Improving Relevance Effectiveness in Data Leakage Detection Using Feature Selection

Adrienn Skrop
Department of Computer Science and Systems Technology
University of Pannonia
Veszprém, Hungary
e-mail: skrop@dcs.uni-pannon.hu

***Abstract*— Data leakage is an uncontrolled or unauthorized transmission of classified information to the outside. Many software solutions were developed to provide data protection. However, none of them can provide absolute protection. The purpose of the research is to design and implement DATALEAK, a data leakage detection system based on information retrieval models and methods. In this paper, a feature selection based information retrieval model is proposed to improve relevance effectiveness of DATALEAK. The paper focuses on dimensionality reduction, where semantic matching of documents is performed in the reduced form of the vector space model.**

Keywords—data leakage; vector space model; feature selection.

I. INTRODUCTION

Data leakage is an event in which classified information has been viewed, stolen or used by somebody who is not authorized to do so. Data leak prevention methods and systems helps ensure that confidential data remain safe and secure [4][6][8]. However, data leakage detection systems cannot provide absolute protection. On the one hand, more than 40 per cent of data breaches are due to insider negligence [3]. On the other hand, 80 to 90 percent of an organization's data is unstructured. Unlike application oriented data, which is usually well structured and has means of protection, unstructured data is loose, out of control and hard to protect.

In [1][2], DATALEAK, a semantic information-retrieval (IR) based application is presented to address the problem of Web data leakage detection. The system uses a vector space model (VSM) based representation to compare documents. Having a high dimensional VSM, it is impractical to calculate similarity measure. In this paper, we propose an approach to increase effectiveness by reducing the dimensionality of the system.

In Section II, the DATALEAK system is presented briefly. Section III presents the feature selection method that is planned to be implemented in the system. Section IV concludes the paper.

II. THE DATALEAK SYSTEM

The goal of the DATALEAK system is to monitor the Web and collect Web documents according to users' preferences. The collected Web documents are compared with user's confidential documents. If a document turns up on the Web that is semantically similar to confidential user documents the system indicates potential data leakage. The DATALEAK system is composed of the following modules. The Search module is responsible for discovering Web pages that might indicate data leakage. The search module is implemented as a conventional keyword-based metasearch engine. The engine uses the hit list of Google and Bing. The Text mining module is responsible for the automated processing of Web pages that were identified by the Search module. The Text mining module converts Web documents into their appropriate mathematical representation. During automated processing relevant keywords or index terms are extracted from Web documents. Using the extracted keywords Web documents can be represented in a vector space. The document collection contains confidential user documents. The Cryptographic Module is responsible for preparing an encrypted version of these documents. It works similarly as the Text mining module. The input can be any user document. The output is a set of keywords. The keywords are used to create a mathematical representation of user documents. The Scoring module matches the mathematical representations of Web documents and confidential user documents. A number of mathematical models can be used to represent documents and to calculate similarity. In the next section, a reduced dimensional vector space model is presented.

III. DIMENSIONALITY REDUCTION BY FEATURE SELECTION

Usually, the similarity of documents is determined using a repetition-based hard similarity metric. This approach ignores all potential semantic correlations between different words. In DATALEAK, not the pure content, but the meaning of Web documents and user documents are compared.

Given a search query, the retrieved Web documents and confidential user documents, the Scoring module computes a relevance score that measures the similarity between these documents. The scoring module uses the VSM representations of documents. In VSM, documents are represented by a vector in an n dimensional vector space, where n is the number of keywords or index terms [7].

The VSM can be created in three steps. The first step is indexing where keywords are extracted from the documents. Many of the words in a document do not describe the content. These words are called stop words, e.g. the, like, is etc...

 ISBN: 978-1-61208-415-2

By using automatic document indexing these non-significant, usually high frequency words are removed from the document, so the document will only be represented by content bearing words, i.e. index terms.

The second step is the weighting of the indexed terms. A term can be assigned a weight that expresses its importance for a particular document. A common weighting scheme for terms within a document is to use the frequency of occurrence, called by term frequency [9]. The term frequency can be used as a content descriptor for the documents and is generally used as the basis of a weighted document vector [10].

The last step is to compare documents according to some similarity measure. The most popular similarity measure is the cosine coefficient, which measures the angle between the document vector and the query vector.

In the data leakage detection system, the basic VSM representation is as follows.

