Sanitization and Anonymization of Document Repositories

Yücel Saygın, Sabanci University, Istanbul, Turkey

ysaygin@sabanciuniv.edu

Dilek Hakkani-Tür, AT&T Labs – Research Florham Park, NJ, USA

dtur@research.att.com

Gökhan Tür, AT&T Labs – Research Florham Park, NJ, USA

gtur@research.att.com

Sanitization and Anonymization of Document Repositories

Abstract

Information security and privacy in the context of the World Wide Web (WWW) is an important issue that is still being investigated. However, most of the present research is dealing with access control, and authentication-based trust. Especially with the popularity of WWW, being one of the largest information sources, privacy of the individuals is now as important as the security of information. In this chapter, our focus is text, which is probably the most frequently seen data type in the WWW. Our aim is to highlight the possible threats to privacy that exist due to the availability of document repositories and sophisticated tools to browse and analyze these documents. We first identify possible threats to privacy in document repositories. We then discuss a measure for privacy in documents with some possible solutions to avoid or at least alleviate these threats.

Keywords : Internet privacy, Security Risk, Data Mining, Text Database, Information Filtering, Data Quality

INTRODUCTION

Information has been published in various forms throughout the history and sharing information has been one of the key aspects of development. Internet revolution and World Wide Web (WWW) made the publishing, and accessing information much easier than it used to be. However, widespread data collection and publishing efforts on the WWW increased the privacy concerns since most of the gathered data contain private information. Privacy of individuals on the WWW may be jeopardized via search engines and browsers, or sophisticated text mining tools that can dig through mountains of web pages. Privacy concerns need to be addressed attention since they may hinder the data collection efforts and reduce the number of publicly available databases that are extremely important for research purposes such as in machine learning, data mining, information extraction/retrieval, and natural language processing.

In this chapter we consider the privacy issues that may originate from publishing data on the WWW. Since text is one of the most frequently and conveniently used medium in the WWW to convey information, our main focus will be text documents. We basically tackle the privacy problem in two phases. The former phase, referred to as *sanitization*, aims to protect the privacy of the contents of the text against possible threats. The latter phase, called *anonymization* Sanitization basically deals with the automatic identification of named entities, such as sensitive terms, phrases, proper names, and numeric values (e.g., credit card numbers) in a given text, and modification of them with the purpose of hiding private information. The

latter phase, called *anonymization*, makes sure that the classification tools cannot predict the owner or author of the text.

In the following sections, we first provide the taxonomy of possible threats. In addition to that, we propose a privacy metric for document databases based on the notion of *k*-anonymity together with a discussion of the methods that can be used for preserving the privacy.

BACKGROUND AND RELATED WORK

Privacy and security issues were investigated in the database community in the context of statistical databases, where the users are limited to statistical queries. In statistical databases, privacy is protected by limiting the queries that can be issued by the user to non-confidential values, or statistical operations (Adam & Brown, 2004). Security leaks that result from the intersection of multiple queries are investigated and the privacy is defined by the concept of *k*-anonymity. A database table provides *k*-anonymity, if it cannot be used to unambiguously identify less than k entities (Samarati & Sweeney, 1998).

Currently data mining community is investigating how data could be mined without actually seeing the confidential values. This is called privacy preserving data mining which was introduced in (Agrawal & Srikant, 2000) for the case of classification model construction. Further research results have been published on various data mining models for preserving privacy (Evfimievski et. al., 2002; Rizvi & Harista, 2002). Privacy preserving data mining on distributed data sources was another interesting research direction, which was addressed in

(Clifton & Kantarcioglu, 2004) and (Vaidya & Clifton, 2004) for association rule mining and classification model construction. Another aspect of privacy issues in data mining is to protect the data against data mining algorithms. This result is due to the fact that data mining tools can be used to discover sensitive information. Hiding sensitive association rules by database sanitization is proposed in (Verykios et. al., 2004; Saygin et. al., 2004). Further research was conducted for data sanitization to protect the data against data mining tools (Oliveira & Zaiane, 2002; Oliveira & Zaiane, 2003). However there is not much work about preserving privacy for natural language databases and its effects except the studies of Ruch *et al.* (Ruch et. al., 2000) and Sweeney (Sweeney, 1996) who have worked on sanitization of medical reports on a limited domain.

