# Content Based Confidentiality Detection Method For Data Leakage Prevention

**Subhashini Peneti, Padmaja Rani B**

***Abstract— protecting confidential data became a challenge for all private and public organizations. According to Gartner report, the majority of data leakages in organizations are due to internal factors. Data Leakage Prevention Systems can protect monitor and identify the confidential data at-rest, in-use and in-motion. This paper presents a Data Leakage Prevention system, to prevent confidential data from leakages using the Term Based Confidentiality Detection Method .The proposed method consists of two phases: training and testing phase. The training phase identifies confidential terms from the documents and testing phase detects the confidentiality of the document.***

***Keywords— confidential, data, data leakages, security, organizations.***

## I. INTRODUCTION

Now a day's modern enterprises depends on data sharing in both inside and outside the organization. Data sharing increased the data leakages [1]. The data which causes data leakages are SSN (social security numbers), medical records, organization financial information and trade secrets. In order to prevent confidential data from the leakages, organizations should focus of their internal hosts and network parameters. Organization security measures can also prevent confidential data from classic threats like viruses, Trojan horses, worms, D/Dos attack and intrusions. In addition organizations are required to meet the terms and regulations of the state and central government. Data leakage is defined as the accidental or unintentional distribution confidential data to an unauthorized entity [2].In information security, identification of data leakage attack has become a critical issue for every organization [2]. Most of the data leakage attacks are caused from employees of the organization. Organization must keep a close monitor to secure their confidential data. Data Leakage Prevention (DLP) solution is one of the recent methodology and technical solution for confidential data security .Data leakage prevention solution provides security to the confidential data from the inside and outside the network. Content-aware DLP solution is a data leakage prevention solution that involves awareness of the content that is being protected. The content-aware DLP solution protects organizations from data leakage threats.Content-aware DLP solutions are able to read the entire content of the document and identify confidential data, including the text found in the organization's documents. Content-aware DLP solutions detect and prevent the confidential data which is available in motion, in use and at rest [13].

## II. RELATED WORK

Different types of DLP solutions are available in the market, some of the solutions protecting confidential data by providing tag to the data i.e. the organization documents are classified in to confidential and non-confidential with tags(confidential and non-confidential), if any employee in the organization sends a new document then identify the tag of the document. The document is blocked, when the document contains a confidential tag otherwise forwarded to the network. DLP solutions are categorized based on the state of the data. In a DLP the data is available in three states [3]: DAR(Data-At-Rest), DIU(Data-In-Use) and DIM(Data-In-Motion). DAR is defined as the whole amount of data available in the organization's data centers. The following table represents the summary of the DLP solutions for Data-At-Rest.

**TABLE 1. SUMMARY OF THE DLP SOLUTIONS FOR DATA-AT-REST**

| Category | Method | Proposed by, year | Issues | References |
|---|---|---|---|---|
| Misuse detection in Information retrieval system | User profiling based on query results and relevance feedback | Rebecca Cathey, Nazil Goharian and David Grossman,2003 | Requires huge amount of profiling | [4] |
| | User profiling based on keyword in queries and search results(relevance feedback) | Ling Ma and Nazil Goharian,2005 | Requires regular administrative references | [5] |

**Revised Manuscript Received on 14 September, 2019.**

**Subhashini Peneti** Assistant Professor Department of CSE Malla Reddy College of Engineering,Hyderabad, India
.**Padmaja Rani B,** Professor, Department of CSE, Jawaharlal Nehru Technological Universty Hyderabad, India.

| | | | | |
|---|---|---|---|---|
| | Assigning privilege levels to users and security level to documents and monitoring user access to documents. | Hyeran Mun, Kyusuk Han and Kwangjo Kim,2014 | If a leak is happening the method is ineffective | [6] |
| Misuse detection in Database | Using machine learning methods to detect abnormal access behavior by analyzing query syntax. | Ashish Kamra, Evimaria Terzi, and Elisa Bertino, 2008 | Lacking query semantics. | [7] |
| | Using machine learning methods to detect abnormal access behavior by analyzing query result-set. | Sunu Mathew, Michalis Petropoulos, Hung Ngo, Shambhu Upadhyaya,2009 | Proposed method work for static database i.e. no updates | [8] |
| Data hidden in files | Awareness | Simon Byers,2004 | | [9] |
| Encryption and access control | Encryption and access control. | Imad M. Abbadi, Muntaha Alawneh,2008 | Encryption can secure confidential | [10] |
| | Application of selective encryption to provide selective access control to outsourced confidential data by third-party partners. | Sabrina De Capitani ,Di Vimercati, Sara Foresti,2010 | Does not detect data leak | [11] |

DIU is the data that user is interacting. The following table represents the DLP solutions for DIU.

