Yanbu Journal of Engineering and Science

Volume 5 | Issue 1 Article 2

2012

# A SURVEY OF PRIVACY PRESERVING DATA MINING ALGORITHMS

Abou el ela Abdou Hussein
*Shagra University, Saudi Arabia*

Nermin Hamza
*Institute of Statistical Studies and Research, Cairo University, Egypt*

Ashraf A. Shahen
*Institute of Statistical Studies and Research, Cairo University, Egypt*

Hesham A. Hefny
*Institute of Statistical Studies and Research, Cairo University, Egypt*

Follow this and additional works at: https://yjes.researchcommons.org/yjes

The journal in which this article appears is hosted on Digital Commons, an Elsevier platform.

## Recommended Citation

Hussein, Abou el ela Abdou; Hamza, Nermin; Shahen, Ashraf A.; and Hefny, Hesham A. (2012), A SURVEY OF PRIVACY PRESERVING DATA MINING ALGORITHMS, *Yanbu Journal of Engineering and Science*: Vol. 5: Iss. 1, 10-18.
DOI: https://doi.org/10.53370/001c.24055
Available at: https://yjes.researchcommons.org/yjes/vol5/iss1/2

This Survey is brought to you for free and open access by Yanbu Journal of Engineering and Science. It has been accepted for inclusion in Yanbu Journal of Engineering and Science by an authorized editor of Yanbu Journal of Engineering and Science.

Yanbu Journal
of
Engineering
and Science
ISSN: 1658-5321

Vol. 5, October 2012 (1433H)
www.yjes.org.sa

# A SURVEY OF PRIVACY PRESERVING DATA MINING ALGORITHMS

Abou el ela Abdou Hussien[1], Nermin Hamza[2], Ashraf A. Shahen[2] and Hesham A. Hefny[2]

*1 Department of Computer Science, Faculty of Science and Arts, Shaqra University, Saudi Arabia*
*2 Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University*

**ABSTRACT**

Data mining is the extraction of vast interesting patterns or knowledge from huge amount of data. In recent years, with the explosive developments in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. Privacy preserving data mining (PPDM) has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes. Privacy-preserving data mining considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm. There are two classic settings for privacy-preserving data mining (although these are by no means the only ones). In the First setting, the data is divided among two or more different parties; the aim is to run a data mining algorithm on the union of parties' databases without allowing any party to view another individual's private data. In the second setting, some statistical data that is to be released (so that it can be used for research using statistics and/or data mining) may contain confidential data; hence, it is first modified so that (a) the data does not compromise anyone's privacy, and (b) it is still possible to obtain meaningful results by running data mining algorithms on the modified data set. In recent years, PPDM has been studied extensively, because of the wide proliferation of sensitive information on the internet. This paper provides a wide survey of different PPDM algorithms and analyses of the representative techniques for PPDM, and points out their merits and demerits. Finally the present problems and directions for future research are discussed.

**Keywords**: *Privacy Preserving Data Mining, Privacy, Randomization, K-Anonymity*

## 1. INTRODUCTION

Data mining is a technique for automatically and intelligently extracting information or knowledge from a large amount of data. However, it can also disclose sensitive information about individuals compromising the individual's right to privacy. Therefore, privacy preserving data mining (PPDM) has becoming an increasingly important field of research.

PPDM is a novel research direction. In the recent years, with the rapid development of the internet, data storage and data processing technologies, PPDM has been drawn increasing attention. In order to make a publicly available system secure, one must ensure not only that private sensitive data have been trimmed out, but also to make sure that certain inference channels have been blocked as well.

Under privacy constraints, the privacy preserving data mining problem was extensively researched. In one hand, The main goals a PPDM algorithm should enforce are:



Received: Jan. 15, 2012 Accepted: April 26, 2012

1. A PPDM algorithm should have to prevent the discovery of sensible information.
2. It should be resistant to the various data mining techniques.
3. It should not compromise the access and the use of non sensitive data.
4. It should not have an exponential computational complexity.

On the other hand using PPDM has a big trade off between security, efficiency, and accuracy Improving any one aspect typically degrades the other two, see Figure 1.