On the other hand, information extraction (IE) has been studied extensively in the natural language processing community. IE is the task of extracting particular types of entities, relations, or events from natural language text or speech. The notion of what constitutes information extraction has been heavily influenced by the *Message Understanding Conferences* (MUCs). The basic information extraction task of *Named Entity Extraction* covers marking names (and determining their types as persons, locations, or organizations), and certain structured expressions (money, percent, date and time; Tur et. al., 1999). An example text from the Broadcast News, whose named entities are marked, is given in Figure 1. These entities can also be marked using XML tags. It may also be useful to make HTML links of entities and co-referencing pronomials, which point to pages, where summary information about the entity, such as the gender of a person, or further references in the text, can be listed.

…The other very big story of the today is in Washington where the White House administration has already been badly shaken up by the possibility that president Clinton and one of his advisors Vernon Jordan obstructed justice…

Figure 1. An example text from Broadcast News corpus. Named entities are marked with colors, green for dates, orange for locations, blue for organizations and red for person names.

A related task to information extraction is the automatic authorship identification, which aims to assign authors to texts. This topic is studied mainly in linguistics, computational linguistics, and stylometrics fields beginning from the pre-computer era (Oakes M., 1998). Commonly used features include vocabulary of the author, usage of some typical words (mostly stopwords) and phrases, and construction style of sentences (e.g., average sentence length, part of speech tags in the first words of the sentences, most common voice, etc.). In our previous work, we have found that using just the word (unigram) frequencies, it is possible to detect the author of a newspaper article by more than 90% accuracy when the candidate set is fixed to 9 authors and using about 100,000 words per author for training (Tur, 2000).

PRIVACY THREATS IN DOCUMENTS

Privacy issues in document repositories arise from the fact that the text contains private information, which can be jeopardized by the adversaries who are curious to know more about individuals for various reasons. Adversaries could turn this information into advantage, such as the private information that may be published in a tabloid.

Since our main concern is document repositories, the main elements we are going to consider are the documents as information sources. There are also what we call Privacy Conscious Entities (PCEs) whose privacy may be jeopardized by the release of

the documents. A PCE could be a person, a company, or an organization. Privacy concerns of PCEs may require that the identifying information of the PCE (person name, company name, etc) should not be seen in a document since the document is related with a sensitive topic. A PCE may not want to be seen as the author of the document, or appear in a document in a certain context, such as being a criminal or being in debt. A PCE may not want to be associated with another entity, such as being a friend of a convict. In doing so, the links between the documents such as references or hyperlinks to a document should also be taken into account.

Type of private information	**Tools that can be used by the adversary**	**Type of threat**	
Explicit	*Browsers, Editors*	*Direct*	
Implicit	*Record Matching, Information Retrieval*	*Via Data Integration*	*Indirect*
	Data Mining Statistical Analysis	*Via Data Analysis*	

Table 1. *Relationships among privacy threats, private information and tools for extracting private information.*

The private information in a text could be grouped into two classes, namely the explicit and implicit information. Explicit information could be the name, salary, or the address of a person that could be viewed by text editors, browsers and can be found by search engines. Implicit information could be the characteristics of a document such as the author, the statistical properties like the frequencies of some words and punctuation marks, or usage of particular

phrases that can be identified with an individual. Data mining and statistics tools are needed to reach such kind of information.

In Table 1, we listed the type of private information that could be obtained by an adversary, the tools that could be used for the corresponding private information type, and the type of threat. As shown in the table, we classify the privacy threats as direct, and indirect. Direct threats occur due to the existence and availability of explicit information in the data such as the name of the person including some extra information regarding the salary of that person. Names of individuals, phone numbers, and salary amounts are just a few examples that form a direct threat to privacy when they are revealed to a third party. Upon the disclosure of the text data, users can see the contents using a browser or an editor. Indirect threats can be of two kinds: one is due to data integration and the other is caused by data analysis. The former type of indirect threats is the integration of different documents in order to infer private information that cannot be revealed by each individual document, when considered alone. The integration can be achieved by finding those documents that are considered "similar" based on the similarity measures defined in the context of information retrieval (Cohen, 2000). Indirect threats via data analysis, instead, are due to the application of machine learning and data mining techniques over the available data. New data mining tools especially for text classification can be used with a training database (which is easy to construct from news groups etc) to infer private information such as the author of a document.

