DETECTION OF CONTEXT TERMS FOR A CONFIDENTIAL TERM USING DIRICHLET SMOOTHING TECHNIQUE IN DATA LEAKAGE PREVENTION

[1]Subhashini Peneti, [2]Padmaja Rani B
[1]Research Scholar, Computer Science Engineering, JNTUH, Hyderabad, India.
[2]Professor, Computer Science Engineering, JNTUH, Hyderabad, India.
Email: [1]subhashinivalluru@gmail.com, [2]padmaja_jntuh@jntu.ac.

***Abstract*— Incorporation of the context terms in the representation of the confidential information is important, because it enables better understanding of the confidentiality of each confidential term. Context terms serve as validators, enabling us to determine whether or not the detected confidential terms are truly indicative of relevant content. Context terms allow us to quantify the degree of relevance. In this paper the context of a term is based on a parameter referred to as the context span. This context span helps to detect context terms in the vicinity of a confidential term both in confidential documents in a cluster. Probability is calculated separately for the confidential and non-confidential documents. For each confidential term, all context terms whose score is positive, are considered as its context.**

***Index Terms*—context terms, confidential terms, Context span Context score.**

I. INTRODUCTION

Data leakage is a threat that comprises the unauthorized disclosure of confidential information. This includes any kind of information such as blue prints, source code, financial or investment plans .etc that is essential to achieve the organizations goals [1].Although information leakage can be originated by both insiders and outsiders, information leakage incident originated by user form within the organization are often much more severe than incidents caused by outsiders[2].Information leakage events usually result in loss of competiveness, economic fees imposed by governments and loss of reputation [3].In order to mitigate the risks posed by information leakage incidents, a few security firms have developed a new kind of security tools usually known as DLP(Data leakage prevention) system. Organizations claim these tools to prevent both accidental and malicious information leakage [4].

II. CONTEXT TERMS DETECTION.

This method detects context terms in confidential documents and non-confidential documents. Before detecting the context terms first detect confidential terms in a cluster.

This method detects confidential terms in confidential documents and non-confidential documents. Create clusters for confidential data and non-confidential document [5]. For cluster creation, the content of the every document is represented by a TFIDF vector [6].

TFIDF (Term Frequency Inverse Document Frequency) value of a particular term indicates the importance of the term in the particular document. After generating the TFIDF vector, apply K-means unsupervised clustering algorithm using the Cosine measure as the distance function to generate clusters [5].These clusters may contain confidential documents and non-confidential documents. K-means is an unsupervised clustering algorithm to classify the dataset into K clusters [7].

Cluster creation requires two sets of documents as input .C represents confidential documents contain the complete organization's confidential documents and N represents the non-confidential documents. This step identifies different subjects represented by all documents (both confidential and non-confidential).

In this paper the cosine measure is used to compute which document (document already in the cluster) is closest to a given document (document to be added).

Create language model separately for confidential and non-Confidential documents of the cluster itself. We denote Confidential_LM and Non-Confidential_LM accordingly. Estimate p (Q/M_d), the probability of the query given the language model of document d [8].The maximum likelihood estimate of the probability of term t under the term distribution for document d is [8].iteratively expands the non-confidential model to include a large number of clusters [5].The effect of each term on the language model diminishes as the number of clusters grows.

Find similarity between the original cluster (Confidential_LM) and the candidate cluster to choose the cluster that are added to language model (similar clusters are added first and lower the required similarity threshold every iteration).If the similarity is above the threshold that candidate cluster non-confidential documents are added to original cluster. Calculate score for each term in the confidential language model.

The score assigned to terms reflects how much more likely they are to be appear in a confidential document than in a non-confidential one[5].All terms whose score is above a predefined threshold are selected as confidential terms.

At the time of calculating the confidential score there is a chance to get infinity (term available in confidential-LM and not in non-confidential-LM), to avoid this problem apply Smoothing techniques at the time of language model. The term smoothing refers to the adjustment of the maximum likelihood estimator of a language model so that it will be more accurate [9].The main purpose of the smoothing is to assign a non-zero probability to the unseen words and improve the accuracy of word probability.

There are many smoothing methods are available[10] , all these methods are trying to discount the probabilities of the words seen in the text, and to then assign the extra probability to the unseen words.

The following formula is the general form of a language model [10]:

$$p(w|d) = \begin{cases} p_s(w|d), & \text{if word } w \text{ is seen} \\ \alpha_d p(w|c), & \text{otherwise} \end{cases}$$

Where P (w|d) is the smoothed probability of a word seen in the document, P (w|c) is the collection language model, and α_d is a coefficient controlling the probability sum to one. In general, α_d may depend on d.

Dirichlet smoothing technique

We apply the Dirichlet smoothing technique for unseen words in the document. When Using the Dirichlet prior for smoothing the α_d in the (5) is document dependent. It is smaller for long documents [12], so can be interpreted as a length normalized component that penalizes long document. In Dirichlet prior μ is the parameter. Dirichlet with parameters

$(\mu p\ (w_1|c), \mu p\ (w_2|c), \mu p\ (w_3|c)\ldots\ldots, \mu p\ (w_n|c))$

Where c (w; d) is the document term frequency, μ is Dirichlet prior parameter p (w|c) is the collection language model of a word w,$\sum_w c(w; d)$ is the total of count of words in the document.

After smoothing, calculate confidentiality score for each term in the confidential_LM. Summing up the scores of all confidential terms gives confidential score of the document for each cluster.

Context terms serve as validators, enabling us to determine whether or not the detected key terms are truly indicative of relevant content. They allow us to quantify the degree of relevance.