**Table 2. Summary of the dlp solutions for data-in-use**

| Category | Method | Proposed by, Year | Issues | References |
|---|---|---|---|---|
| Detecting Malicious Insiders using Honey pots | Web-based service generated and distributed decoy documents to registered users and | Brian M. Bowen, Shalom Hershkop, Angelos D. Keromytis and Salvatore J. Stolfo, 2009 | Produces overheads | [12] |
| | Insertion of a honey table into database which is able to attract malicious users. | Antanas Čenys, Darius Rainys, Lukas Radvilavičius, and Nikolaj Goranin,20 | Honey table Creation | [13] |
| | Distributes confidential data objects to several agents | Maya Bercovitch, Meir Renford, Lior Hasson Asaf ,Lior Rokach and Yuval | Limited to specific situations or scenarios. | [14] |

The data that is being transmitted through the network is called as Data in Motion (DIM). The following table presents the DLP solutions for DIM.

**Table 3.Summary Of Dlp Solutions For Protecting Data-In-Motion**

| Category | Method | Proposed by, year | Issues | References |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| | Different rules are framed , these rules determine the | W.W. Cohen ,1996 | Produce high False Positive Rate | [15] |
| | Automatic text classification algorithms for classifying enterprise documents as either confidential or non-confidential | Michael Hart, Party's Manadhata and Rob Johnson,2011 | Not able to classify encrypted and multimedia content | [16] |
| | Analysis of the email exchange among groups of members of the organization to | Polina Zilberman, Asaf Shabtai and Lior | False positive leaks because of a short history between a | [17] |
| | Context based confidentiality detection method(CoBA | Gilad Katz, Yuval Elovic | Produce high False Positive Rate | [18] |
| Network/Web-based protection | Monitoring user's access to information on an intranet using network-based sensors which generates information-use events. These events are combined with contextual information and processed by various rule-based and statistical detectors that may issue alert. | Caputo,2009 | Not able to classify encrypted and multimedia content | [19] |

It is observed that, the above all methods which are proposed for email leakage protection are suffered from a high false positive rate. In contrast to the above methods, we proposed a method to prevent confidential data from the leakages with low false positive. The proposed method "Term Based Confidentiality Detection Method" protects data in-motion and used content-based inspection for preventing data leakages. CoBAn (**Co**ntext **B**ased **An**alysis) method is taken up as a baseline method[18].

## III. TERM BASED CONFIDENTIALITY DETECTION METHOD

Term Based Confidentiality Detection Method is a content-based confidentiality detection method which identifies the confidentiality of the test document based on the content of the document. This method classifies organizational documents into two classes: Confidential and Non-confidential by using the content of the document.

TBCDM consists of two phases: training and detection phase. During the training phase, clusters of documents with confidential terms are generated, in the detection phase, each tested document is assigned to clusters and its content are then matched to each cluster's respective confidential terms to determine the confidentiality of the document.

The proposed method maintains the confidential and non-confidential documents separately and identifies the confidential data through confidential terms. Confidential terms serve as an initial indication of the presence of confidential content in the document. The proposed method uses language modeling technique for confidential terms identification.

Term based confidentiality detection method consists of two phases

1) Training phase and
2) Testing phase.

### *A. Training Phase*

The objective of the training phase is to represent the confidential content of documents using the training repository. The training repository consists of both confidential and non-confidential documents of the organization.

The main operation in the training phase is language model creation. The language model is created separately for confidential clusters set and non-confidential clusters set. Language model consist of the following steps:

1. Document pre-processing
2. Unsupervised cluster creation
3. Confidential terms detection

#### *a) Document Pre-processing*

In general the document repository is not available in understandable format; with the help of data pre-processing the document repository is converted into understandable format [21].

#### *b) Unsupervised Cluster Creation*

To find the confidentiality of a document, we need to identify the various subjects represented by organization documents. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled. The most common unsupervised learning method is cluster analysis [22].

For cluster Creation, the content of the every document is represented as a vector using TF-IDF (Term Frequency-Inverse Document Frequency).
In information retrieval, TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection[20].

TF-IDF (term) = TF (term) *IDF (term).

Where TF is the term frequency of a term in a document and IDF is the inverse document frequency of a term.

After TF-IDF calculation of each document in the training repository, the next step is to generate confidential and non-confidential clusters. For cluster creation K-means unsupervised algorithm with cosine measure as the distance function is used.

*c) Confidential Terms Detection*

The objective of the confidential terms detection method is to identify the terms, which indicate with a high probability in a confidential cluster and low probability in a non-confidential cluster, we referred these terms as confidential terms [23].
The Confidential terms have two purposes:

- They serve as initial indicators of relevant content; and
- They help to find context terms.