PRESERVING PRIVACY IN DOCUMENTS

In this section we first propose a measure for privacy in document repositories using the notion of *k*-anonymity. We then address the privacy problem in two phases. The former phase is called *sanitization*. It deals with the automatic extraction of named entities, such as sensitive terms, phrases, proper names, and numeric values from a given text. Extracted terms are then replaced with dummy values, or more generic terms depending on the privacy requirements. The latter phase, known as *anonymization*, makes sure that the classification tools cannot predict the owner or author of the text. We should note that all the sanitization and anonymization techniques can be applied to spoken language as well. For example, in the case of anonymization, the purpose may be to hide the identity of the speaker.

k-Anonymity as a Privacy Measure in Documents

Privacy in documents can be assessed using the notion of k-anonymity that has been heavily investigated in the context of statistical disclosure control (Samarati & Sweeney, 1998). *k*-anonymity is related to the identification of entities (individuals, companies etc) in a released data where the confidential attributes are hidden [15, 19]. For example, the data in a hospital regarding patients that includes the patient name, surname, social security number (SSN), postal code (ZIP), birth date, sex, diagnosis, and the medication should not be disclosed without removing the name, surname, and SSN columns which identify a person. However, when we remove the name, surname, and SSN, one may think that ZIP, birth date, sex, diagnosis, and the medication can be safely released without the risk of privacy violation. It turns out that this is not the case because when combined with publicly available data such as voter list, we may recover the SSN, name, and surname information from the voter list database using the ZIP, birth date, and sex columns from the released data. In tabular data sources, a set of attributes (such as the

ZIP, birth date, and sex) is called *quasi-identifier* if it could be used in connection with public data to identify a person (Sweeney, 1996). Quasi-identifiers are used as a base for measuring the anonymity provided by the released data with respect to publicly available data sources. The degree of anonymity is called *k*-anonymity in general and it is formally defined in (Sweeney, 1996) for tabular data as follows:

Definition (*k*-anonymity*). Let $T(A_1, A_2, \ldots, A_n)$ be a table with attributes A_1 through A_n, and QI be a quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI if and only if each sequence of values in T[QI] appears at least with k occurrences in T[QI] where T[QI] denotes the projection on attributes QI maintaining duplicates .*

k-anonymity property makes sure that the disclosed data cannot be used to distinguish a person among a group of k or more individuals. In relational data model, which was heavily adopted by statistical database researchers, the data is structured, and therefore the attribute values related to an individual are in the same row clearly identifying their relationship to the individual. However in case of unstructured text, a major issue is to find the terms that identify an individual, or that are related to an individual. Similar to the definition in (Sweeney, 2002) we define the set of named entities that relate to the same individual and that can be used to identify an entity in relation to other publicly available text as *quasi-identifier named entities*. Quasi-identifier named entities can be date, sex, address and so on. We need to make sure that these terms cannot be used to differentiate between a set of entities, where the set size is k. In case of authorship detection we need to make sure that the author of a document cannot be identified among less than k-authors to provide *k*-anonymity. We define *k*-anonymity in case of the authorship of a document as follows:

Definition (*k*-anonymity of authorship*). Let D_P be a set of documents whose authors are known, D_C be a set of documents whose authorship is confidential, and A be the set of authors of the documents in $D_P \cup D_C$. A document $d_i \in D_C$ satisfies k-anonymity of authorship with respect to D_P, if D_P can not be used to form a prediction model that will reduce the set of possible authors to A_P where $A_P \subseteq A$, and $|A_P| < k$.*

Text Sanitization

The aim of sanitization is to protect the privacy of individuals given their privacy requirements. The first step of the sanitization process is to extract personally identifying information such as the name, SSN of a person, or company name if we would like to protect the privacy of a company. However spotting the personally identifying information may not be enough, we also need to find the quasi-identifier named entities that could be used to identify individuals by linking to other documents such as the ZIP, birth date, and gender.

