Data Leakage Prevention System with Time Stamp

Subhashini Peneti
Research scholar, Computer Science Engineering, JNTUH
Hyderabad, Telangana 500 085
subhashinivalluru@gmail.com.

B.Padmaja Rani
Professor, Computer Science Engineering, JNTUH
Hyderabad, Telangana-500 085
padmaja_jntuh@jntuh.ac.in.

***Abstract*---Because of huge usage of data, necessity of the Data leakage prevention is growing day by day. Data Leakage Prevention system decided that particular data (confidential or non-confidential) is permitted to access or not. In Data leakage Prevention, time stamp is very important for giving permission to access a particular data, because in a particular period of time the data is confidential after the time stamp the same data could be non confidential, here we developed an algorithm for data leakage prevention with time stamp.**

***Keywords*—: Data Leakage; Data states; Data leak channels; Data Leakage Detection and Prevention.**

I. Introduction

In information security data leakage threat has become an important issue especially data leakage caused by insider threat [1]. Most of the computer attacks are from authorized users of the system. With the wide spread of internet the insider threat is more serious.

Sending confidential data to an unauthorized party called as a data leakage. To prevent the data leaving from the outside of the organization private network called as data leakage prevention. Data leakage prevention system is collection of sub systems which helps to identify the confidential data and also prevents the data leakages [1].

Data leakage prevention systems consider two parameters for preventing data leakage. One parameter is data states. In general data is available in three states i.e. rest state, use state and move state another parameter is deployment scheme i.e. where we are deploy our Data Leakage Prevention system Based on these two parameters the Data Leakage Prevention solutions are changed.

II. Time Stamp Based Data Leakage Prevention

In Data Leakage Prevention identification of confidential data is very important, along with the data one more parameter i.e. time stamp also considered as an important aspect in the Data leakage Prevention. For example in an educational system question paper is confidential until on or before examination date once exam over that is public and treated as a non-confidential [2].

In our time stamped based DLP two phases are there
Learning Phase
Detection Phase

A. Learning Phase

In Learning Phase the documents are trained as confidential documents with time stamp. Fig 1 represents the pictorial representation of learning phase.

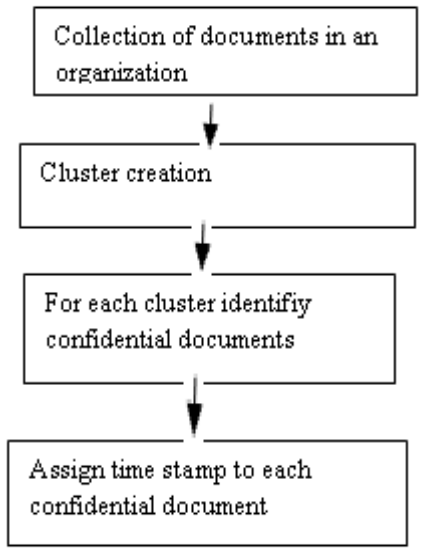

Fig. 1. Learning Phase

Algorithm for confidential documents with time stamp [3]

978-1-5090-2552-7

Input:
Document corpus

Output:
Confidential Documents with time stamp

1. Collection of Confidential and non confidential documents of an organization
2. Create clusters using Kmeans with cosine similarity function
3. For each cluster identify the key terms based on their frequency.
4. For each key term calculate the score
5. Assign time stamp for a document based on deadlines of organization schedule

1. Collection of Confidential and non confidential documents of an organization

For develop an effective Data leakage Prevention solution the organization should manage the data in proper manner i.e. which is confidential and non confidential... Our method uses both confidential and non confidential documents .These two are the inputs for our method, at the time confidential terms identification we construct probability for both confidential and non confidential documents.

