International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Privacy-Preserving in Data Mining using Anonymity Algorithm for Relational Data

Karan Dave[1], **Chetna Chand**[2]

[1]Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

[2]Assistant Professor, Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

Abstract: ***Data mining is the process of analyzing data from different perspectives. To summarize it into useful information, we can consider several algorithms. To protect data from unauthorized user in this case is a problem to solve. Access control mechanisms protect sensitive information from unauthorized users. But if the privacy protected information is not in proper format, again the user will compromise the privacy and quality of data. A privacy protection mechanism can use suppression and generalization of relational data to anonymize and satisfy privacy requirements, e.g., k-anonymity and l-diversity, against identity and attribute disclosure. However, privacy is achieved at the cost of precision of authorized information. In this paper, we propose an accuracy-constrained privacy-preserving access control framework. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k-anonymity or l-diversity. An additional constraint that needs to be satisfied by the PPM is the imprecision bound for each selection predicate. The techniques for workload-aware anonymization for selection predicates have been discussed in the literature. However, to the best of our knowledge, the problem of satisfying the accuracy constraints for multiple roles has not been studied before. In our formulation of the aforementioned problem, we propose heuristics for anonymization algorithms and show empirically that the proposed approach satisfies imprecision bounds for more permissions and has lower total imprecision than the current state of the art.***

Keywords: Data mining, Data Integrity, Data privacy, Anonymization, K anonymity, L diversity.

1. Introduction

The problem of data privacy is getting increasingly crucial for our society. This can be proved by the very fact that the accountable management of sensitive knowledge is expressly being mandated through laws. The challenges of privacy-aware access control are similar to the problem of workload-aware anonymization. In our analysis of the related work, we focus on query-aware anonymization. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [9]. Databases within the globe area unit are typically massive and sophisticated. The challenge of querying such infuse in a very timely fashion has been studied by the database, data processing and knowledge retrieval communities, however seldom studied within the security and privacy domain.

The concept of privacy-preservation for sensitive data can require the enforcement of privacy policies or the protection against identity disclosure by satisfying some privacy requirements. We investigate privacy-preservation from the anonymity aspect. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements with minimal distortion of micro data. The anonymity techniques can be used with an access control mechanism to ensure both security and privacy of the sensitive information. The privacy is achieved at the cost of accuracy and imprecision is introduced in the authorized information under an access control policy [1].

2. Overview

Data Mining

Data mining is a recently emerging field, connecting the three worlds of Databases, Artificial Intelligence and Statistics. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it. Data mining, otherwise known as knowledge discovery, attempts to answer this need. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses. As a field, it has introduced new concepts and algorithms such as association rule learning.

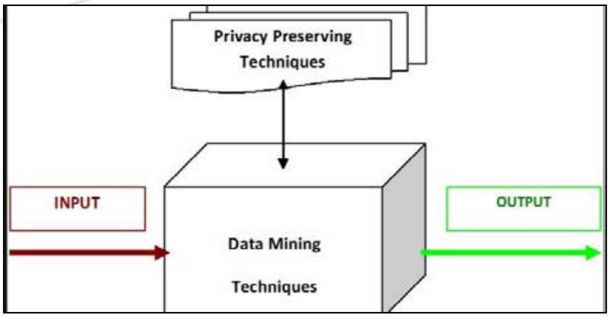
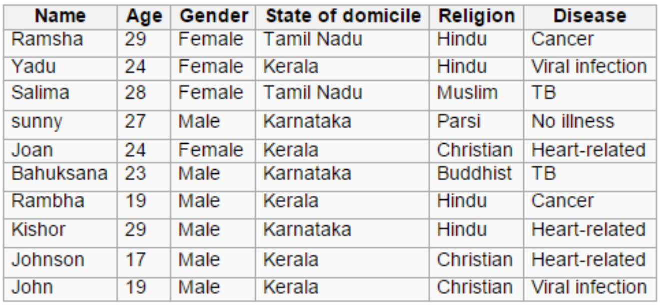


Figure 1: anonymization with data mining

It has also applied known machine-learning algorithms such as inductive-rule learning (e.g., by decision trees) to the setting where very large databases are involved. Data mining techniques are used in business and research and are becoming more and more popular with time. Confidentiality issues in data mining. A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it be scientific, or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research; as can even competing businesses with mutual interests. Despite the potential gain, this is often not possible due to the confidentiality issues which arise