Sanitization depends on the corresponding task. There are 3 known methods for partial access to databases (Conway & Strip, 1976), which can also be used for sanitization:

1. *Value distortion* alters the confidential values to be hidden with random values.
2. *Value dissociation* keeps these values but dissociates them from their actual occurrence. This can be achieved, for example, by exchanging the values across sentences.
3. *Value-class membership* exchanges the individual values with disjoint, mutually exhaustive classes. For example, all the proper names can be changed to a single token *Name*.

The simplest form of sanitization is modifying the values of named entities or replacing them with generic tokens as in the value-class membership approach. If the named entities are not already marked using XML tags, we can utilize automatic named entity extraction methods, which are well studied in the computational linguistics community. The concept of *k*-anonymity can be assured for text sanitization while determining the generic tokens. For example, people names can be generalized until they map to at least k-people. For the case of numeric values such as salary, a concept hierarchy can be exploited. The salary can be mapped to a more generic value, which refers to at least k people in a specific context. (e.g., low, average, high, and astronomic linguistic hedges in the concept hierarchy) even when quasi-identifier information is used. For the case of addresses, we can ensure that the address maps to k different people in the company or a district where at least k distinct addresses exist.

The generic tokens can also preserve the non-sensitive information to ensure readability of the text. For example, the gender or identification of the people can be marked in the token for the resolution of further (pronominal) references (i.e., <PERSON> versus <PERSON, GENDER=MALE>). An even harder task would be associating references during sanitization as in the example below where <DATE2> is extended as <DATE2=DATE1+3 days>. A sample text and its sanitized version is provided in Figure 2.

Dear Dr. Blue,
Your patient, Mr. John Brown, stayed in our service from 05/05/1999 to 05/08/1999.
Mr. Brown, 72 year old, has been admitted to emergency on 05/05/1999. His tests for the cytomegalovirus and the EBV were negative. Therefore, Dr. George Green performed an abdominal CT scan. Mr. Brown will be followed in ambulatory by Dr. Green…

Dear Dr. <PER1>,
Your patient, Mr. <PER2>, stayed in our service from <DATE1> to <DATE2=DATE1+3>.
Mr. <PER2>, 72 year old, has been admitted to emergency on <DATE1>. His tests for the cytomegalovirus and the EBV were negative. Therefore, Dr. <PER3> performed an abdominal CT scan. Mr. <PER2> will be followed in ambulatory by Dr. <PER3>…

Figure 2. A modified example text from a medical record (Tur et. al., 1999) and its sanitized version.

Another example is the task of automatic call classification, which is an emerging technology for automating the call centers. During the development of call routing systems, previous customer calls are recorded and then transcribed. A human annotator examines the transcriptions, and assigns them a call-type from a set of pre-defined call-types. This data is then used to train automatic speech recognition and call classification components. Figure 3 presents an example dialog between an automated call center agent and a user. As it is clear from this example, these calls may include very sensitive information, such as the credit card and phone numbers, that needs to be sanitized before this data can be shared or made public.

System: How may I help you?
User: Hello. This is John Smith. My phone number is 9 7 3 area code 1 2 3 9 6 8 4. I wish to have my bill, long distance bill, sent to my Discover card for payment. My Discover card number is 2 8 7 4 3 6 1 7 8 9 1 2 5 7 hundred and it expires on first month of next year.

System: How may I help you?
User: Hello. This is <NAME>. My phone number is <PHONE_NUMBER>. I wish to have my bill, long distance bill, sent to my <CREDIT_CARD> for payment. My <CREDIT_CARD> number is <CREDIT_CARD_NUMBER> and it expires on <DATE>.

Figure 3. An example dialog from a automatic call center recording and its sanitized version.

One problem with text sanitization is that the performance of the state of the art information extraction techniques is still far from being perfect (especially when employed for previously

unseen text or domains). In order not to miss any confidential information, one may choose high recall for low precision, which may end up with more number of falsely sanitized portions of text. With the value-class membership approach, the missed named entities will be easy to recognize. Thus, if the named entity extraction performance is low, using value distortion or dissociation methods would be more appropriate. Another problem is the domain dependency of the confidential information. For example, some named entities may be confidential for only some domains, such as drug names in the medical reports vs. pharmaceutical company customer care center recordings, requiring a list of the entities that should be sanitized.

