Prevention and Detection of Sensitive Information Exposure by Advanced Context Based Data Mining Techniques

Shubham Pampattiwar

Savitribai Phule Pune University, Pune Institute of Computer Technology, Dhankawadi, Pune, Maharashtra, India

Abstract: ***In today's world, data security has become a major issue of concern to be addressed. Information Exposure or Leakage is one of the prominent way through which data security is compromised. The data exposure may occur accidentally or intentionally. So a model is required to prevent such kind of scenario. Existing methodologies work on specific phrases or are statistical based techniques. Keyword-based methods are not accurate as they do not take context into consideration, and statistical methods ignore the analysis of the text. Thus a new context based approach is introduced to prevent data leakage. This approach anchors the advantages of both the statistical approach as well as keyword based approach. This approach is basically divided into two phases-Drilling stage and Identification stage. The first one deal with the clustering of documents and the latter phase determines the confidentiality. As this approach offers the best of both worlds, it appears that this model is the superior one.***

Keywords: Information exposure, Context, Security, Data mining, Clustering

1. Introduction

Data Exposure is defined as the accidental distribution important data to an unauthorized entity. Sensitive data in companies and organizations include intellectual property (IP), financial information, patient information, personal credit-card data, and other information depending on the business and the industry. Data Leakage is enhanced by the fact that transmitted data (both inbound and outbound), including emails, instant messaging, website forms, and file transfers among others, are largely unregulated and unmonitored on their way to their destinations. Furthermore, in many cases, sensitive data are shared among various stakeholders such as employees working from outside the organization's premises (e.g., on laptops), business partners, and customers. This increases the risk that confidential information will fall into unauthorized hands.[1]

In order to avoid the accidental distribution of sensitive data, Information Exposure Prevention (IEP) systems are introduced. A IEP system helps is avoiding the distribution of important information in wrong hands by monitoring the data and detecting the exposure beforehand and hence plays a crucial role in prevention of Data Leakage. Conventionally there exist only two types of IEPs - Content based and Behaviour based.

The content based IEP involves the Rule and Classifier based approach. In this system the importance of data or the sensitiveness of the data is determined on the frequency of occurrence of some predefined keywords. For instance these systems act as spam detectors for E-mail filtering.
The behaviour based approach is based on the detection and identification of anomalies in behaviour. For instance, a suspected employee's communication in and out of the organisation can be monitored with the hope of finding out some irregular pattern or behaviour.

Consider a simple example, here there is a normal exchange of mails between the Public Relations officer of the organisation and a journalist, but this mail is majorly non-confidential apart from a very small paragraph giving away high importance technical information (data leakage). All the above mentioned methodologies will fail. The content based ones will fail because these systems are way too non-flexible and very rate of raising false alarms, the behaviour based methods will also fail because there was nothing abnormal in the employee's behaviour, it was an official intended mail.

In contrast to the above methods, the context based approach is able to tackle this problem. The proposed method takes the best of both the keyword and classifier-based approaches. First, we use clustering to group together documents of similar content. Then, we extract the confidential content via language modelling. And lastly we determine whether the document contains confidential information or not.

2. Design of the Proposed System

The proposed method majorly consists of two stages. The first stage is the Drilling Stage and the second one is the Identification stage. The Drilling stage consists of grouping of documents via unsupervised clustering and language modelling methodologies. And then during the Identification stage a confidentiality based score is calculated which helps to identify whether the document consists of sensitive information or not.

2.1 Drilling Stage

This stage deals with the representation of the sensitive data of the document. This depiction should consist of the confidential key terms as well the context [2] in which they materialize or frequently appear. There are two inputs needed for this stage: set of sensitive(S) documents and set of non-sensitive(NS) documents. Stemming, Tokenization[3] and removal of stop words is done for all the input documents and eventually these are transformed into weighed vector terms. Clustering of these documents is done via usage of unsupervised K-means algorithm (cosine distance function[4]).