During indexing, a set of keywords $C = \{c_1, ..., c_i\}$ are extracted from confidential documents and another set of keywords $W = \{w_1, ..., w_j\}$ are extracted from Web documents. Web documents and user confidential documents are represented using the VSM over $C \cup W$ as follows. Given a finite set $T = C \cup W$ of index terms $T = \{t_1, ... ,t_n\}$ any Web document D_j is assigned a vector $\mathbf{v}_j$ of finite real numbers:

$$\mathbf{v}_j = (w_{ij})_{i=1,...,n} = (w_{1j}, ..., w_{ij}, ..., w_{nj}) \quad (1)$$

Confidential user documents U_k also have to be represented as a vector v_k of finite real numbers, as follows:

$$\mathbf{v}_k = (w_{ik})_{i=1,...,n} = (w_{1k}, \ldots, w_{ik}, \ldots, w_{nk}) \quad (2)$$

A Web document D_j is represented to a user having confidential document U_k if they are similar enough, i.e., a similarity measure S_{jk} between the Web document vector $\mathbf{v}_j$ and the confidential user document vector $\mathbf{v}_k$ is over some threshold K. The threshold K should be chosen to represent a required level of lower bound for the similarity of two documents.

The disadvantage of this representation is that, thanks to the Web documents' potential diversity, the dimensionality of the vector space can be fairly high causing costly computation of the similarity measure. A solution to this problem can be to reduce the size of the vector space. Feature subset selection is a technique that is used in supervised and unsupervised classification or regression problems for reducing the attribute space of a feature set. The purpose of feature selection is to identify significant index terms and eliminate irrelevant ones [5]. In this paper, feature selection is proposed to be used in VSM to reduce the dimensionality of the vector space. In order to emphasis the importance of user documents, keywords are extracted only form user documents. These keywords will be used to form the dimensions of the vector space. User documents and Web documents are represented over this limited VSM. Besides reducing the dimensionality of the vector space, feature selection also might contribute to improve precision, i.e., to ensure that the model becomes specialized enough to represent confidential user documents. Fig. 1 visually represents the proposed approach as opposed to the previously presented vector space based matching. We consider that the automatic indexing process has associated a set of keywords $C = \{c_1, ..., c_i\}$ to confidential user documents. The user is allowed to modify the set of keywords by adding new, content bearing keywords $E = \{e_1, ..., e_j\}$, e.g. synonyms; and removing improper keywords $R = \{r_1, ..., r_k\}$.

The result is a VSM over $C + E - R$ as follows. Given a finite set $T' = C + E - R$ of new index terms $T' = \{t'_1, ... ,t'_m\}$, $m < n$, any confidential user document U_k is assigned a vector $\mathbf{v}'_k$ as follows:

$$\mathbf{v}'_k = (w_{ik})_{i=1,...,m} = (w_{1k}, \ldots, w_{ik}, \ldots, w_{mk}) \quad (3)$$

Any Web document D_j is assigned a vector v_j over this new reduced dimensional VSM as follows:

$$\mathbf{v}'_j = (w_{ij})_{i=1,...,m} = (w_{1j}, ..., w_{ij}, ..., w_{mj}) \quad (4)$$

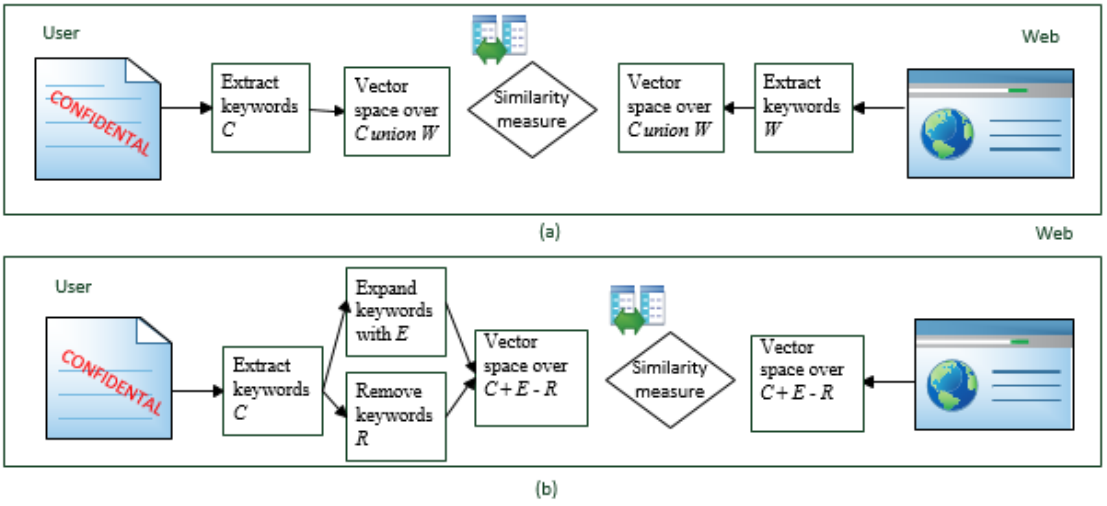

Figure 1. Block diagram showing (a) the vector space based matching, as opposed to (b) the reduced dimensionality based matching problem.