Text Anonymization

Text anonymization aims at preventing the identification of the author (who is also considered to be the owner) of a given text. In case of speech, the speaker is considered to be the owner. Text anonymization is therefore necessary to protect the privacy of the authors. For example, the identity of the reviewers of the scientific papers would be preferred to be anonymous. This is also the case for authors of the papers in blind reviews.

We know that using a state-of-the-art classifier, it is possible to identify the author of a text with a very high accuracy. The features that can be used are the words and phrases (n-grams) in the text, the total number of tokens, total number of different word types, total number of characters, and the number of word types that occur once. We have identified through our experiments with a limited set of articles from a newspaper that each author uses a characteristic frequency distribution over words and phrases. We ~~will~~ use *k*-anonymity of authorship as the privacy metric for anonymization that was defined in Section 4.1. For the anonymization process we may assume a fixed set of documents such as a digital library which collects all the major works of a

given set of authors. In this case, authorship information for some documents are known and some of them are not known. However, we should also consider the case when the adversary is able to find another set of documents for the authors for example by searching the internet, where the number of documents that could be found is practically infinite.

Text classification techniques first parse the text to obtain the features. Each document is represented as a feature vector where each feature may be represented by a real number. In case of a fixed document set, let D_P be the set of documents where the authorship information is public, and D_A be the set of documents where the authorship information is confidential. An adversary could train a classification model using D_P to predict the authorship information of a document in D_A. Since D_P is known and fixed, anonymization can work on both D_P and D_A. The basic approach for anonymization is to modify the documents in D_P and D_A in order to change their feature vectors so that the data mining tools can not classify the document accurately. The most general model that an adversary may use is a classification model that returns probabilities $P(a_j|d_i)$ for each author, a_j, for a given document, d_i. In this way each author will have a certain probability of being an author for a specific anonymous document. The basic approach that can be used for achieving *k*-anonymity is to change the probability of the real author so that (s)he falls into one of top 1…k positions randomly selected among the top-k authors with highest probability. Probabilities are then changed by updating the documents in D_P and D_A. This process should be performed in such a way that the original meaning and coherence of the document is preserved. When D_P is not fixed then the model that could be constructed by the adversary can not be known in advance which complicates the anonymization process. In this case the approach would be to update the anonymous documents in such a way that their feature vectors

look alike to obscure the adversary. We can achieve this by changing the feature vectors such that at least k of the documents with different authors have the same feature vector which can be done by taking the mean of k feature vectors of documents with different authors and assigning the mean as the new feature vector.

The anonymization method heavily depends on the features of the classifier used for authorship identification by the adversary. If the classifier only uses unigram word distributions, then anonymization can be achieved simply by changing the words with their synonyms or by mapping them to more generic terms as done by sanitization. If the classifier uses a different feature set, such as the distribution of the stop-words (such as “the” or “by”) or words from closed class part of speech (word category) tags (that is almost all words which are not noun, verb, or adjective) then revising the sentences would be a solution as in text watermarking (Atallah et. al., 2002). If the classifier uses other features such as passive or active voice, specific clauses, average length of sentences, etc. they need to be addressed specifically. If the text anonymization task has no information about the features of the classifier adversary is using, then the optimal solution would be assuming that it uses all the features we can think of and anonymize the text accordingly.

Discussion of a System for Preserving Privacy in Document Repositories

A system for anonymization and sanitization is depicted in Figure 4. As can be seen in the figure, sanitization and anonymization can be viewed as a layer between the mediums of interaction with the user and document repository. Users may create a document using an editor, and upon his/her request, the document may be sanitized before it is stored. The same process works in the opposite direction as well. When a user wants to view a document, the document could be

sanitized (if it is stored in its original form in a trusted document repository) before it could be viewed.

