reduce model, we select the following reference standards to make a comparison with our method. Standard classification is true in "Blood" and "Gramma" database, so we choose data set from the two sets. What's more, we make a comparison with reference [21].

There are two usability indicators measuring data mining results: recall and precision. But F-measure can integrate recall and precision, hence we use F-measure to demonstrate cluster usability. The greater F-measure is, the similarity stronger two clusters results have. That is to say, our algorithm with adding noise has a small effect on cluster usability.

This paragraph introduces the calculation for F-measure. We use CR as the cluster results of reference standards, CR' as the cluster results of our method, cluster number is K. U_i is the $i - th(i \le i \le K)$ cluster set in CR. V_i is the $i - th(i \le i \le K)$ cluster set in CR'. Ce_i is the coincident record number between U_i and V_i . $|U_i|$ and $|V_i|$ is the record number in U_i and V_i respectively. The precision and recall of i - th cluster is P_i and R_i respectively. Therefore,

$$R_i = \frac{Ce_i}{|U_i|}.$$
(5)

$$P_i = \frac{Ce_i}{|V_i|}.\tag{6}$$

Then we calculate weighted harmonic mean value F_i of P_i and R_i .

$$F_i = \frac{2R_i P_i}{R_i + P_i}.$$
(7)

Finally, we compute weighted mean value for F_i of every cluster. Assuming that NT is the data set record total number, so the usability measurement F-measure of cluster result is:

$$F - measure = \sum_{U_i \in CR} \frac{|U_i|}{NT} F_i.$$
(8)

We write the similarity between our method and standard data set as F1, as well as similarity between our method and [21] as F2. In each running process, the added noise obeys the Laplace random distribution, so the results have randomness. For the cluster results, we select average value of ten times. For "Blood" and "Gamma", when differential privacy protection ε changes, F1 and F2 are as figure2 and figure3. (F is the usability measurement F-measure value).

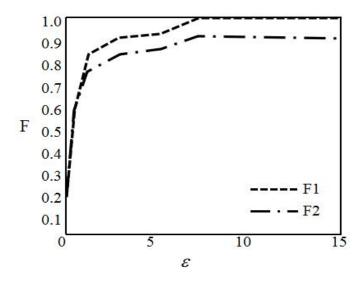


FIGURE 2. F changes with different ε in "Blood"

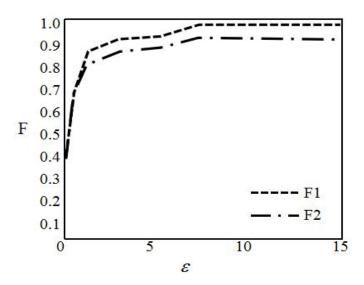


FIGURE 3. F changes with different ε in "Gamma"

From fig2 we can know that, when ε is small ($\varepsilon < 7$), F will remarkably increase. When $\varepsilon > 7$, the results begin to flatten. Fig3 shows that F will reach to a high level in a short time with our method when $\varepsilon > 3$. Therefore, the new scheme is a better choice for privacy protection.

4. **Conclusions.** This paper utilizes Map-Reduce model to realize parallel distributed k-means cluster and at the same time uses Laplace to implement difference privacy protection of the algorithm. Finally, it improves the timeliness and privacy of k-means algorithm. This new method greatly perfects social privacy safety. In the future, we will study more advanced algorithms to reduce Map-Reduce Job running time. And we will

learn that how can we use integrated index mechanism and Laplace mechanism to reduce the disturbance of each iteration implementation.

Acknowledgment. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

- T. Menzies. Data Mining, Recommendation Systems in Software Engineering, Springer Berlin Heidelberg, pp.39-75, 2014.
- [2] C. M. Keet, A Lawrynowicz., C. D'Amato, et al. The Data Mining Optimization Ontology, Web Semantics Science Services & Agents on the World Wide Web, vol 32, pp. 43-53, 2015.
- [3] P. L. Lin, P. W. Huang, A. Lai, et al. An efficient clustering analysis method for image segmentation with noise International Conference on Machine Learning & Cybernetics. IEEE, pp. 493-498, 2015.
- [4] Ye. F. Fuzzy Granules Clustering Analysis Based on Quotient Space Journal of Nature, 2015, vol. 485, no. 7397, pp. 177-178, 2015.
- [5] L. I. Wen-Hua. Network Intrusion Detection Model Based on Clustering Analysis. Journal of Computer Engineering, vol. 37, no. 17, pp. 234-237 2011.
- [6] L. Lin, P. Capak, C. Laigle, et al. Clustering analysis and large-scale environments of galaxies out to z 3 in SPLASH, [J]. Iau General Assembly, pp. 22, 2015.
- [7] E. Brentari, L. Dancelli, M. Manisera, Clustering ranking data in market segmentation: a case study on the Italian McDonald's customers' preferences [J]. Journal of Applied Statistics, pp.1-18, 2016.
- [8] L. Cao, C. Zhang, T. Joachims, et al. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]// ACM SIGKDD, 2015.
- [9] Y. Zhang, E. Pan, L. Song, et al. Social Network Aware Device-to-Device Communication in Wireless Networks, [J]. IEEE Transactions on Wireless Communications, vol. 14, no. 1 pp.177-190, 2015.
- [10] M. P. De, E. Ferrara, D. Rosaci, et al. Trust and compactness in social network groups, [J]. IEEE Transactions on Cybernetics, vol. 45, no. 2, pp.205-216, 2015.
- [11] X. Y. Liu, B. Wang, X. C. Yang, Survey on Privacy Preserving Techniques for Publishing Social Network Data, [J]. Journal of Software, 2014.
- [12] Y. Wang, L. Xie, B. Zheng, et al., High utility K-anonymization for social network publishing, [J]. Knowledge & Information Systems, vol 41, no. 3, pp. 697-725, 2014.
- [13] Anchalia, P. Prajesh, Improved MapReduce k-Means Clustering Algorithm with Combiner, [C]// Uksim-Amss International Conference on Computer Modelling and Simulation. IEEE, pp.386-391, 2014.
- [14] C. Wang, Y. Peng, M. Tang, et al., MapCheckReduce: An Improved MapReduce Computing Model for Imprecise Applications, Proc. of International Congress on Big Data (BigData Congress). IEEE Computer Society, pp. 366-373, 2014.
- [15] L. Holleran, M. Ahmed Anderson-Schmidt. H, et al. On the evolution of user interactions in facebook, [J]. Proceedings of the Acm Workshop on Online Social Networks, vol. 39, no.4, pp.37–42, 2009.
- [16] J. Cheng, M. C. Fu, J. Liu, K-isomorphism: Privacy preserving network publication against structural attacks, Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010, pp. 459-470, 2010.
- [17] C.Lin, P. Wang, H. Song, et al., A differential privacy protection scheme for sensitive big data in body sensor networks, [J]. annals of telecommunications - annales des tlcommunications, pp.1-11, 2016.
- [18] Y. Wang, Preserving differential privacy in complex data analysis, [J]. Dissertations & Theses -Gradworks, 2015.
- [19] X. Ping, T. Q. Zhu, X. F. Wang, A Survey on Differential Privacy and Applications, [J]. Chinese Journal of Computers, vol.37, no.1, pp.101-122, 2014.
- [20] L. P. Ding, G. Q. Lu, Survey of differential privacy in frequent pattern mining, [J]. Journal on Communications, vol. 35, no. 10, pp. 200-209, 2014.
- [21] S. Gopalani, R. Arora, Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means, [J]. International Journal of Computer Applications, vol. 113, no. 1, pp. 8-11, 2015.