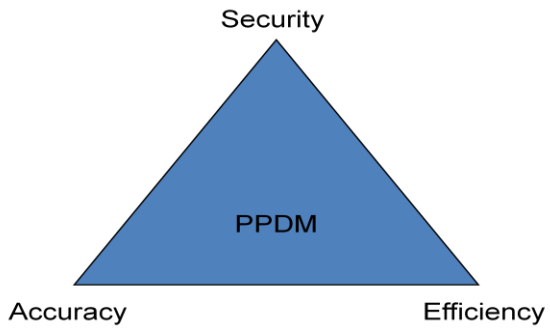


Fig. 1. Tradeoffs Security Accuracy & Efficiency

This paper provides a wide survey of different PPDM techniques and analyses the representative methods for PPDM. It points out their merits and demerits. The rest of this paper is organized as follows. In section 2, we introduce the classification of privacy preserving methods. In section 3, we discuss the anonymization method. In section 4, we analyze the method of randomization for privacy preserving on the original data. The encryption method is discussed in section 5. Finally, section 6 contains the conclusions and future work.

## 2. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

The goal of PPDM is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. The topic of privacy preserving data mining has been extensively studied by the data mining community in recent years. A number of effective methods for PPDM have been proposed. Most methods use some form of transformation on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. We classify them into the following three categories:

**The anonymization method:**

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The representative anonymization method is k-anonymity. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records in order to uniquely identify the records. Many advanced methods have been proposed, such as, p-sensitive k-anonymity, (a, k)-anonymity, l-diversity, t-closeness, M-invariance, Personalized anonymity, and so on. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent [1, 2].

**The randomization method:** Randomization method is a popular method in current privacy preserving data mining studies. It masks the values of the records by adding noise to the original data.

The noise added is sufficiently large so that the individual values of the records can no longer be recovered. However, the probability distribution of the aggregate data can be recovered and subsequently used for privacy-preservation purposes. In general, randomization method aims at finding an appropriate balance between privacy preservation and knowledge discovery.

Representative randomization methods include random-noise-based perturbation and Randomized Response scheme. The randomization method is more efficient. However, it results in high information loss [3-5].

**The encryption method:** Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. There are two different distributed privacy preserving data mining approaches such as the method on horizontally partitioned data and that on vertically partitioned data. The encryption method can ensure that the transformed data is exact and secure, but it is much low efficient [6-7].

## 3. THE ANONYMIZATION APPROACH

There is a table of the form: D (Explicit_ Identifier, Quasi_Identifier, Sensitive_ Attributes, Non_Sensitive_ Attributes), where Explicit_ Identifier is a set of attributes ,such as name and social security number( SSN), containing information that explicitly identifies record owners; Quasi_Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive_Attributes consists of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive_Attributes contains all attributes that do not fall into the previous three categories [1]. The four sets of attributes are disjoint.

Most works assume that each record in the table represents a distinct record owner. Anonymization refers to the PPDM approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed .Even with all explicit identifiers being removed, Sweeney [2] showed a real-life privacy threat to William Weld, former governor of the state of Massachusetts. For example, Table 2 is an anonymous table of Table1. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. For example , the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. Suppression involves not releasing a value at all. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data. In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression.

TABLE 1: ORIGINAL TABLE

| Name | Race | Birth | Sex | Zip | Disease |
|---|---|---|---|---|---|
| Alice | Blank | 1965 | M | 02141 | Flu |
| Bob | Blank | 1965 | M | 02142 | Cancer |
| David | Blank | 1966 | M | 02135 | Obesity |
| Helen | Blank | 1966 | M | 02137 | Gastritis |
| Jane | White | 1968 | F | 02139 | HIV |
| Paul | White | 1968 | F | 02138 | Cancer |

TABLE 2. ANONYMIZATION OF TABLE 1

| Race | Birth | Sex | Zip | Disease |
|---|---|---|---|---|
| Blank | 1965 | M | 0214* | Flu |
| Blank | 1965 | M | 0213* | Cancer |
| Blank | 1966 | M | 0213* | Obesity |
| Blank | 1966 | M | 0213* | Gastritis |
| White | 1968 | F | 0213* | HIV |
| White | 1968 | F | 0214* | Cancer |