Large Datasets with efficient anonymity

Datasets containing micro-data, that is, information about specific individuals, are increasingly becoming public in response to “open government” laws and to support data mining research. Some datasets include legally protected information such as health histories; others contain individual preferences and transactions, which many people may view as private or sensitive. Privacy risks of publishing micro-data are wellknown. Even if identifiers such as names and Social Security numbers have been removed, the adversary can use background knowledge and cross-correlation with other databases to re-identify individual data records. Famous attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with a public voter database [25] and privacy breaches caused by (ostensibly anonymized) AOL search data [16]. Micro-data are characterized by high dimensionality and sparsity. Each record contains many attributes (i.e., columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no “similar” records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the “fat tail” phenomenon: individual transaction and preference records tend to include statistically rare attributes. Our contributions. Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the probability of successful de-anonymization, the other on the amount of information recovered about the target. Unlike previous work [25], we do not assume a priori that the adversary’s knowledge is limited to a fixed set of “quasi-identifier” attributes. Our model thus encompasses a much broader class of de-anonymization attacks than simple cross-database correlation.

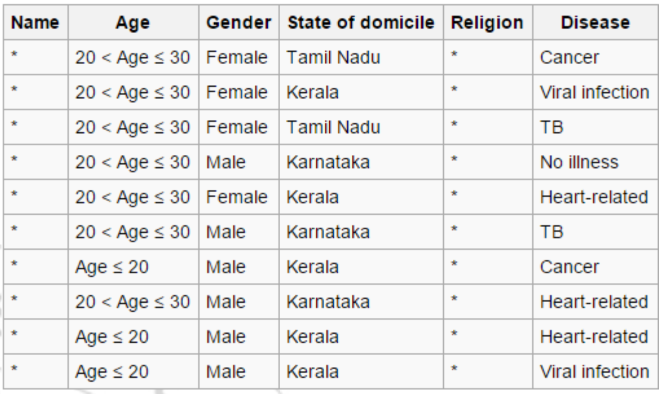


Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Figure 2: data set of patients in hospital

here are 6 attributes and 10 records in this data. There are two common methods for achieving k-anonymity for some value of k.

- Suppression: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. In the anonymized table below, we have replaced all the values in the 'Name' attribute and all the values in the 'Religion' attribute have been replaced by a '*'.
- Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' , etc.



Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

Figure 3: Anonymized data set of patients in hospital

This data has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'State of domicile' since for any combination of these attributes found in any row of the table there are always at least 2 rows with those exact attributes. The attributes available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple occurs in at least k records for a dataset with k-anonymity

3. Background Theory

3.1 Document Summarization

Eigenword Vector Summarization

An *Eigenword* is a real-valued vector "embedding" associated with a word that captures its meaning in the sense that distributional similar words have similar eigenword. This page contains links to several sets of eigenword They are computed as the singular vectors of the matrix of co-occurrence of words and their contexts, and used in a variety of spectral NLP methods and applications

Each sentence feature has its unique Contribution and combing them would be advantageous. Therefore we investigate combined sentence features for extractive summarization. [2] Currently, most successful multi-document summarization systems [5] follow the extractive summarization framework. These systems first rank all the sentences in the original document set and then select the most salient sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries, some intensive post-processing approaches are also appended on the extracted sentences. Two Summary Construction Methods are applied first one is Abstractive method where summaries produce generated text from the important parts of the documents and second is Extractive Method where summaries identify important sections of the text and use them in the summary as they are.

The sentence similarity calculation remains central to the existing approaches. The indexing weights of the document terms are utilized to compute the sentence similarity values. Elementary document features are used to allocate an indexing weight to the document terms, which include the document length, term frequency, occurrence of a term in a background corpus. Therefore, the indexing weight of the other terms appearing in the document remains independent and the context in which the term occurs is overlooked in assigning its indexing weight for the documents. This results in “context independent document indexing.” To the authors’ knowledge, no other work in the existing literature addresses the problem of “context independent document indexing” for the document summarization task.

