A new approach for k-anonymity based on tabu search and genetic algorithm

論 文
10-4-3

Cui Run, Hyoung-Joong Kim*, and Dal-Ho Lee

Abstract

Note that *k*-anonymity algorithm has been widely discussed in the area of privacy protection. In this paper, a new search algorithm to achieve *k*-anonymity for database application is introduced. A lattice is introduced to form a solution space for a *k*-anonymity problem and then a hybrid search method composed of tabu search and genetic algorithm is proposed. In this algorithm, the tabu search plays the role of mutation in the genetic algorithm. The hybrid method with independent tabu search and genetic algorithm is compared, and the hybrid approach performs the best in average case.

Keywords : database, privacy, heuristic, algorithm

I. INTRODUCTION

Sometimes organizations have to publish micro-data for special usages such as statistical analysis and health condition research. In order to protect individual privacy, known identifiers (e. g., name or social security number) must be removed. However, that is not enough. In addition, this process must consider about the possibility of combining certain other attributes with external data to uniquely identify individuals. Such kind of attributes combination, called quasi-identifiers, can also locate the individual because they make up a unique mark to be distinguished. For example, an individual might be re-identified by joining the released data with another (public) database on age, sex, and salary.

The k-anonymity model is just one of the most popular ways to solve the privacy protection problem. It provides modification to the tuples in the database to remove the quasi-identifiers. It makes sure that each record is indistinguishable from at least other records. The idea is simple, but it has been proved that how to get k-anonymity property in the database is an NP hard problem. That is, k-anonymity clustering is quite challenging.

Up to now, there are many excellent researches in k-anonymity area [1-5]. Most of them are based on the analysis of the data and try to find some patterns for the data in order to achieve k-anonymity fast. However, for the complexity of the data, there is no “common” model in most cases, which means that information loss is brought in with the pattern achieved higher after the modification applied to the database. To solve the problem above, some self-adoptive methods are invented. One of the efficient examples is heuristic algorithm [6-8]. As thek-anonymity is an NP hard problem, heuristic algorithm is quite suitable to be applied in such cases. However, it is a pity that most of the heuristic including genetic algorithm pay attention to the data modification in record level other than full domain. This provides more chances for the analysts to get useful information from the modified records. In such a

접수일자 : 2011년 10월 05일
심사일자 : 2011년 11월 20일
최종완료 : 2011년 12월 17일
*교신저자, E-mail : khj-@korea.ac.kr

kind of method, two similar records can have a high probability to be modified to different internals and such a situation introduces higher distinguishable ability for the two processing records. Full domain consideration is important in k-anonymity problem.

The conception of lattice is introduced into this area in [9] and [10] in order to enhance the full domain modification property. Each node in a lattice in k-anonymity represents a way of modification. With a lattice, how to find a suitable solution has been changed into how to search in the lattice node space to get a suitable node. In [11], the authors provide an efficiency binary search algorithm (OLA) to find the optimal node in lattice space. However, it focuses on lattice space with monotonic property. If the nodes are not monotones, it can also give good solutions but no support in theory.

Based on the previous work mentioned above, we provide a new heuristic search method in lattice solution space; this approach is a combination of traditional tabu search method and genetic algorithm. It inherits the strong "climbing" ability of tabu search and the multiple start point property of genetic algorithm. We compare the performance of this new approach separately with the tabu search and genetic algorithm. This paper shows our method performs better in most of cases.

Some basic concepts are described as follows:

Quasi-Identifier Attribute Set: *A quasi-identifier set Qis a minimal set of attributes in table T that can be joined with external information to re-identify individual record(with sufficiently high-probability)* [3].

Equivalence Class: *A table T consists of a multi-set of tuples. An equivalence class for T with respect to attributes X1, …, Xdis the set of all tuples in T containing identical values (x1, … , xd) for X1, …, Xd* [4].

k-Anonymity Property: *Relation T is said to satisfy the k-anonymity property (or to be k-anonymous with respect to attribute set Q if every count in the frequency set of T with respect to Q is greater than or equal to k* [3].