2. Create clusters using Kmeans with cosine similarity function.

The data available in the organization is unlabeled data we should divide that data into a groups called clusters. We used Kmeans with cosine similarity for cluster creation. K means is an unsupervised algorithm, based on the k value initially it creates k clusters and randomly assigns documents to each cluster and iteratively assign documents to the clusters based on the cosine similarity value.

Kmeans algorithm

Input:
Training dataset
K value

Output
K clusters

1. Randomly we choose k documents as initial documents.
2. Place k documents in k clusters
3. For each document in the training data set
4. find cosine similarity value for training dataset document and cluster document
5. If cosine similarity is greater than a threshold add training dataset document to particular cluster
6. Calculate the centroid of the cluster.
7. End for.

3. For each cluster identify the key terms based on their frequency

This is the main steps for our method, for each cluster we should identify the key terms using the concept called language modeling technique. In this method we used a formula for a language model i.e. the term frequency of a particular term divided by total number of terms in that document. For each cluster language model is created for both confidential and non confidential documents. If the language model value of a term is greater than a threshold value then consider that term as a key term.

4. For each key term calculate the score

For each key term in a cluster find the score using following formula

Score (key term) =
Language model value in confidential document divided by language model value in non confidential.

5. Assign time stamp for a document.

Based on the deadlines of the organization assign time stamp for each document

B. Detection Phase

In the detection phase the tested document is compared with confidential score and time stamp, if the time stamp of the tested document is greater than or equal to the time stamp in the above table then that document is treated as a confidential and it is blocked. Fig 2 represents the pictorial representation of detection phase.

Algorithm for detection phase [4]

1. D- document to be test
2. Identify similar clusters using cosine similarity
3. For each cluster identify documents
4. For each document check the timestamp
5. If time stamp>=time stamp set in the learning phase
6. Calculate confidential score of a document
7. If score > threshold
8. D is a confidential document
9. Else D is a non confidential document

1. In our method the documents are in text format, so we perform data preprocessing for all documents in training dataset and tested dataset. As a part of data preprocessing apply stop words (unnecessary words like is, the, numbers…) and stemming algorithm.
2. With the help of Cosine similarity function identify the similar cluster for the tested document.
3. Extract the documents from the similar clusters. Now the documents contain confidential score and time stamp.
4. The current date of the tested document is compared with the time stamp already available in the documents extracted from the step 3.
5. If the time stamp greater or equal in the learning phase documents then calculate the score of tested document if score greater than a threshold .
6. The tested document considered as a confidential and could be blocked for sending to outside.