Our first aim is to find the terms with a high probability of appearing in confidential documents and a low probability of appearing in non-confidential documents. In order to find the confidential terms, performs two operations:

1. Probability calculation of each term in confidential and non-confidential clusters (CR, NR).
2. Confidential score calculation for each term of CR

*1) Probability Calculation*

The probability of a term indicates the occurrences of a term in the cluster [83]. The higher the probability of a term indicates the more occurrence of the term in the cluster. Calculate the probability of each term available in the confidential and non-confidential clusters using the following formula,

$p_{lm}(t/c)=(t_f)/N(t,c)$ (1)

Where $p_{lm}(t/c)$ is the probability of a term t in cluster C,$tf_t$ is a term frequency of a term t in a cluster C and N (t, c) is the total number of terms in C. In a cluster if a term appears more than one time then add all the probabilities of a term and maintain only one probability .Once the probability calculated for all terms of confidential and non-confidential clusters, calculate confidential score for each term of CR.

*2) Confidential Score Calculation*

To calculate the confidential score of a term we need two inputs: Confidential clusters with term probabilities ($CR_{LM}$) and Non-confidential clusters with term probabilities ($NR_{LM}$). In order to find the occurrence of a term in non-confidential and evaluating its effect towards confidentiality can be measured by finding similar Non-Confidential clusters to a corresponding confidential cluster. The term occurrence and its weight age among similar non-confidential documents can be measured in one of the following three ways. Three methodologies are proposed for finding the confidentiality of a term.

1. In method I, the combined effect of a term occurrence in non- confidential clusters compared to term occurrence in a confidential cluster is considered. Hence, the sum of all probability of occurrence of a term in non-confidential is considered because if the term appears in more than one non-confidential cluster its combine effect on confidentiality is less.
2. In method II, the average effect of a term occurrence in all non- confidential clusters compared to the term occurrence in a confidential cluster. For this reason the average of all probability of occurrence of a term in non-confidential is considered.
3. In method III, the effect of a term occurs maximum in non-confidential clusters that are similar to confidential cluster is considered. Hence the maximum occurrence of a term out of all probability occurrence of that term in non-confidential cluster compared to confidential cluster is considered.

**Method I:** Confidentiality of a term is measured in terms of term probability of occurrence in confidential documents divided by the summation of term probability of occurrences in all similar non-confidential clusters.

$$\left[ confidential\_score(t) = \frac{p\left(t/_{cr}\right)}{p\left(t/_{nr1}\right)+ p\left(t/_{nr2}\right)+ p\left(t/_{nr3}\right)+.....+ p\left(t/_{nrn}\right)} \right] \quad (2)$$

Where Cr is the analyzed confidential cluster, is the probability of a term t in Cr, nr is the non-confidential cluster that is similar to Cr and $p\left(t/_{nr}\right)$ is the probability of a term t in nr.

Let $p\left(t/_{cr}\right) = p_{cr}$ (3)

$$p\left(t/_{nr1}\right)= p_{nr1}, \; p\left(t/_{nr2}\right)= p_{nr2}$$
$$, p\left(t/_{nr3}\right)= p_{nr3}, ..... \; p\left(t/_{nrn}\right)= p_{nrn} \quad (4)$$

Using eq. (3) and eq. (4)

$$confidential\_score(t) = \frac{p_{cr}}{p_{n_{r1}} + p_{n_{r2}} + p_{n_{r3}} + .....+ p_{n_{r3}}} \quad (5)$$

$$confidential\_score(t) = \frac{p}{\sum_{i=1}^{n} p_{nri}} \quad (6)$$

**Method II:** The confidentiality of a term is measured in terms of term probability of occurrence in confidential documents divided by an average of term probability of occurrence in all non-confidential clusters that are similar to a corresponding confidential cluster.

$$confidential\_score(t) = \frac{p(t/cr)}{\frac{p(t/nr1)+p(t/nr2)+p(t/nr3)+.....+p(t/nrn)}{N}} \quad (7)$$

Where Cr is the analyzed confidential cluster, $p(t/cr)$ is the probability of a term t in Cr, nr is a non-confidential cluster that is similar to Cr, $p(t/nr)$ is the probability of a term t in nr and N is the number non-confidential clusters that contain term t.

Using eq. (3) and eq.(4)

$$confidential\_score(t) = \frac{p_{cr}}{\frac{p_{nr1}+p_{nr2}+p_{nr3}+.....+p_{nrn}}{N}} \quad (8)$$

$$confidential\_score(t) = \frac{N * p_{cr}}{p_{nr1}+p_{nr2}+p_{nr3}+.....+p_{nrn}} \quad (9)$$

**Method III:** Confidentiality of a term is measured in terms of term probability occurrence in confidential documents divided by maximum value of term probability occurrence of a term t in all similar non-confidential clusters.

$$confidential\_score(t) = \frac{p(t/cr)}{Max[p(t/nr1), p(t/nr2), p(t/nr3),....., p(t/nrn)]} \quad (10)$$

Where Cr is the analyzed confidential cluster, $p(t/cr)$ is the probability of a term t in Cr, nr is the non-confidential cluster similar to Cr and $p(t/nr)$ is the probability of a term t in nr.