Information Loss: *It represents the difference between the original and modified record. There are many different ways to calculate the value of this variable.*

In Section II, we explain how to construct a lattice solution space for the database and introduce some necessary preparations such as information loss. In Section III, the details of the hybrid algorithm are shown and in Section IV experiment results are given to compare the new method with other traditional heuristic algorithms such as genetic and tabu method. Finally, we give the conclusion for our work and provide some possible directions on how to improve the method in the future.

II. Lattice generation and evaluation matrix

A. Value Generalization Hierarchies

Lattice in the k-anonymity problem is based on the conception of value generalization hierarchies. It is a tree structure and its internal nodes are intervals and its leaves are values appeared in the corresponding attribute.

Each value generalization hierarchy is corresponding with one of the attributes in the database. The leaves are really values appeared in this column and the upper node is interval which can contain all the values or intervals in its sub-node. All the interval nodes located in the same height of the tree are disjoint. Here is an example. Consider the attribute Age, where the possible values appeared in the

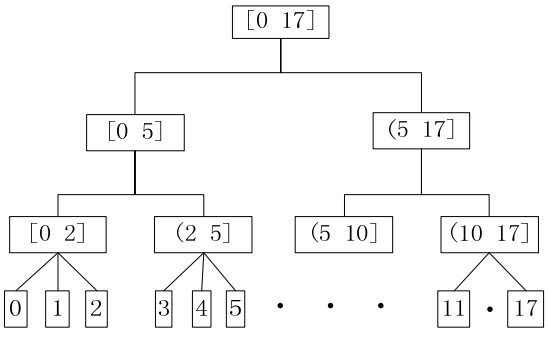

Fig. 1. Example of Value Generalization Hierarchies

column are 1 to 17. The corresponding value generalization hierarchy is shown inFig. 1.

B. Lattice Generation

Samarati [12] and Sweeney [13] have formulated mechanisms for k-anonymity property using the ideas of generalization and suppression. In their works, they showed basic knowledge about the construction witha small scale example. However, in heuristic method, we can process much more attributes at the same time by increasing the size of candidate solutions easily. The variables we use for the construction of lattice solution space are as follows:

Table. 1. Somedefinitions in the paper

N: the amount of attributes we use in the database.
A_i: the ith attribute in the database. $1 \le i \le N$.
H_i: the height of the value generalization hierarchy for A_i.
L: the lattice set for solution nodes.
T: temporary set for storing nodes.
Mark: variable to mark the levels of the nodes in lattice
NeighborSet: this set is defined as follows: for a node $[x_1,x_2,\cdots,x_n]$, we check the node $[x_1-1,x_2,\cdots,x_n]$, $[x_1,x_2-1,\cdots,x_n]$, $\cdots[x_1,x_2,\cdots,x_n-1]$. Among these nodes, if one of the values inside is less than 0, we delete it. All the elements remained after checking form the *NeighborSet*of node checked

Here we show how to get the lattice of solution space in k-anonymity problem.

Table. 2. Algorithm: Lattice-Construction

1. $L = \{[H_1, H_2, \ldots, H_N]\}$, *Mark* = 1, assign the *Mark* value to all the nodes in $L, T = \Phi$
2. Find all the nodes with highest level value in L. For each node, calculate its *NeighborSet* and add the set to T.
3. IF T is not $[0,0,0,\ldots,0]$
   a) *Mark* = *Mark* + 1,
   b) For each Node x in T, find its *NeighborSet*, add the set to L, assign the value of *Mark* to their level value.
   c) Go to 2.
4. LSE, add $[0,0,0,\ldots,0]$ to L
5. END algorithm.

After applying the algorithm above to the attributes, we can get a solution space in the form of lattice.