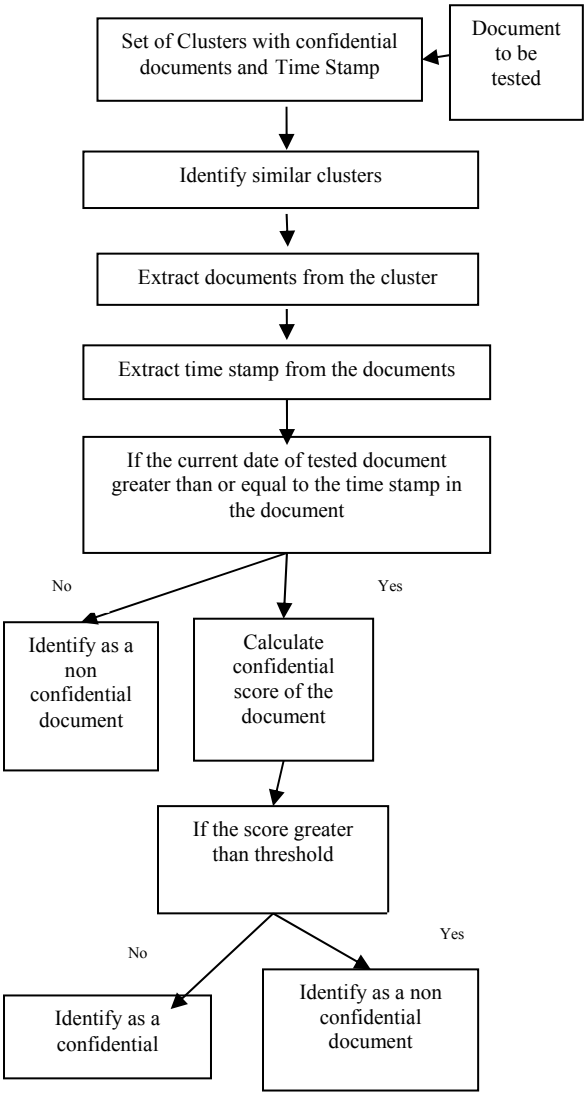

Fig. . 2 Detection Phase

III. Evaluation

We evaluate our method on two data sets i.e. Enron data set and Reuters news article dataset.

A. Enron email dataset

This dataset contains different folders of 150 users.
In this dataset normal meeting emails are considered as non-confidential documents, remaining all emails is considered as confidential documents *[5]*. For each document first we performed tokenization, remove stop words and apply stemming algorithm.
After data preprocessing calculate TFIDF value for each term in the confidential and non confidential document

.Clusters are created with TFIDF value using Kmeans with cosine similarity function. We made a small modification in the basic kmeans algorithm

For example assume k value is 3 and training dataset are d1,d2,d3,d4,d5,d6,d7,d8,d9,d10.

Randomly we choose d3 d7 d9 as initial documents
C1=d3
C2=d7
C3=d9

Find cos (d3, d1) = 0.78
cos (d7, d1) = 0.34
cos (d9, d1) = 0.45
add d1 to cluster 1 because d1 is more similar to C1 compared with C1and c3.

TABLE1. CONFIDENTIAL DOCUMENTS WITH TIME STAMP

Document Name	*Confidential score*	*Time stamp*
Email_1.txt	0.986	30/01/2016
Email_2.txt	0.976	23/02/2016
Email_3.txt	1.245	25/02/2016
Email_4.txt	1.008	10/10/2010

B. Reuters News Article dataset

In the year 2000 Reuters made available a large collection of Reuters news stories for use in research and development. [6] Dataset contains 21578 documents of different categories, for our evaluation we used Economic category. Economic consist 16 categories, out of 16, the TRADE category we used as confidential purpose remaining 15 are used as a non-confidential.

TABLE 2. CONFIDENTIAL DOCUMENTS WITH TIME STAMP

Document Name	*Confidential score*	*Time stamp*
Reuters_Economic_1.txt	0.986	30/01/2016
Reuters_Economic_2.txt	0.976	23/02/2016
Reuters_Economic_3.txt	1.456	08/09/2017
Reuters_Economic_4.txt	1.839	30/06/2005

The performance of our method is evaluated using TPR (True positive rate) and FPR (False Positive Rate). TPR is how many confidential documents are identified correctly as a confidential and FPR is how many non confidential documents are wrongly identified as a confidential, the ultimate goal or our method is maximizing TPR and minimize FPR.

Iv. Conclusion

Data Leakage Prevention with time stamp method best suited for both large and small dataset. Our method used by different application where we need to match the content of the documents. Documents with complete confidential or non-confidential content are 100% detected by our method. In future we extend our method for detecting small portions of confidential content in non-confidential documents.

References

[1] Shabtai, A., Elovici, Y., Rokach, L: A survey of data leakage detection and prevention solutions. Springer Briefs in Computer Science, Springer, 2012.

[2] Katz, Gilad. , Elovici, Yuval. , Shapira, Bracha: CoBAn: A Context based model for data leakage prevention. Information Science 262(2011) 107-128.

[3] Zilberman, Polina. , Shabtai, Asaf. , Rokach, Lior: Analyzing Group Communication for Preventing Data Leakage via Email .IEEE, 2011.

[4] Ponte, J., Croft, W: 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.

[5] https://www.cs.cmu.edu/~./enron/

[6] https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection