Data Leakage Detection Using Information Retrieval Methods

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

Abstract— **Data leakage is an uncontrolled or unauthorized transmission of classified information to the outside. It poses a serious problem to companies as the cost of incidents continues to increase. Different solutions have been developed to prevent data loss; however none of them can provide absolute protection due to insider negligence. It is essential to discover data leakage as soon as possible, thus the purpose of this research is to design and implement a data leakage detection system based on different, semantic driven information retrieval models and methods. After describing briefly the idea, the architecture of the system and potential methods are discussed.**

Keywords—data leakage; interaction information retrtieval; hyperbolic information-retrieval; cryptography.

I. INTRODUCTION

Data or information leakage can be defined as an uncontrolled, unauthorized transmission of classified information to the outside or simply an unauthorized dissemination of information. Data leakage can be described by many closely related terms, as the following definitions show. Information leak can be an uncontrolled, unauthorized transmission of classified information from a data center or computer system to the outside. Such leakage can be accomplished by physical removal of data storage devices or even plain old human memory [5]. Data breach is also an unauthorized dissemination of information. It may be due to an attack on the network or outright theft of paper documents, portable disks, USB drives or laptops [12]. An information exposure is the intentional or unintentional disclosure of information to an actor that is not explicitly authorized to have access to that information [13]. Data exfiltration, also called data extrusion, is the unauthorized transfer of data from a computer. Such a transfer may be manual and carried out by someone with physical access to a computer or it may be automated and carried out through malicious programming over a network [11]. Industrial espionage is the theft of trade secrets by the removal, copying or recording of confidential or valuable information in a company for use by a competitor [24]. As the definitions indicate, data leakage can occur in many forms and in any place. Thus, a number of solutions have been developed to prevent data loss. On one hand, encryption may prevent lost or stolen data from being viewed or used by non-authorized individuals. On the other hand, different data leakage products help monitor, manage, and protect data to minimize the risks of data loss and ensure compliance with security policies [26]. However, none of these solutions can provide absolute protection. According to a Symantec study, more than 40 per cent of data breaches were estimated to be due to insider negligence [25]. The purpose of this research is to design and implement a data leakage detection system based on different information retrieval models and methods.

In Section 2, the problem of data leakage is presented. Section 3 shows a potential architecture of a data leakage detection system. Section 4 presents the methods that are planned to be used in the system. Section 5 concludes the paper and presents ideas for future work.

II. DATA LEAKAGE PREVENTION

Data leakage is an incident in which sensitive, protected or confidential data has potentially been viewed, stolen or used by an individual unauthorized to do so. Information and data leakage is on the rise for multiple reasons, e.g., the poorly performing economy, frequent job changes, market advantage achieved by acquisition of trade secrets. This situation poses a serious problem to companies and organizations. The number of leakage incidents and the cost they inflict continues to increase. In most cases, the end product is not as valuable as obtaining the means of production, the research and development, or the know-how. Data loss can be caused by malicious intent or by unintentional mistake. Both cases can diminish a company's brand, reduce shareholder value, and damage the company's goodwill and reputation [15]. Data leakage prevention has been studied both in academic research areas and in practical application domains e.g., [1][16]. A number of methods and systems have been developed to prevent data leakage. For example, in [17], IRILD, an information retrieval based cyclical hashing approach for information leak detection is presented. Cyclical hashing is employed to split the document into multiple parts and generate fingerprints for these parts. This series of fingerprints are checked against the series of fingerprints of outgoing documents. In [18] the problem of giving sensitive data to a set of supposedly trusted third parties is discussed. Data allocation strategies were proposed, that improve the probability of identifying leakages caused third party agents. In [19] a framework is presented for detecting sensitive data exfiltration by an insider attack. However, data leakage detection systems cannot provide absolute protection. Thus, if we cannot

 ISBN: 978-1-61208-364-3

prevent data loss it is essential to discover data leakage as soon as possible.