Bayardo and Agrawal [22] presented an optimal algorithm that starts from a fully

generalized table and specializes the dataset in a minimal k-anonymous table. LeFevre et al. [29] described an algorithm that uses a bottom-up technique and a priori computation. Fung et al. [14] presented a top-down heuristic to make a table to be released k-anonymous. As to the theoretical results, Sweeney [10] proved the optimal k-anonymity is NP-hard and provided approximation algorithms for optimal k-anonymity. However, Machanavajjhala et al. [11] pointed out that the user may guess the sensitive values with high confidence when the sensitive data is lack of diversity, and introduced the l-diversity method. Xiao and Tao [12] presented a new generalization framework based on the concept of personalized anonymity. Their technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the original data. Subsequently, several models such as p-sensitive k-anonymity [13], (a, k)-anonymity [15], t-closeness [16], and M-invariance [17], etc. were proposed in the literature in order to deal with the problem of k-anonymity.

The existing k-anonymity solutions based on generalization and suppression techniques suffer from high information loss and low usability mainly due to reliance on pre-defined generalization hierarchies or full order imposed on each attribute domain. References [18-21] provided a kind of new algorithm based on clustering technique, which reduced greatly the amount of information loss resulting from data generalization for implementing data anonymization.

## 4. RANDOMIZED RESPONSE TECHNIQUES

The method of randomization can be described as follows. Consider a set of data records denoted by $X = \{x_1 \ldots x_N\}$.For $x_i \varepsilon X$, we add a noise component which is drawn from the probability distribution fy(X).These noise components are drawn independently, and are denoted $y_1 \ldots y_N$ Thus, the new set of distorted records are denoted by $x_1 + y_1 \ldots x_N + y_N$. We denote this new set of records by $z_1 \ldots z_N$. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. Thus, if X is the random variable denoting the data distribution for the original record, Y is the random variable describing the noise distribution, and Z is the random variable denoting the final record, we have:

$$Z = X + Y, \quad X = Z - Y$$

Now, we note that N instantiations of the probability distribution Z are known, whereas the distribution Y is known publicly. For a large enough number of values of N, the distribution Z can be approximated closely by using a variety of methods such as kernel density estimation. By subtracting Y from the approximated distribution of Z, it is possible to approximate the original probability distribution X. In practice, one can combine the process of approximation of Z with subtraction of the distribution Y from Z by using a variety of iterative methods Such iterative methods typically have a higher accuracy than the sequential solution of first approximating Z and then subtracting Y from it.

The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. Although information from each individual user is scrambled, if the number of users is

significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items. Randomized Response technique was first introduced by Warner as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A, queries are sent to a group of people Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models:
The first step is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm. The model of randomization is shown in Figure 2.

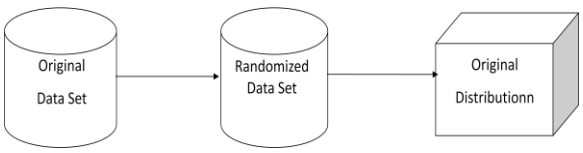


Fig. 2. The Model of Randomization

It is shown that given the distribution of random noises, reconstructing the distribution of the original data is possible. Subsequently, Evmievski et al. proposed an approach to conduct privacy preserving association rule mining [23]. Kargupta et al. [3] proposed a random matrix-based spectral filtering technique to recover the original data from the perturbed data. Huang et al. further proposed two other data reconstruction methods: PCA-DR and MLE-DR in [4]. In addition, several distribution reconstruction algorithms have been proposed in correspondence to different randomization operators [24-26]. The basic idea of most algorithms is to use Bayesian analysis to estimate the original data distribution based on the randomization operator and the randomized data. For example, the expectation maximization (EM) algorithm [25] generates reconstructed distribution that converges to the maximum likelihood estimate of the original distribution.
In data mining community, Rizvi and Haritsa presented a MASK scheme to mine association rules with secrecy constraints [24]. Du and Zhan proposed an approach to conduct privacy preserving decision tree building [5]. Guo et al. addressed the issue of providing accuracy in terms of various reconstructed measures in privacy preserving market basket data analysis [27]. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Therefore, the randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process. However, it results in high information loss.

## 5. ENCRYPTION TECHNIQUE

Another branch of PPDM which using cryptographic techniques was developed. This branch became hugely popular [6] for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [7] has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of PPDM.