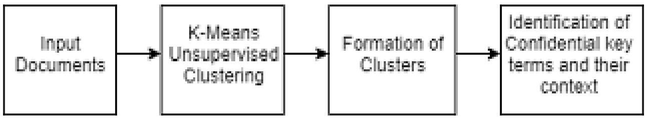

Input Documents
K-Means Unsupervised Clustering
Formation of Clusters
Identification of Confidential key terms and their context

Figure 1: Drilling Stage

The Drilling Stage mainly consists of two phases - Identification of Confidential Key terms and Identification of Context terms.

2.1.1. Identification of Confidential key terms

The sole purpose of Confidential key terms is that they act as initial indicators of important/sensitive information and these terms also serve as the axis around which the context terms are generally materialized.

The Identification of Confidential Key terms is a very important step. This identification is done via a method called the language modelling technique[6]. The language modelling technique enables us to represent the value of Confidential key term in terms of probability. It gives a statistical perception to the methodology.

Initially for each cluster, create a separate language model for sensitive(S) and non-sensitive(NS) documents denoted by LM-S and LM-NS respectively.

$$LM\left(\frac{t}{d}\right) = \frac{tf}{N}$$

where “t” is the term, “d” is document, “tf” is the term frequency and “N” is the total number of terms in the documents.

For each term “t”, calculate a score,

$$score(t) = \frac{LM\text{–}S(t)}{LM\text{–}NS(t)}$$

This score indicates how much more likely the term will appear in sensitive document than in a non-sensitive document. The terms whose score will be greater than one will only be utilised in the Identification Stage. Language modelling performance can be improvised via data smoothing techniques.

2.1.2. Identification of Context Terms

For each of the Confidential key terms, its respective context terms should be identified. The purpose of these context terms[5] is that they act as validators, helping in detecting whether the identified Confidential key terms are precise indicators of sensitive data. These enable us to quantify the significance of the key terms. The analysis of shared context also enables to identify whether the key terms are connected to each other or not.

Firstly, find all the instance of the Confidential key terms (KT) in all the Sensitive as well as Non-Sensitive documents.

For every instance, calculate a context span. A context span is calculated by extracting the terms around the KT using a suitable window size W.(W2 terms before the KT and W2 terms post occurrence of KT) .

Each of these extracts will now be treated as a document (E-S and E-NS).Now create a language model for each Context term in Sensitive and Non-Sensitive.

$$LM(Context) = \frac{No\ of\ Documents\ containing\ Context\ term}{No\ of\ Documents\ containing\ Confidential\ Key\ term}$$

Calculate the probability of each of these context terms appearing near the KT in Sensitive as well as Non-Sensitive documents. Lastly, Evaluate the score of each context term by,

$$score(context\ term) = LM_{E\text{–}S}(context\ term) - LM_{E\text{–}NS}(context\ term)$$

If the value of the score obtained for the context term is more than the pre-defined threshold value, then the context term is Identified as the Context term of the Confidential Key term.

2.2 Identification Stage

The main objective of this stage is to identify the confidentiality of the inspected document. This stage has a few challenges. First being the identification of whole Confidential Document and the other being identification of little parts of confidential text enclosed/embedded in a larger non-sensitive document.

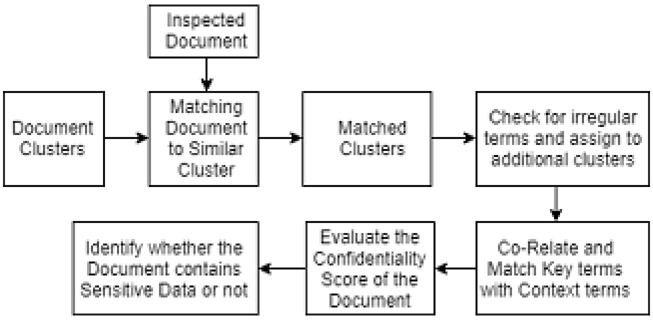

Inspected Document
Document Clusters
Matching Document to Similar Cluster
Matched Clusters
Check for irregular terms and assign to additional clusters
Co-Relate and Match Key terms with Context terms
Evaluate the Confidentiality Score of the Document
Identify whether the Document contains Sensitive Data or not

Figure 2: Identification Stage

After applying tokenization, stemming and removal of stop words, the inspected document is transformed into a vector and finds similar cluster via the cosine distance measure. All the clusters whose similarity is above the predefined threshold are selected for further processing.