III. DETECTING DATA LEAKAGE

The purpose of this research is to design and implement a cloud technology based data leakage detection system using different information retrieval models and methods. Cloud-implemented systems and services are available from anywhere and from any device. The only condition is that the device should have Internet access. Further benefits of clouds computing are: reduced IT costs, scalability, flexibility, etc. [23]. The research is expected to result in a system that is suitable for detecting sensitive information on the Web. The implemented data leakage detection system goes beyond the currently available services for comparing the contents of the documents. For example, plagiarism checking services examine and compare the documents word by word to find copy-paste content. These methods have the disadvantage that it can only take into account the words in the documents. In these methods, information about the semantics is not included.

The goal of the data leakage system is to monitor the Web and collect information according to users' preferences. Figure 1 shows the model of the system. The Web data sources are compared with user's confidential documents. Semantically meaningful similarity of data sets might indicate data leakage. Usually, the similarity of documents is determined using a repetition-based hard similarity metric S_H of any two words w_i and w_j:

$$S_H(w_i, w_j) = \begin{cases} 1, if\ w_i = w_j \\ 0, otherwise. \end{cases} \quad (1)$$

This approach ignores all potential semantic correlations between different words. In our system not the pure content, but the meaning of Web documents and user documents are compared. The system may consist of a number of modules. In this section, these modules are introduced briefly. The modules are represented in Figure 2.

The document collection contains sensitive, protected or confidential information. In order to protect these documents, encryption is required. Thus, the Cryptographic module resides on the clients' server. All the other services reside in the cloud. The Cryptographic module is responsible for preparing an encrypted version of the documents. In order to do it, an adequate mathematical model is necessary.

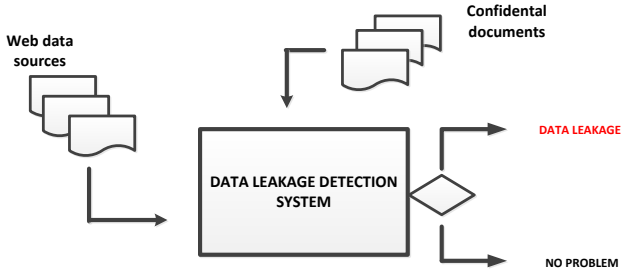

Figure 1. Data leakage detection system.

In information-retrieval the vector space model [20] is the basis of many systems. It is a simple and intuitively appealing framework for implementing term weighting and ranking. In this model, documents are assumed to be a part of an *n* dimensional vector space, where *n* is the number of index terms. In this model, every document is represented by a vector of index terms [6]. Applying the vector space model as the basis of the Cryptographic module the question arises: how to choose index terms. A previous idea is to let the user define index terms.

Text mining module will use the output of the Cryptographic module to define search Queries. Search queries are determined using the predefined vector space. Ontology may be used to add more semantics to search queries. Queries are submitted to the Search module. The Search module is responsible for discovering Web pages and collecting relevant data. It can be implemented as a conventional keyword-based metasearch engine.

Metasearch engines are search engines that search other engines. They submit the search query to several other search engines and return a summary of the results. This strategy gives the search a broader scope than searching with a single search engine [14]. The Search module incorporates a Crawler module that investigates the structure of Web sites, determines those pages of Web sites that contain relevant data, and indexes these pages using keywords. Text mining module converts Web documents into their vector space representations.

The scoring module matches the mathematical vector space representations of Web documents and user documents. Similarity is defined as a kind of distance. Based on this mathematical distance, the system can determine whether Web documents and user documents are close enough. If Web documents are close enough to confidential user documents data leakage warning signs appear. A number of mathematical models can be used to calculate similarity. In Section 4, two different approaches are proposed.

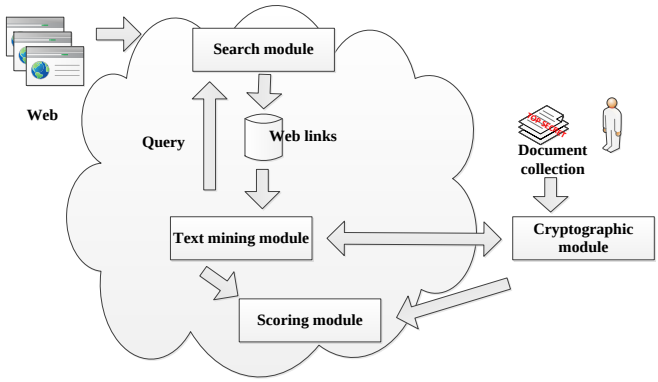

Figure 2. Data leakage detection modules.