TABLE 3: MERITS AND DEMERITS OF PPDM

| Techniques of PPDM | Merits | Demerits |
|---|---|---|
| ANONYMIZATION | This method is used to protect respondents' identities while releasing truthful information. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. | There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables.<br>The second limitation is that the k-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. |
| RANDOMIZED RESPONSE | The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in PPDM. The randomization method is more efficient. However, it results in high information loss. | Randomized Response technique is not for multiple attribute databases. |
| CRYPTOGRAPHIC | Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. | This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records. |

In general, distributed data mining involves two forms: horizontally partitioned data and vertically partitioned data. Horizontally partitioned data means that each site has complete information on a distinct set of entities, and an integrated dataset consists of the union of these datasets. In contrast, vertically partitioned data has different types of information at each site; each has partial information on the same set of entities.

In horizontally partitioned data sets, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. Many of these techniques use specialized versions of the general methods discussed in for various problems. The work in discusses the construction of a popular decision tree induction method called ID3 with the use of approximations of the best splitting attributes. Subsequently, a variety of classifiers have been generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier and the SVM Classifier with nonlinear kernels . An extreme solution for the horizontally partitioned case is discussed in [8], in which privacy preserving classification is performed in a fully distributed setting, where each customer has private access to only their own record. A host of other data mining applications have been generalized to the problem of horizontally partitioned data sets [28]. These include the applications of association rule mining, clustering, and collaborative filtering.

For the vertically partitioned case, many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms. For example, the methods in [8] discuss how to use to scalar dot product computation for frequent item set counting. The process of counting can also be achieved by using the secure size of set intersection as described in

[9]. Another method for association rule mining uses the secure scalar product over the vertical bit representation of item set inclusion in transactions, in order to compute the frequency of the corresponding item sets. This key step is applied repeatedly within the framework of a roll up procedure of item set counting. It has been shown that this approach is quite effective in practice. The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees, SVM Classification, Naïve Bayes Classifier, and k-means clustering. Table 3 shows merits and demerits of PPDM techniques.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we carried out a wide survey of the different approaches for privacy preserving data mining (PPDM), and analyses the major algorithms available for each method and points out the existing drawback. While all the proposed methods were only approximate to our goal of privacy preservation, we need to further enhance those approaches or develop some efficient methods. To address this issue, we prefer that the following problems should be widely studied:

(1) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.

(2) Side-effects are unavoidable in data sanitization process. How to reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring the side-effects resulted from data processing.

(3) In distributed PPDM areas, efficiency is an essential issue. We should try to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.

(4) The way to deploy privacy-preserving techniques into practical applications is also required to be further studied.

## REFERENCES


[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, andA. Zhu. Anonymizing tables. In Proc. of the 10th International Conference on Database Theory (ICDT), pages 246{258, Edinburgh, UK, January 2005.

[2] L. Sweeney."k-anonymity: a model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570, 2002.

[3] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, “On the Privacy Preserving Properties of RandomData Perturbation Techniques”, In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.

[4] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland, USA, pp.37-48, 2005.

[5] W. Du, Z. Zhan, “Using Randomized Response Techniques for Privacy Preserving Data Mining”, In Proceedings 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.505-510, 2003.

[6] Laur, H. Lipmaa, and T. Mieli' ainen."Cryptographically private support vector machines". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624, 2006.

[7] PINKAS, B. 2002. Cryptographic techniques for privacy-preserving data mining. ACM SIGKDD Explor. Newsl. 4, 2, 12–19.

[8] Yang Z., Zhong S.Wright R. Privacy-Preserving Classification of Customer Data without Loss of Accuracy. SDM Conference, 603-610, 2006.

[9] Clifton C. Kantarcioglou M., Lin X., Zhu M.Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations, 4(2), 2002.

[10] L. Sweeney, “Achieving k-Anonymity Privacy Protection Using Generalization and suppression”,

International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.571-588, 2002.

[11] A. Machanavajjhala, J. Gehrke, D. Kifer, "l-Diversity: Privacy Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data, pp.24-35, 2007.

[12] X.K. Xiao, Y.F. Tao, "Personalized Privacy Preservation", In Proceedings of the ACM Conference on Management of Data (SIGMOD), pp.229-240, 2006.

[13] T. Truta, B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property", In roceedings of the 22nd International Conference on Data Engineering Workshops, pp. 94-103, 2006.

[14] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.

[15] R.C.Wong, J.Y.Li, A.W. Fu, "(α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing", In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.754-759, 2006.