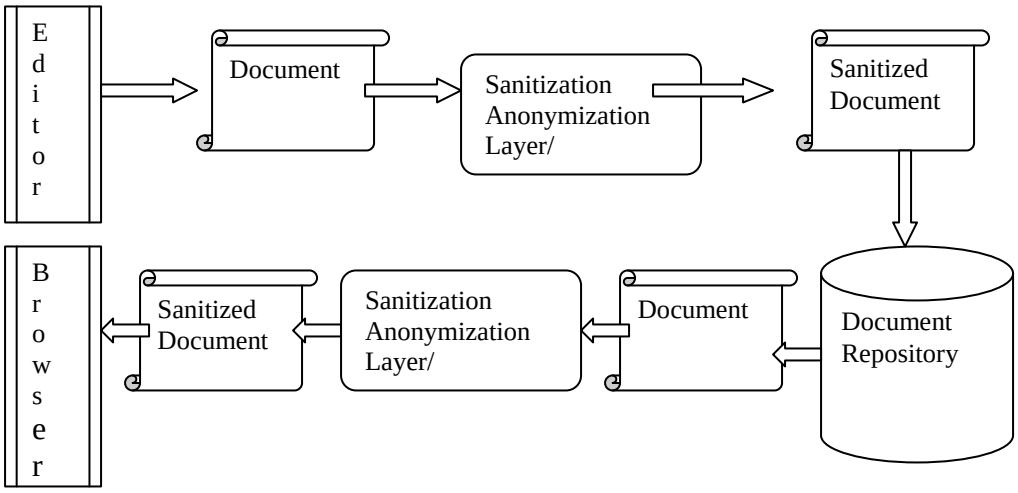

Figure 4. Sanitization and anonymization for documents

Data Quality Measures for Text

We need to make sure that the data quality is still preserved after the sanitization. Data quality could be measured in terms of readability and the ability to use the sanitized text for the corresponding task. For example, if the data is going to be used for text classification, it is necessary to perform sanitization without deteriorating the classification performance. If the task is information retrieval, sanitization methods should not interfere with the indexing and document matching methods.

Information hiding aims at inserting additional information into any kind of digital media (Katzenbeisser and Petitcolas, 2000). This information can be intended to be a concealed

message to be read only for specific people and not by other parties. (steganography), a code identifying/protecting the owner (watermarking), or a code identifying the receiver (fingerprinting). The availability of media in digital form made the unauthorized copying and distribution of these very easy, increasing the concerns and therefore research for protecting the copyright. One recent study on text watermarking digitizes the sentence using its syntactic parse tree and embeds the hidden information into the tree by changing its structure and regenerates a new sentence with the same meaning from this structure (Atallah et. al., 2002). The subject of this chapter is not to insert hidden information into text, instead to hide the sensitive information in the text without distorting its meaning. However evaluation measures can be shared across information hiding, sanitization, and anonymization tasks, since all have the requirement that they should not change the meaning and the coherence of the original text during the update process. Possible information theoretic measures for data quality are Kullback Leiblar distance and change in conditional entropies whose details can be found in (Cover & Thomas, 1991).

FUTURE TRENDS

Standards for privacy policies and privacy preference specifications are being developed under the W3C with the Platform for Privacy Preferences (P3P) Project (http://www.w3.org/P3P/). A method for implementing the P3P standard was proposed using database technology by Agrawal *et. al.* in (Agrawal et. al., 2003). As a future research direction, sanitization and anonymization tools should consider the privacy preferences of users and privacy policies of data collectors. Another important aspect is the development of online techniques for sanitization and anonymization. This is becoming more important especially with the emerging portals with online email scanning capabilities.

Threats that occur due to data integration from multiple sources need further investigation. Simply preprocessing the data may not be enough to ensure privacy. Data integration, record linkage can be used to identify individuals from data sources sanitized by different mechanisms and different policies. For example the same type of text collected from different sources may be released in sanitized form. However, one may sanitize the names, and one may sanitize sensitive data values due to inconsistent sanitization policies. Standardization of sanitization mechanisms and policies for specific data types is needed for ensuring privacy in large scale distributed document repositories.