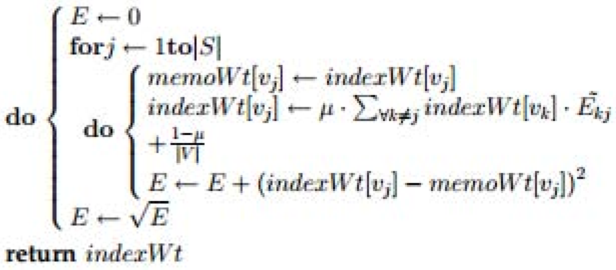
A document contains both the background terms as well as the content-carrying terms. In the sentence similarity analysis the traditional indexing schemes cannot distinguish between these terms. The higher weight is given by the context sensitive document indexing model to the topical terms where it is compared with the non topical terms and thus influences the sentence similarity values in a positive manner. Using the lexical association between document terms the system considers the problem of “context independent document indexing. The content carrying words will be highly associated with each other in a document, while the background terms will have very low in association with the other terms in the document. The association between terms is stated in this paper by the lexical association and is computed through the corpus analysis.

3.2 Word Indexing

- **Sentence Similarity based**

Sentence similarity assessment is key to most NLP applications. This paper presents a means of calculating the similarity between very short texts and sentences without using an external corpus of literature. This method uses WordNet, common-sense knowledge base and human intuition. Results were verified through experiments. These experiments were performed on two sets of selected sentence pairs. We show that this technique compares favorably to other word-based similarity measures and is flexible enough to allow the user to make comparisons without any additional dictionary or corpus information. We believe that this method can be applied in a variety of text knowledge representation and discovery applications.

- **Context based**



```
do {
    E ← 0
    for j ← 1 to |S|
    do {
        memoWt[v_j] ← indexWt[v_j]
        indexWt[v_j] ← μ · Σ_{∀k≠j} indexWt[v_k] · Ê_kj + (1−μ)/|V|
        E ← E + (indexWt[v_j] − memoWt[v_j])^2
    }
    E ← √E
}
return indexWt
```

Figure 2: Context Base word Indexing equation

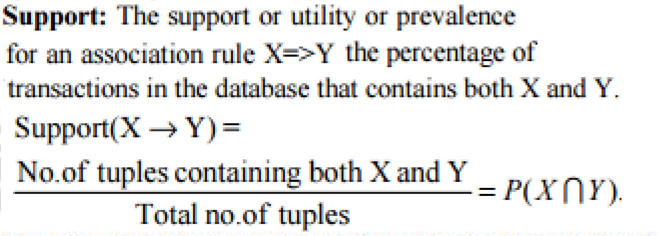
Given the lexical association measure between two terms in a document from hypothesis H2, the next task is to calculate the context sensitive indexing weight of each term in a document using hypothesis H3. A graph -based iterative algorithm is used to find the context sensitive indexing weight of each term. Given a document Di, a document graph G is built. Let G = (V,E) be an undirected graph to reflect the relationships between the terms in the document Di. V = {Vj|1 ≤ j ≤_ |V|} denotes the set of vertices, where each vertex is a term appearing in the document. E is a matrix of dimensions |V| × |V|. Each edge ejk ε E corresponds to the lexical association value between the terms corresponding to the vertices vj and vk. The lexical association between the same terms is set to 0.

4. Data Anonymization Algorithm

4.1 K-Anonymity

To count the support of all these combinations and to store them the count-tree is used, based on the count tree algorithm. The tree assumes an order of items and their generalizations, based on their frequencies (supports)in D.

Definition 3.2.1



Support: The support or utility or prevalence for an association rule X=>Y the percentage of transactions in the database that contains both X and Y.

$$\text{Support}(X \to Y) = \frac{\text{No.of tuples containing both X and Y}}{\text{Total no.of tuples}} = P(X \cap Y).$$

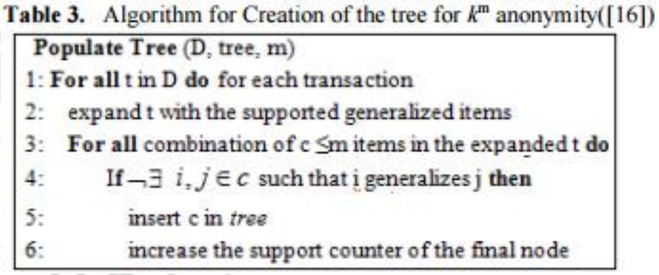
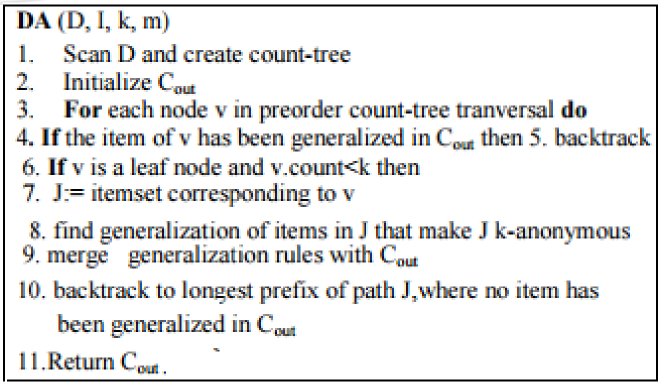