[16] N.H. Li, T.C. Li, "t-Closeness: Privacy beyond k-Anonymity and l-Diversity", In roceedings of the 23rd International Conference on Data Engineering, pp.106-115, 2007.

[17] X.K. xiao, Y.F. Tao, "M-Invariance: Towards Privacy Preserving Re-Publication of Dynamic Datasets", In Proceedings of the ACM Conference on Management of Data (SIGMOD), pp.689-700, 2007.

[18] G. Loukides, J.H. Shao, "An Efficient Clustering Algorithm for k-Anonymisation", nternational Journal of Computer Science And Technology,vol.23, no.2, pp.188-202, 2008.

[19] J.L. Lin, M.C. Wei, "Genetic Algorithm-Based Clustering Approach for k-nonymization", International Journal of Expert Systems with Applications,vol.36, no.6, pp.9784-9792, 2009.

[20] L.J. Lu, X.J. Ye, "An Improved Weighted-Feature Clustering Algorithm for k-Anonymity", In Proceedings of the 5th International Conference on Information Assurance and Security, pp.415-419, 2009.

[21] Z.H. Wang, J. Xu, W. Wang, B.L. Shi, "Clustering-Based Approach for Data Anonymization", Journal of Software,vol.21, no.4, pp.680-693, 2010.

[22] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), pages 217{228, Tokyo, Japan, 2005.

[23] Evfimievski, A.Srikant, R.Agrawal, and Gehrke J,"Privacy preserving mining of association rules". In Proc.KDD02, pp. 217228, 2002.

[24] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[25] D. Agrawal, C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", In Proceedings of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp.247-255, 2001.

[26] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discoveryin Databases and Data Mining, pp.217-228, 2002.

[27] S.J. Rizvi, J.R. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th VLDB conference, pp.1-12, 2002.

[28] M. Kantarcioglu and C. Clifton."Privacy-preserving distributed mining of association rules on horizontally partitioned data", In Proc.of DKMD'02, 2002.

# بحث في خوارزميات صيانة خصوصية البيانات المنقبة

أبوالعلا عبده أبوالعلا[1]، نرمين حمزة[2]، أشرف شاهين[2] ، هشام حفني[2]

1 قسم علوم الحاسب ، كلية العلوم والآداب ، جامعة شقراء ، المملكة العربية السعودية
2 قسم علوم الحاسب والمعلومات ، معهد الدراسات الاحصائية ، جامعة القاهرة، جمهورية مصر العربية

**الملخص:**

تنصب عملية تنقيب البيانات على استخراج المعلومات الهامة من كميه هائلة من البيانات . في السنوات الأخيرة ومع التقدم الهائل في شبكة المعلومات العنكبوتية والزيادة العالية في قدرات التخزين وتكنولوجيا تشغيل البيانات أصبحت صيانة خصوصية البيانات تمثل تحديا كبيرا في عملية تنقيب البيانات. ولقد أصبحت عملية صيانة تنقيب البيانات أكثر ثراء لأنها تسمح المشاركة في البيانات الخاصة والحساسة وإجراء عمليات التحليل عليها. يوجد تقسيم تقليدي لصيانة خصوصية تنقيب البيانات. أولهما يتم تقسيم البيانات على أساسه إلى قسمين أو أكثر من قواعد البيانات ويتم تشغيل خوارزم تنقيب البيانات على القسمين معا بدون السماح لاى قسم رؤية البيانات الشخصية والخاصة بالقسم الأخر. ثانيهما يقوم على أن بعض البيانات الإحصائية والمعلنة من الممكن أن تحتوى على بعض البيانات السرية والخاصة لذا يتم تعديلها بحيث لا يسمح بالتساهل في خصوصية أي شخص وكذلك يمكن الحصول على نتائج منطقية عن طريق تشغيل خوارزميات التنقيب على البيانات بعد تعديلها. حاليا يتم دراسة تأمين خصوصية تنقيب البيانات بشكل مطرد للحفاظ على البيانات الحساسة في الشبكة العنكبوتية ونقوم في ورقة العمل هذه باستعراض الطرق المختلفة المستخدمة في صيانة خصوصية البيانات المنقبة وعمل تحليل لهذه الطرق وعرض لمميزاتها وعيوبها وأخيرا استعراض مشاكل واتجاهات البحث في المستقبل.