CONCLUSION

In this chapter, we identified the privacy issues in document repositories and pointed out some approaches to tackle the privacy protection problem. The ideas we presented aim to identify the problem and propose some initial solutions to it combining the existing technology from natural language processing and data mining. The initial ideas we presented will hopefully lead to more research in this direction and the development of tools between the users of documents and the storage medium that will ensure the privacy requirements. With the privacy issues in text identified, tools for protecting the privacy can be developed, which will lead to the release of more text data without the need of money and time consuming text preprocessing done by humans. In sum, text sanitization and anonymization will both ensure privacy of individuals and serve to increase the richness of data sources on the web.

REFERENCES

Adam, N.R., Wortmann, J.C. (1989). Security-Control Methods for Statistical Databases: A Comparative Study. *ACM Computing Surveys*, 21(4), 515-556.

Agrawal, R., & Srikant, R. (2000). Privacy Preserving Data Mining. In *Proceedings of SIGMOD Conference*, 45-52.

Agrawal R., Kiernan J., Srikant R., Xu Y. (2003) Implementing P3P Using Database Technology, In *Proceedings of the 19th International Conference on Data Engineering*, Bangalore, India, March.

Atallah M., Raskin V., Hempelmann C., Karahan M., Sion R., Topkara U., Triezenberg K. (2002) Natural Language Watermarking and Tamperproofing. *Information Hiding*, pp. 196-212.

Clifton C., Kantarcioglu M. (2004) Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. *IEEE Transactions on Knowledge and Data Engineering*. Vol.16 No.9.

Cohen W. (2000) Data integration using similarity joins and a word-based information representation language. ACM Transactions on Information Systems. Vol. 18 No 3, pp 288-321

Conway R., Strip D. (1976) Selective Partial Access to a Database, In *Proceedings of the ACM Annual Conference*.

Cover M., Thomas A. (1991) *Elements of Information Theory*. John Wiley and Sons. New York.

Evfimievski A., Srikant R., Agrawal R., Gehrke J. (2002) Privacy Preserving Mining of Association Rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, AB, Canada, pp217-228.

Hakkani-Tür D., Tur G., Stolcke A., Shriberg E. (1999) Combining Words and Prosody for Information Extraction from Speech. In *Proceedings of the EUROSPEECH'99, 6th European Conference on Speech Communication and Technology*, Budapest, Hungary.

Katzenbeisser S., Petitcolas F. (ed.) (2000). *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House Inc., Norwood, MA.

Oakes M. (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press.

Oliveira S., Zaïane O. (2002) Privacy Preserving Frequent Itemset Mining. In *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining*, Maebashi City, Japan, pp.43-54.

Oliveira S., Zaïane O. (2003) Protecting Sensitive Knowledge By Data Sanitization. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* Melbourne, Florida, USA, pp. 613-616.

Rizvi S., Haritsa J. (2002) Privacy-Preserving Association Rule Mining. In *Proceedings of 28th International Conference on Very Large Data Bases. VLDB*, Hong Kong, China.

Ruch P., Baud R.H., Rassinoux A.M., Bouillon P., Robert G. *Medical Document Anonymization with a Semantic Lexicon*. Journal of American Medical Informatics Association (Symposium Suppl), 2000.

Sweeney L. (2002) k-Anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty , Fuzziness and Knowledge-based Systems, vol. 10 (5) pp. 557-570.*

Sweeney L. (1996) Replacing Personally-Identifying Information in Medical Records, the Scrub System. In *Proceedings of the American Medical Informatics Association Meeting*. Washington, DC: Hanley & Belfus, Inc, pp:333-337

Samarati P., Sweeney L. (1998) Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression, SRI Technical Report.

Saygin Y., Verykios V. S., & Clifton C. (2001). Using Unknowns to Prevent the Discovery of Association Rules. *SIGMOD Record*, 30(4), 45-54.

Tur G., (2000) Automatic Authorship Identification. Technical Report. http://www.research.att.com/~gtur/pubs.html

Vaidya J., Clifton C. (2004) Privacy Preserving Naïve Bayes Classifier for Vertically Partitioned Data. In *Proceedings of the 2004 SIAM Conference on Data Mining*, Lake Buena Vista, Floria, USA.

Verykios, V. S., Elmagarmid, A., Bertino, E., Saygin, Y., & Dasseni, E. (2004) Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering* 16(4), 434-447.