IV. METHODS

In this section, methods that can be used in the scoring module are presented. Given a search query, the retrieved Web documents and confidential user documents, the scoring module computes a relevance score that measures the similarity between the Web documents and user document. Different methods will be implemented to satisfy different information needs or user preferences.

Web documents and user documents are represented using the vector space model. In this model, the vector space has to be defined first. On one hand, text mining methods may be used to identify index terms in confidential user documents. On the other hand, users may be asked to define index terms themselves. Considering the sensitive nature of user documents, the last solution may be better. The vector space representation of documents is as follows.

Given a finite set T of index terms $T = \{t_1, ..., t_i, ..., t_n\}$ defined by the user, any Web document W_j is assigned a vector $\mathbf{v}_j$ of finite real numbers, as follows:

$$\mathbf{v}_j = (w_{ij})_{i=1,...,n} = (w_{1j}, ..., w_{ij}, ..., w_{nj}) \quad (2)$$

The weight w_{ij} is interpreted as an extent to which the index term t_i characterizes a Web document. Confidential user documents also have to be represented as a vector. An appropriate safe approach can be to create an artificial document, i.e., the weights of index terms are determined by the user. As a result, a confidential user document is assigned a vector $\mathbf{v}_k$ of finite real numbers, as follows:

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

A Web document W_j is represented to a user having confidential document C_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.,

$$S_{jk} = s(\mathbf{v}_j, \mathbf{v}_k) > K \quad (4)$$

In the classical vector space model (VSM) [3], different weighting schemes and similarity measures can be used, e.g., Cosine measure, Dice's coefficient [21]. However, the categoricity of the system can be varied by both changing the weighting scheme and similarity measure at the expense of a costly computation. To avoid costly computation, we propose the use of a VSM over the Cayley-Klein Hyperbolic Geometry [22].

In the classical VSM, the feature space is mathematically modelled by the orthonormal Euclidean space. In the hyperbolic information-retrieval model, the vector space is defined over the Cayley-Klein Hyperbolic Geometry. In hyperbolic IR (HIR), the similarity measure is derived from the hyperbolic distance. The hyperbolic similarity measure $S_{j,k}$ is defined as follows [4]:

$$S_{j,k} = \sigma_{j,k} = \left(ln\left(e \cdot \frac{r+\sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}}{r-\sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}} \right) \right)^{-1} \quad (5)$$

where

$$r > max_{\boldsymbol{v}_j} d_E(\boldsymbol{v_j v_k}) \quad (6)$$

and

$$d_E(\boldsymbol{v_j v_k}) = \sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}, \quad (7)$$

represents the Euclidean distance of the vectors. Given a VSM based on the Cosine measure using the term-frequency-normalized weighting, this scheme can be replaced with this hyperbolic IR model producing exactly the same answers and ranking. For technical disciplines, the usage of the term-frequency-normalized weighting scheme is recommended as yielding good results [2]. In the hyperbolic model, the categoricity of the system can be varied by only modifying the radius of the hyperbolic space and without using a different weighting scheme and similarity measure. Experiments demonstrated that categoricity in HIR can be varied more than O(n) faster, where n is the number of index terms, than in the VSM [4]. Thanks to the variable categoricity of the measure, the degree of similarity is easily variable in the system. This property can be utilized to vary similarity measure to satisfy different information needs or user preferences.

Besides the VSM, other techniques are considered to be used to determine the similarity of documents. Interaction information-retrieval (I^2R) model [7][8] may prove to be applicable too. Clustering is a well-known technique applied in IR. It is typically used to group documents to be searched. A special case of clustering is adaptive clustering, i.e., a clustering in which the cluster structure is being developed under or is being influenced by an interaction with the user. One way of conceiving adaptive clustering is to adopt a connectionist-based view.