Table 3. Algorithm for Creation of the tree for k^m anonymity([16])

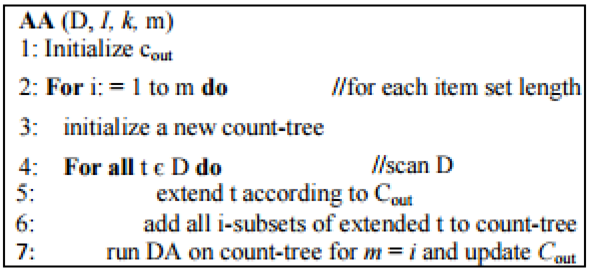
```
Populate Tree (D, tree, m)
For all t in D do for each transaction
    expand t with the supported generalized items
    For all combination of c ≤m items in the expanded t do
        If ¬∃ i, j ∈ c such that i generalizes j then
            insert c in tree
            increase the support counter of the final node
```

4.2 Direct Anonymization Algorithm



```
DA (D, I, k, m)
1.  Scan D and create count-tree
2.  Initialize C_out
3.  For each node v in preorder count-tree tranversal do
4. If the item of v has been generalized in C_out then 5. backtrack
 6. If v is a leaf node and v.count<k then
 7. J:= itemset corresponding to v
  8. find generalization of items in J that make J k-anonymous
 9. merge  generalization rules with C_out
10. backtrack to longest prefix of path J,where no item has
    been generalized in C_out
11.Return C_out .
```

4.3 Apriori based Anonymization Algorithm



```
AA (D, I, k, m)
Initialize c_out
For i: = 1 to m do              //for each item set length
   initialize a new count-tree
   For all t ϵ D do                  //scan D
         extend t according to C_out
         add all i-subsets of extended t to count-tree
      run DA on count-tree for m = i and update C_out
```

Figure 3: model for concept based analysis of data

The process has involved the above stated steps. Basically they all have one ir the other conceptual technique based on text mining and data mining. We are proposing to use Bernoulli morel and context based similarity indexing for words because the process does not take much time and become efficient than the earlier one.

5. Proposed Algorithm

Input: A set T of n records; the value k for k-anonymity and the value l for l-diversity

Output: A Partition P = {P1, P2...Pk}

1. Sort all records in T by their quasi-identifiers;
2. Let K := [n/k];
3. Select K distinct records based on their frequency in sensitive attribute values;
4. Let Pi := {r_i} for i = 1 to K;
5. Let T := T / {r1, r2...rk}; 6. While (T ≠ φ) do
6. Let r be the first record in T ;
7. Order {Pi} according to their distances from r;
8. Let i = 1;
9. Flag = 0;
10. While ((i< K) or ((s(r) ∈ s(Pi)) and (|s(Pi| < l))
11. Let s(Pi) be the set of distinct sensitive attribute values of Pi;
12. Let s(r) be the sensitive attribute value of r;
13. if((|Pi| < K) or ((s(r) ∈ s(Pi)) and (|s(Pi| < l))
14. then add r to Pi;
15. Update centroid of Pi;
16. Flag = 1;
17. Else i := i + 1;
18. If (Flag = 0) add r to the nearest cluster;
19. Let T := T /{r};
20. End of while

6. Conclusion

I have followed the strategies and methods available and written in the base and research papers. After studying it in detail and searching and learning the idea behind privacy preservation by maintaining accuracy constraints, I have followed the l-diversity method. It has many advantages over the previous k-anonymous algorithm. It does not impede the flow of information. I have followed the approach of randomization. But other approaches can also be studied and taken further as cryptographic and statistical disclosure control. The group based anonymization process to preserve privacy in data sets by reducing granularity of a data representation is displayed

6.1 Expected outcome

- Accuracy constrained dataset with higher predictive privacy and limit the sensitive attributes.
- To limit the gain of some prior belief B_0 to the chosen limit B_1
- To add new data table contents with ease to the existing one so that the data privacy cannot be refrained
- To measure the distance between two probabilistic distributions and thus maintaining accuracy with privacy.
- To modify l-diversity for the above stated gains.