A Web document D_j is represented to a user having confidential document U_k if they are similar enough, i.e., a similarity measure S_{jk} between the Web document vector $\boldsymbol{v}'_j$ and the confidential user document vector $\boldsymbol{v}'_k$ is over some threshold K, i.e.,

$$S_{jk} = s(\boldsymbol{v}'_j, \boldsymbol{v}'_k) > K \quad (5)$$

The expansion of the vector space with new keywords is similar to query expansion. The goal is to improve effectiveness by matching related terms. Instead of adding new keywords manually, a variety of automatic or semi-automatic expansion techniques can be used [11]. Semi-automatic techniques require user interaction to select best expansion terms. In this application, the combination of two techniques is considered to be used. One technique is to use general or domain specific term taxonomies, e.g. WordNet, to determine a set of semantically similar keywords E_1 (e.g. synonyms and hyponyms). Another technique is relevance feedback, which relies on user interaction to identify relevant documents. Having the hit list produced by the Search module, the user indicates which documents are similar enough (relevant) and which documents are non-relevant. Only relevant documents are sent to Text mining module to extract a set of keywords E_2. Finally, the co-occurring keywords, i.e. $E = E_1 \cup E_2$ are selected to be added to the VSM.

IV. CONCLUSION

In this paper, we introduced a vector space based matching approach to address the problem of data leakage detection. The idea is to reduce the dimensionality of the vector space and improve relevance effectiveness by feature selection. Feature selection is based on confidential user documents in order to achieve better precision. Semantic matching of Web documents is performed against confidential documents in the reduced form of the vector space model to reduce complexity.

ACKNOWLEDGMENT

This research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004 - National Research Center for Development and Market Introduction of Advanced Information and Communication Technologies.

REFERENCES

[1] A. Skrop, "Data Leakage Detection Using Information Retrieval Methods," in: Schmidt, A., Yarali, A. (eds.). *IMMM 2014*, The Fourth International Conference on Advances in Information Mining and Management. IARIA. Paris, France, July 20-24. 2014. pp. 74-78. ISBN: 978-1-61208-364-3.

[2] A. Skrop, "DATALEAK: Data Leakage Detection System," MACRo2015, The 5th International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics. Tirgu Mures, Romania, March 6-7, 2015, pp. 115-126. ISSN: 2247-0948.

[3] D. S. Wall, "Organizational security and the insider threat: Malicious, negligent and well-meaning insiders," Technical report, Symantec, 2011.

[4] E. Gessiou, Q. H. Vu, and S. Ioannidis, "IRILD: an Information Retrieval based method for Information Leak Detection," in Proceedings of European Conference on Computer Network Defense, 2011, pp. 33–40, IEEE.

[5] I. Guyon and E. André, "An introduction to variable and feature selection," The Journal of Machine Learning Research" vol(3), 2003, pp. 1157-1182.

[6] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," Knowledge and Data Engineering, IEEE Transactions on, vol. 23(1), 2011, pp. 51-63.

[7] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval: The Concepts and Technology behind Search (2nd Edition). ACM Press Books, Addison-Wesley Professional, 2011, ISBN: 0321416910.

[8] Y. Liu, C. Corbett, K. Chiang, R. Archibald, B. Mukherjee, and D. Ghosal, "SIDD: A framework for detecting sensitive data exfiltration by an insider attack," In System Sciences, 2009, HICSS'09, pp. 1-10, IEEE.

[9] H. P. Luhn, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development 2 (2), 1958, pp. 159-165 and 317.

[10] G. Salton, Gerard and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, 24.5, 1988, pp. 513-523.

[11] E. N. Efthimiadis, "Query expansion," Annual review of informasion systems and technology (ARIST), vol. 31. 1996, pp.121-187.

 ISBN: 978-1-61208-415-2