C. Suppression

In full domain k-anonymity method, suppression is used to deal with the remained records in small size equivalent classes All these small scale data will be suppressed in one equivalence class to achieve k-anonymity. As the suppression can introduce higher information loss, we must control the percent of the suppression in the full domain data modification. We prefer to use value generalization hierarchies, but we cannot avoid suppression.

D. Evaluation Matrix

There are many different kinds of Information Loss Matrix, but there is no common rule for that. Thus, users can choose any kind of traditional Information Loss Matrix or design their own method. In this paper, to achieve a non-monotonic property, we adopt the Discernability Metric (DM) as in Eq. 1, where fiis the size of equivalence classes:

$$DM = \Sigma_{f_i \ge k}(f_i)^2 + \Sigma_{f_i < k}(f_i \cdot n) \quad (1)$$

DM value means the distinguishability or the records. If DM is bigger, it represents that we can distinguish data more easily and less information loss occurs. In this case, only few data modification is required.

III. Hybrid method details

Now we will introduce our new search method in the lattice space. This method is a combination of tabu search method and genetic algorithm.

The traditional tabu search has a strong capability of “climbing”. It can jump out of the local optimal solution and has a possibility to achieve the best solution point. However, this climbing capability is highly limited by the starting point of the tabu search. For some problems of tabu search which gets the starting point with greedy method, it always shows that the starting point is optimal. As we use a good

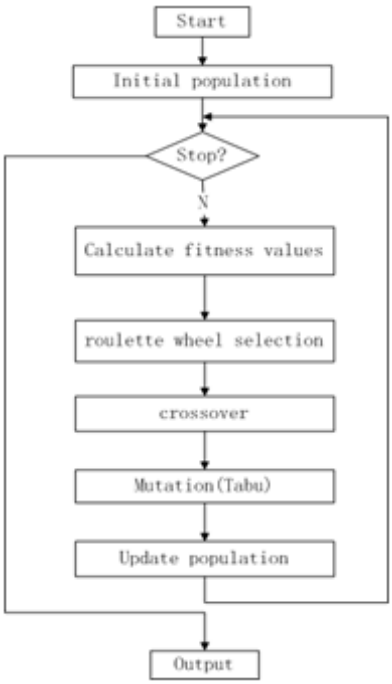

Fig. 2. Flowchart of Genetic Algorithm

Fig. 3. Flowchart of combined method

starting point search strategy, it may limit the chance to achieve the really optimal solution at the same time. For genetic method, it owns a good multiple starting point properties. With the large size of populations, it can also find a very good solution but with high level of randomness. If you are lucky enough, you will find a much better solution than you imagine. It cannot produce a stable best or nearly best solution for the problem processing.

In our method, a tabu search is embedded into the traditional genetic algorithm to implement a local search from multiple starting points getting from genetic method. Here, the tabu search performs the role of mutation in the genetic algorithm.

Fig. 2shows the basic framework of the genetic part in our approach and the framework of the tabu part is shown inFig. 3.

In Genetic Algorithm part, the setting for each part is as follows:

Initial solution: We randomly generate an initial population which serves as multiple start points of the whole algorithm within lattice space. A uniform distribution is used in random procedure.

Stop condition: A value of 100 is appointed as the limitation of the cycle time. If it arrives at the limitation, the algorithm will stop and output the best solution node that can be achieved.

Fitness value: Each node in lattice space corresponds with a strategy of database modification. In this paper, DM value becomes larger with less information loss; we can use the inverse of DM value to evaluate the information loss. The fitness is defined as follows:

$$fitness = 1 / DM \tag{2}$$

where DM value is defined inEq 1.

Roulette wheel selection: This part is a traditional strategy to choose the candidates with the fitness value. Assuming each node, x, fxis the corresponding fitness value. Then, in the selection wheel, it will own a chosen probability Px defined as:

$$P_x = f_x / (\sum fitness\ of\ initial\ population) \tag{3}$$

Then, we can choose the number of candidates randomly chosen according to the possibility as above.