In the data leakage detection system, adaptive clustering can be implemented as follows. Any Web document is represented by an object. An object o_i, i = 1, 2, ..., M, is assigned a set of identifiers. Identifiers are predefined index terms t_{ik}, k = 1, 2, ..., n_i. There are weighted and directed links between any pair (o_i, o_j), $i \neq j$, of objects.

One weight is the relative frequency [7][8] – denoted by w_{ijp} – of a term given a Web document, i.e., the ratio between the relevance r_{ijp} of index term t_{jp} in Web document object o_i, and the length n_i of o_i, i.e., the total number of index terms assigned to o_i:

$$w_{ijp} = \frac{r_{ijp}}{n_i}, p = 1, \ldots, n_j \quad (8)$$

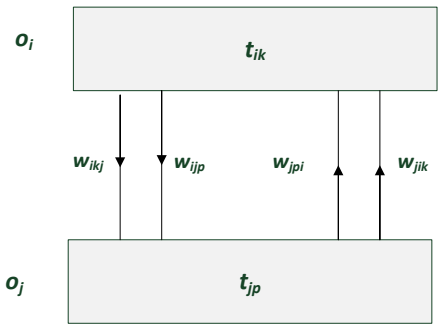

Figure 3. Connections between object pairs.

The other weight is the extent to which a given index term reflects the characteristic of a Web document, i.e., the inverse document frequency [7][8]. If r_{ikj} denotes the relevance of index term t_{ik} in o_j, and df_{ik} is the number of Web documents that can be indexed by t_{ik}, then w_{ikj} is given by the inverse document frequency formula, and thus, represents the extent to which t_{ik} reflects the characteristic of o_j:

$$w_{ikj} = r_{ijp} log \frac{2M}{df_{ik}}. \quad (9)$$

The other two connections - in the opposite direction - have the same meaning as above: w_{jik} corresponds to w_{ijp}, while w_{jpi} corresponds to w_{ikj}. Figure 3 shows the connections between the object pairs.

The Web documents are represented as an interconnected network; every document is linked to every other document. The confidential user document is conceived as being an object, too. It is interconnected with the already interconnected Web documents causing some of the already existing connections to change because of the change of M and df_{ik}. The objects are conceived as being a network of artificial neurons in which a spreading of activation takes place according to a winner takes all strategy. The activation is initiated at the confidential document, and spreads over along the strongest connection thus, passing on to another neuron, and so on. The strength of the connection between any pair (o_i, o_j), $i \neq$ i, of objects, and thus, between the confidential document and another Web document object o_i is defined as follows [7][8]:

$$K_{ij} = \sum_{p=1}^{n_j} w_{jpi} + \sum_{k=1}^{n_i} w_{jik} \quad (10)$$

After a number of steps, the spreading of activation reaches an object that was already affected. This is analogous to a local memory recalled by the confidential document. Those Web documents are retrieved by the system which belongs to this circle. These retrieved documents may indicate data leakage. The corresponding Web documents are ranked in the order of maximal activation. The advantages of the interaction model are as follows [9][10]. On one hand, this method also avoids costly computation. The complexity of weights computation is polynomial. The retrieval process takes polynomial time. On the other hand, the interaction retrieval method allows for a relatively high precision within 50%−70%. Standard test collections based evaluation showed that this method is useful when high precision is favored at low to middle recall values [9].

V. CONCLUSION

Data leakage prevention and protection might ensure that sensitive or confidential information remains safe and secure. Many software solutions were developed to provide data protection. However, malicious attacks and insiders' negligence cannot be completely eliminated. Thus, it is essential to discover data leakage as soon as possible.

In this paper we introduced a semantic information-retrieval based approach to address the problem of data leakage detection. The idea of the system is to monitor the Web, collect information according to users' preferences and indicate data leakage based on semantic similarity of documents. A modular system architecture that composed of Web searching, text mining and scoring was proposed. A connectionist and a hyperbolic IR model were suggested to be implemented in the scoring module, because these models avoid costly computation. The system is under implementation. After implementing the system, experiments have to be carried out to evaluate its effectiveness and precision.