The process of detecting the context terms for a single confidential term(c_{term}) is as follows:

- Find all the instances of c_{term} both in confidential and non-confidential documents.
- For each instance, extract the terms around c_{term}, using a sliding window of size X ($\frac{x}{2}$ terms before the position of the term and $\frac{x}{2}$ terms after it). Every term in this text excerpt (aside from the c_{term}

itself) will be considered as a possible context term. The size of the sliding window is denoted using a parameter referred to as context span.

- Each such excerpt will now be considered as a document. Excerpts from confidential and non-confidential will be denoted as $d_{confidential}$ and $d\ non_{-confidential}$ documents.
- Calculate the probability of each context term appear near the confidential term c_{term} both in the $d_{confidential}$ and $d_{non-confidential}$ documents. Next calculate the score of each possible context using the following formula:

$$lm_{d_{confidential}}(t_{context}) - lm_{d_{non-confidential}}(t_{context})$$

- The formula for calculating the score of each context term is presented in (7) denotes the probability of a context term to appear near a c_{term} in $d_{confidential}$ and $d_{non-confidential}$ documents.
- If the score of possible context term exceeds a predefined threshold, it will be defined as a context term of the confidential term c_{term}.

Here, unlike in (4), subtract rather than divide the non-confidential probability in 970. The reason behind this is [5] that division can cause large fluctuations in the values of the context terms.

3. Evaluation

In this section, we apply confidential term detection method on Enron emails dataset [14] this dataset contains thousands of emails. Extract each email from the dataset and create separate text document. Extract terms from each email document, remove stop words and apply stemming algorithm. In this paper we are using Porter stemming algorithm. After stemming, the email documents are ready for creating TFIDF vector. Table 1 represents the TFIDF vector representation of the documents.

Clustering is done with K-means clustering method with K value set as 5. A small modification is made in the basic K-means algorithm. In the original K-means if the cosine value between the existing document(document already in the cluster) and the selected document (document to be added) is above the threshold that document is added to that particular cluster, because of this some of the email documents are not matching to any one of the five clusters so that we removed threshold concept .First we calculate similarity value for a document and remaining all added the document to the cluster which is having the maximum cosine value(between 0 and 1).

Next, create language model for all confidential documents and non-confidential documents to each cluster called confidential_LM and non-confidential_LM respectively. With the help of the (4) calculate the confidential score for all terms in the confidential_LM, only the terms whose confidential score is greater than 1 are used to determine the confidentiality of the document. After completion of this step cluster wise a list of confidential terms and confidential scores are available.

Table 1. Cluster1 Confidential Terms and Scores

Term	Score
depart	infinity
peak	0.189
proposit	0.188
messag	infinity
document	1.016

confidential	1.092
sent	2.112
rick	infinity
declar	0.149
interpret	1.092
cats	1.092

Table.2 Cluster1 Context Terms for a Confidential Term **dept** and Scores using smoothing

Term	Score
polit	-0.5
close	0.0
releas	0.3333
larg	0.390
connect	1.0
intern	1.0
seek	-0.5
rick	-0.666
defin	0.56
set	1.0
global	1.0

4. Conclusion and Future work

In this paper we present a method for context terms detection using language model with Dirichlet prior smoothing technique. These terms indicates the confidential content of the document in the organization. At the time of language model creation we add non-confidential documents to the confidential language model this helps to identify the confidential terms in non-confidential documents.

In future, a graph can be created for confidential terms and corresponding context terms, cluster wise. Context terms serve as validators, enabling us to determine whether or not the detected terms are truly indicative of relevant content. Context terms allow us to quantify the degree of relevance (many context terms=higher relevance).Graph representation is capable of capturing the structure of the document as well as its content. These graphs represent the confidential content only and not the whole document

References

1. Jorge Blasco, Julio Cesar Hernandez. Bypassing information leakage protection with trusted application. Science Direct Computer & Security. Volume 31, Issue 4, Pages 557-568.
2. Moore Ap,Cappelli DM,Caron Tc.Insider theft of intellectual property for business advantages: A preliminary model. First international workshop on managing insider security threats. West Lafayette, USA: Purdue University; 2009, p.1-22.
3. Rantala RR.Cybercrime against business, 2005.Technical Report.U.S.Department of Justice.2008.
4. McCromic M. Data theft: A prototypical insider threat. Insider attack and cyber security 2008.p.53-68
5. Gilad, Katz, Yuval, Elovici, Bracha, Shapira. CoBAn: A Context based model for data leakage prevention. Information Science 262(2014) 137-158.
6. Zilberman, p, Shabtai, A. Analyzing Group Communication for Preventing Data Leakage via Email .IEEE, 2011.
7. http://croce.ggf.br/dados/K%20mean%20Clustering1.pdf
8. Ponte, J., M., Croft, W., B., A language modelling approach to information retrieval. In proceeding of the 21st annual international ACM SIGIR conference on research and development in information retrieval.1998, ACM:Melbourne,Australia.pp. 275-281.
9. Lavrenko, V., Croft W., B., Relevance based language models. In proceedings of the 24thannual international ACM SIGIR conference on research and development in information retrieval. 2001. ACM: New Orleans, Louisiana, United States.pp.120-127.
10. Zhai, C., Lafferty J. A study of smoothing methods for language models applied to adhoc information retrieval. In proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval .2001.ACM.
11. Chen, S., Goodman, F. An empirical study of smoothing techniques for language Modelling. Tech Report.TR-10-98.Harward University.
12. https://www.cs.cmu.edu/~./enron