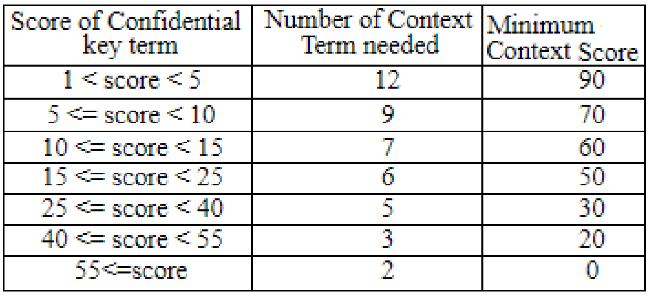
The above methodology works well for identification of sensitive document as a whole but fails for the case when confidential key terms are embedded in non-sensitive documents. In order to tackle this problem we must first identify the irregular terms in the inspected document which are less likely to appear considering the cluster to which the document was assigned. In order to detect such terms, create a language model for both inspected document and the cluster it was assigned to, after this find the score for each term in the inspected document by,

$$\forall\, t \in D \ , \ t_{score} = \frac{p(t/D)}{p(t/C)}$$

where t is the term in the inspected document, C is for Clusters. More the score of the term, more it is less likely to belong to the assigned cluster. If the score for a particular term is greater than a predefined threshold value then the term is an irregular term. After this, the irregular terms are matched against the Confidential key terms of the cluster and if match is obtained then the corresponding cluster gets added to the set of candidate/similar clusters, against which the confidentiality score of the document is calculated.

Finally the confidential score for the inspected document is calculated by adding the scores of all the confidential key terms according to the following criteria:

Table 1: Criteria Table



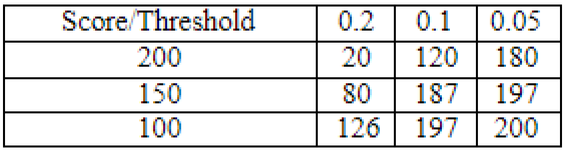
Score of Confidential key term	Number of Context Term needed	Minimum Context Score
1 < score < 5	12	90
5 <= score < 10	9	70
10 <= score < 15	7	60
15 <= score < 25	6	50
25 <= score < 40	5	30
40 <= score < 55	3	20
55<=score	2	0

If the Confidential Key term's score evaluated for the inspected document for a particular cluster is greater than the define threshold, the document is identified as a sensitive and important document.

3. Evaluation of the Design

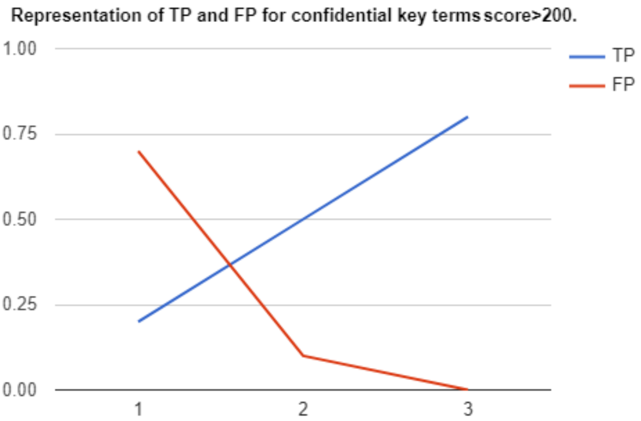
Evaluation is done a dataset generated from the Reuters news articles[7]. This dataset consists of 21578 documents, distributed in 22 files each of the first 21 (reut2-000.sgm-reut2-020.sgm) files consist of 1000 documents and the last one (reut2-021.sgm) consists of 578 documents. This data set has five categories and the category of economics was chosen for evaluation purposes. Economics contains 16 categories and out of which trade was considered as confidential one. Sensitive Documents are 350(200 for drilling and 150 for Identification) and Non-Sensitive are 750(550 for drilling and 200 for identification).