Our future work includes the investigation of automatic summarization methods that can be implemented in the Cryptographic module to extract keywords from sensitive documents. Another open problem is the extension of the Cryptographic module with query expansion techniques so that module can handle semantically related forms of keywords.

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. Shabtai, Y. E. Asaf, and R. Lior, A survey of data leakage detection and prevention solutions. Springer, 2012, ISBN: 978-1-4614-2052-1.

[2] C. T. Meadow, Text Information Retrieval Systems. Academic Press, 2000, ISBN: 0124874053.

[3] G. Salton, "Automatic phrase matching," In: Hayes, DG, Ed., Readings in Automatic Language Processing. American Elsevier Publishing Company, Inc., New York, 1966, pp. 169-188.

[4] J. Góth, and A. Skrop, "Varying retrieval categoricity using hyperbolic geometry," Information Retrieval, vol. 8(2), 2005, pp. 265-283.

[5] M. E. Kabay. *Glossary of Computer Crime Terms*. [Online]. Available from: http://www.mekabay.com/overviews/glossary [retrieved: May, 2014]

[6] 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.

[7] S. Dominich, "Connectionist interaction information retrieval," Information processing & management, vol. 39.2, 2003, pp. 167-193, doi: 10.1016/S0306-4573(02)00046-8.

[8] S. Dominich, "Interaction information retrieval," Journal of Documentation, vol. 50.3, 1994, pp. 197-212, doi: 10.1108/eb026930.

[9] S. Dominich, "Connectionist interaction information retrieval," Information Processing & Management, vol. 39, 2003, pp.167-193, doi.: 10.1016/S0306-4573(02)00046-8.

[10] S. Dominich, A. Skrop, and Zs. Tuza, "Formal Theory of Connectionist Web Retrieval," Soft Computing in Web Information Retrieval, Studies in Fuzziness and Soft Computing, vol. 197, 2006, pp. 163-194.

[11] TechTarget. *WhatIs.com. Definition: data exfiltration.* [Online]. Available from: http://whatis.techtarget.com/ [retrieved: May, 2014]

[12] The Computer Language Company Inc. *Encyclopedia.Definition of: data breach.* [Online]. Available from: http://www.pcmag.com/encyclopedia/term/61571/data-breach [retrieved: May, 2014]

[13] The MITRE Corporation. *Common Weakness Enumeration (CWE) is a list of software weaknesses. CWE-200: Information Exposure.* [Online]. Available from: http://cwe.mitre.org/data/definitions/200.html [retrieved: May, 2014].

[14] W. B. Croft, D.Metzler, and T. Strohman, Search engines: Information retrieval in practice (p. 283). Reading: Addison-Wesley, 2010.

[15] CISCO. *Cisco Data Loss Prevention.* [Online]. Available ffrom: http://www.cisco.com/c/en/us/products/security/email-security-appliance/dlp_overview [retrieved: May, 2014]

[16] Symantec. *Phishing – The latest tactics and potential business impacts.* White paper. Oct 11, 2012, [Online]. Available from: http://whitepapers.itnews.com.au/content22479

[17] 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.

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

[19] 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.

[20] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18(11), 1975, pp. 613-620.

[21] W. B. Frakes, and R. Baeza-Yates, Information retrieval, Data Structures and Algorithms. Prentice Hall, 1992, ISBN: 0-13-463837-9.

[22] J. Bolyai, APPENDIX: The Theory of Space. Akadémiai Kiadó, Hungary, Budapest, 1987, ISBN: 9630515121.

[23] T. Velte, A. Velte, and R. Elsenpeter, Cloud Computing, A Practical approach. McGraw Hill Professional, 2009, ISBN: 0071626956.

[24] InvestopediaUS. *Industrial Espionage.* [Online]. Available from: http://www.investopedia.com/terms/i/industrial-espionage.asp [retrieved: May, 2014].

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

[26] CDW. *Data loss prevention.* Online]. Available from: http://www.cdw.com/content/solutions/data-loss-prevention.aspx [retrieved: May, 2014].

 ISBN: 978-1-61208-364-3