Crossover: In this part, we randomly choose pairs from the candidates. Cut the pairs from a random point between two attributes and exchange the second halves of the solutions to get a new pair of solutions. The procedure will be repeated until we

get enough new candidates.

Update population: After the tabu part, which performs the role of mutations in the genetic algorithm, we will get a new group of solutions. The original population will be replaced by a new group of solutions in this step.

The procedures in the tabu search part are shown as follows:

Initial point: In this step, for each node passed by crossover part, the tabu search will deal with it as a starting point and search the local area around it.

Generation limitation check: The circle time will be checked in this procedure. If the limitation is arrived, the tabu part will be ended.

Neighbor Search: The concept of neighbor is defined as follows: For any pair of nodes x and y, we subtract the corresponding attributes values from x to y. If the results are all 0s but only one nonzero integer, x and y are neighbors.

For example, x = [1 2 3], y = [1 3 3], x-y= [0 -1 0], then, x and y are neighbors. In this step, we will produce all the possible neighbors for the nodes we are checking and calculate their fitness values at the same time. The union of neighboring sets from different nodes being checked will be sent to the next step to process it.

Choose Candidates: This procedure will choose the candidates for the circle of the tabu search. There are two kinds of nodes: k-anonymity node and non-k-anonymity node. All thek-anonymity will share a possibility 0.7 and all non-k-anonymity nodes will share the left 0.3. However, among each subclass of nodes, the candidates are equally chosen. In this step, we always remember the best node solution for k-anonymity. Candidate in the tabu list will not be considered.

Update Tabu list: After the candidate chose, the tabu list will be updated. The new candidates will enter into tabu list. Some nodes in the tabu list will be removed if their living time arrives.

Output: The best solution for the tabu search will be produced.

Ⅳ. Experiment result and analysis

This section evaluates the performance of our new search method for different k values (1 to 15). At the same time we also compute the performance of traditional genetic algorithmand tabu search separa-tely. The test database is the Pima Indians Diabetes Database, which is consisted of 768 records and 9 attributes.

In genetic part of our method, we set the population size as 20 and the circle time as 20.In tabu part of our algorithm, we set the number of candidates is 20; size of tabu list 6; living time7;circle time 20.In traditional tabu search and genetic algorithm, we adopt the same setting except that the circle time is 300.

We repeat the algorithm for 50 times and get the average results as shown in Fig. 4and Fig. 5 for different databases.

From the analysis of experiment results, we can see that in most the cases, our new approach is better than single tabu search or genetic algorithm. It has a stable performance than the two methods. For small values of k, the differences among the three methods are small, but as the increase of k the differences increase at the beginning and decrease later.

Ⅴ. Conclusion

This paper presents a hybrid search method composed of tabu search and genetic algorithm. The experimental results show that the proposed heuristic approach is a good method to search for the solutions in lattice space. Even though the method is simple, it achieves better perform than the other two traditional heuristic method.

There are still many things to do in the future work. We can try to find some efficient aspiration criterion in the tabu part. Also in the crossover part, we can replace it with many other kinds of method such as two point crossover. Since the DM value

is not the only way to evaluate the effort of the algorithm, we can design other kinds of special evaluating matrix to satisfy specific command.