Table 2: Evaluation Table

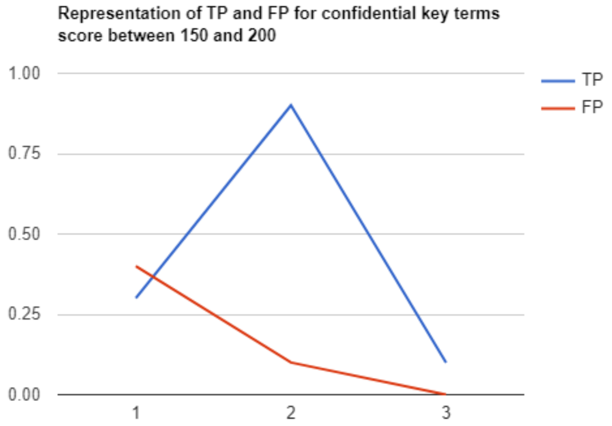


Score/Threshold	0.2	0.1	0.05
200	20	120	180
150	80	187	197
100	126	197	200

The performance was evaluated using True Positive Rate (TPR), False Positive Rate (FPR).TPR is the percentage of sensitive documents correctly identified, FPR represents the percentage of non-sensitive documents mistakenly classified as sensitive documents. The final objective of the method is maximizing TPR and minimizes FPR.



Evaluation Line Graph 1:Score > 200



Evaluation Line Graph 2: Score between 150 and 200

4. Conclusion

The proposed solution is for prevention of information exposure to the unauthorized entities. This method is designed for identification of Documents containing sensitive information and also for identification of small embedded portions of sensitive information in a non-confidential documents. And in both these scenarios, the proposed methodology maximizes the True Positive Rate, Hence an efficient solution.

References

[1] Data Leakage Prevention Implementation and challenges http://www.niiconsulting.com/innovation/DLP.pdf.
[2] Gilad Katz, Yuval Elovici & Bracha Shapira, (2014)"CoBAn: A Context based model for data leakage prevention", Science Direct, Information Science, pp137-158.
[3] Carvalho & Cohen W. W, (2007)"Preventing information leaks in emails", in proceeding of SIAM International conference on data mining.
[4] Salton G & Buckly (1988)"Term-weighting approaches in automatic text retrieval",Information Processing and Management, pp 513-523.

[5] Gilad Katz, Bracha Shapira & Nir Ofek, (2013) "CoBAn: A context Based Approach for Text Classification, http://www.ise.bgu.ac.il/engineering/upload/23944/technical_report.pdf.

[6] Lavrenko, V & W. B.Croft,(1998)"A language modeling approach to information retrieval", In the proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval ACM, Melbourne,Australia,pp275-281.

[7] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[8] J.I Helfman, C.I Isbell & Ishmail, (1995)"Immediate identification of important information", AT&T Labs Technical report.

[9] J.Staddon & P.Golle (2008),"A content-driven access control system", in proceedings of the 7th symposium on identity and trust on the Internet, ACM, Gaithersburg, Marylands, pp26-35.

[10] W.W Cohen & Y.Singer, (1999)"Context-sensitive learning methods for text categorization", ACM transactions on Information sytermmms, pp141-173.

[11] H.Drucker & D.Wu, (1999)"Support vector machines for spam categorization", IEEE transaction on neural networks.

[12] J.Song & H.Takakura, (2013)"Toward more practical unsupervised anomaly detection System", Information Science, pp4-14.

Author Profile

Shubham Pampattiwar has completed Bachelor's Of Engineering in Computer Engineering from Pune Institute of Computer Technology (PICT), Pune in 2017. He is currently an IT professional exploring the practical aspect of Computer Science. His area of Interests includes Machine Learning, Data Science, Predictive analysis, Deep Learning, Information Retrieval and Data Security.