6.2 Performance evaluation

- Comparison of levels of anonymization among various datasets.
- Demonstrating it in the form of graph
- Preparing a table of analyzed data to show various results
- Preparing a table of analyzed data to show various results

References

[1] Aggarwal, G., Feder, G., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D. and Zhu, A.: Achieving Anonymity via Clustering, In Proc. of ACM PODS, (2006), pp.153-162.

[2] Aggarwal, G., Feder, G., Kenthapadi,R., Motwani, R., Panigrahy, D., Thomas, and Zhu, A.: Approximation Algorithms for k-Anonymity, .Journal of Privacy Technology, (2005). International Journal of Advanced Information Technology (IJAIT) Vol. 2, No.5, October 2012 13

[3] Atzori, M., Bonchi, F., Giannotti, F., and Pedreschi, D.: Anonymity Preserving Pattern Discovery, VLDB Journal, accepted for publication, (2008).

[4] Bayardo, R. J. and Agrawal, R.: Data Privacy through Optimal k-Anonymization, In Proc. of ICDE, (2005), pp.217-228.

[5] Ghinita, G., Karras, F P., Kalnis, P., and Mamoulis, N.: Fast Data Anonymization with Low Information Loss, In VLDB, (2007), pp.758-769.

[6] Ghinita, G., Tao, Y., and Kalnis, P.: On the Anonymization of Sparse High-Dimensional Data, In Proceedings of ICDE, (2008).

[7] Han, J., Pei, J., and Yin, Y.: Mining frequent patterns without candidate generation, In Proc. of ACM SIGMOD, (2000), pp.1-12.

[8] Iyengar, V.S.: Transforming Data to Satisfy Privacy Constraints, In Proceedings of SIGKDD, (2002), pp.279-288.

[9] Privacy Preserving Data Mining∗ Yehuda Lindell Department of Computer Science Weizmann Institute of Science Rehovot, Israel. lindell@wisdom.weizmann.ac.il

[10] Privacy Preserving Data Mining Cynthia Dwork and Frank McSherry. 2012

[11] k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1 LATANYA SWEENEY School of

Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA E-mail: latanya@cs.cmu.edu-Received May 2002

[12] ℓ-Diversity: Privacy Beyond k-Anonymity Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer Muthuramakrishnan Venkitasubramaniam, Department of Computer Science, Cornell University {mvnak, johannes, dkifer, vmuthu} @cs.cornell.edu – release - 2012

[13] Privacy Preserving Suppression Algorithm for Anonymous Databases Ebin P.M 1, Brilley Batley. C 2 1,2 AMIE, Assistant Professor Department of Computer Science & Engineering, Hindustan University, Chennai, India pmebin74@gmail.com .(IJSR), India Online ISSN: 2319- 7064. Volume 2 Issue 1, January 2013

[14] A Survey on Security and Accuracy Constrained Privacy Preserving Task Based Access Control Mechanism for Relational Data Pratik Bhingardeve 1, D. H. Kulkarni21, 2 Pune University, Smt. Kashibai Navale College of Engineering, Vadgaon (BK), Pune-411041, India – IJSR-Feb-2013

[15] https://en.wikipedia.org/wiki/K-anonymity

[16] IJRITCC ISSN: 2321-8169 Volume: 3 Issue: 4 Security Management Methods in Relational Data Suhasini Gurappa .Metri PG Student, CSE Dept Cambridge institute of technology ,Bangalore ,India.

[17] Zahid Pervaiz, Walid G.Aref, Arif Gafoor, "Accuracy constrained privacy preserving access control mechanism for relational databases" IEEE Transaction on Knowledge Engineering, vol.26, No.4, April 2014, pp.795-807 .

[18] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload Anonymization Techniques for Large-Scale Datasets," ACMTrans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.

[19] A. Machanavajjhala, D. kifer, j. Gehrke, and M. Venkitasubramaniam,"L-Diversity: Privacy Beyond k-anonymity," ACM Trans.Knowledge Discovery from Data,vol. 1, no. 1, article 3, 2007.

[20] S. Rizvi, A. Mendelzon, S. Sudarshan, and P. Roy, "ExtendingQuery Rewriting Techniques for Fine-Grained Access Control,"Proc. ACM SIGMOD Int'l Conf. Management of Data, pp.1-562,2004.

[21] E. Bertino and R. Sandhu, "Database Security-Concepts, Approaches, and Challenges," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-Mar. 2005.

[22] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov. 2001.