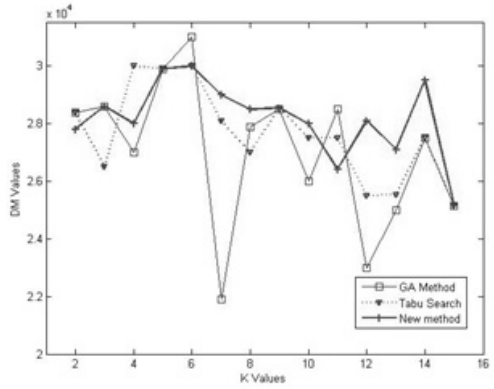

Fig. 4. Average performance results for Pima Indians Diabetes Database

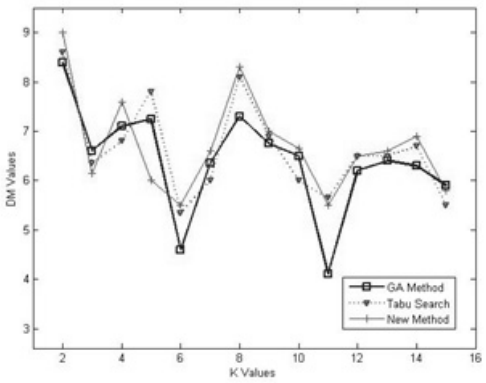

Fig. 5. Average performance results for Poker Database

Acknowledgment

This work was supported by the CTRC project under the auspice of Ministry of Culture, Sport, and Tourism.

References

[1] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proceedings of the 22nd IEEE International Conference on Data Engineering*, pp. 25-25, 2006.

[2] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp. 49-60, 2005.

[3] H. Zhu and X. Ye, "Achieving k-anonymity via a density-based clustering method," in *Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-Age Information Management Conference on Advances in Data and Web Management*, vol. 4505, pp. 745-752, 2007.

[4] Z. H. Wang, J. Xu, W. Wang, and B. Shi, "Clustering-based approach for data anonymization," *Chinese Journal of Software*, vol. 21, no. 4, pp. 680-693, 2010.

[5] H. Park and K. Shim. "Approximation algorithms for k-anonymity," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 67-68, 2005.

[6] V. Iyengar, "Transforming data to satisfy privacy constriants," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery andData Mining*, pp. 279-288, 2002.

[7] M. Lunacek. D. Whitley, and I. Ray, "A crossover operator for the k-anonymity problem," in *Proceedings of the Geneticand Evolutionary Computation Conference*, pp. 1713-1720, 2006.

[8] R. Chaytor, "A better problem representation for k-anonymity," in *Proceedings of the 1st ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD*, pp. 52-61, 2007.

[9] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proceedings of the 21st International Conference on Data Engineering*, pp. 217-228, 2005.

[10] C. C. Charu, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 901-909, 2005.

[11] K. E. Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J. P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, "A globally optimal k-anonymity method for the de-identification of health data," *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 670-682, 2009.

[12] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010-1027, 2001.

[13] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557 - 570, 2002.

[14] M. E. Nergiz, C. Clifton, and A. E. Nergiz, "Multirelational k-anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 8, pp. 1104-1117, 2009.

[15] A. Friedman. R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The International Journal on Very Large Data Bases*, vol. 17, no. 4, pp. 789-804, 2008.

Biography

Cui Run

He is a Ph.D. student in the Graduate School of Information Management and Security, Korea University, Seoul, Korea. He received his bachelor's degree in computer science from Harbin Institute of Technology, Harbin, China in 2008. His research interests include database security, parallel and distributed computing, and data mining.

Hyoung Joong Kim

He received his B.S., M.S., and Ph.D. degrees from Seoul National University, Korea, in 1978, 1986, and 1989, respectively. He joined the faculty member of Kangwon National University, Korea, 1989. He is currently a Professor of Korea University, Korea. He published numerous technical papers including more than 40 peer-reviewed journal papers covering distributed computing and multimedia computing. He served Guest Editor of several journals including IEEE Transactions on Circuits and Systems for Video Technology. He is a Vice Editor-in-Chief of the LNCS Transactions on Data Hiding and Multimedia Security. His main research interests include security engineering.

Dal Ho Lee

He received his B.S., M.S., and Ph.D. degrees from Seoul National University, Korea, in 1982, 1985, and 1992, respectively. He is currently a Professor of the department of electronic engineering at Kyungwon University. In 1997, he was a visiting scholar at University of Southern California, Los Angeles, CA, USA. His research interests include filtering techniques, inertial navigation systems, and security engineering.