Using eq. (3) and eq. (4)

$$confidential\_score(t) = \frac{p_{cr}}{Max[p_{nr1}, p_{nr2}, p_{nr3},....., p_{nrn}]} \quad (11)$$

**Comparison of three methods:**

Considering Method I and II,

In Method I,

$$confidential\_score(t) = \frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}}$$

In Method II,

$$confidential\_score(t) = \frac{p_{cr}}{\frac{p_{nr1}+p_{nr2}+p_{nr3}+.....+p_{nrn}}{N}}$$

Since

$$\sum_{i=1}^{n} p_{nri} > \frac{p_{nr1}+p_{nr2}+p_{nr3}+.....+p_{nrn}}{N} \quad (12)$$

$$\frac{p}{\sum_{i=1}^{n} p_{nri}} < \frac{N * p_{cr}}{p_{nr1}+p_{nr2}+p_{nr3}+.....+p_{nrn}} \quad (13)$$

Considering Method I and III

In Method I,

$$confidential\_score(t) = \frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}}$$

In method III,

$$confidential\_score(t) = \frac{p_{cr}}{Max[p_{nr1}, p_{nr2}, p_{nr3},....., p_{nrn}]}$$

Let

$$Max[p_{nr1}, p_{nr2}, p_{nr3},.....p,_{nrn}] = p_j \quad (14)$$

Since

$$p_{nr1}+p_{nr2}+p_{nrj-1}+p_{nrj}+p_{nrj+1}+.....+p_{nrn} > p_j$$

$$\frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}} < \frac{p_{cr}}{p_j} \quad (15)$$

From the above observations(eq.(13) and eq. (15),When method 1 is compared with method II and method III , the method II and method III results in high confidential score hence results in more FPR than method I. So, for further research evaluation Method I is chosen as a proposed method for confidential score calculation.

*B. Testing Phase*

The main objective of the detection phase is to detect the confidentiality level of the test document in terms of a score [84]. In order to find the confidential score of the tested document, apply data pre-processing steps, transformed into TF vector format, find the similar confidential clusters and calculate the confidentiality value of the test document.

Testing phase consists of the following steps:

1. *Assign the test document to a relevant confidential cluster:* The purpose of this step is to identify which of the confidential clusters to be used to determine the confidentiality level of the tested document. This is done by using the cosine similarity measure (as discussed in section 3.3.2.3); all the confidential clusters whose similarity is greater than a predefined threshold are selected.

2. *For each of the similar cluster, identify all the confidential terms that appear both in the tested document and clusters.*

Scan the text of the test document and identify the terms that appeared both in test documents and similar clusters. Consider the terms of similar clusters, whose score is greater than 1 to be included in the document's confidentiality calculation.

3. *Calculate the document's confidentiality score*.

For each similar cluster, calculate the score by summing up the scores of all the confidential terms. If more than one similar clusters available then the confidentiality of the tested documents is calculated by a summation of scores of all the similar clusters.

Confidentiality value (similar cluster) = summation of scores of all confidential terms that appear both in the tested document and similar cluster.

Confidentiality value (tested document) = summation of all similar clusters score.

4. *Determine whether the document is confidential.*

If the confidentiality value of the tested document is above the threshold then that document marked as a confidential and it is blocked.

*C. Algorithms*

This section presents different algorithms used for the proposed method. Training phase and Detection phase algorithms are the two main algorithms in the proposed method.

*a) Training Phase Algorithm*

The training phase algorithm is to create confidential clusters with confidential terms by taking confidential, non-confidential documents as an input.

*Training phase algorithm*

*Input*

1. List of confidential documents
2. List of non-confidential documents

*Output*

CTR – List of confidential clusters, each cluster with confidential terms and their scores.

*Steps:*

1: CD ← Read set of confidential documents

2: ND ← Read set of non-confidential documents
3: T ← Read confidential score threshold value
4: Perform clustering on confidential documents with K-means unsupervised clustering algorithm and assign set of confidential clusters to CR

// CR is a list of confidential clusters

5: Perform clustering on non-confidential documents with K-means unsupervised clustering algorithm and assign set of confidential clusters to NR

//NR is a collection of K non-confidential clusters

6: For each cluster cr CR

Calculate probability of each term and write into $CR_{LM}$

End for

// $CR_{LM}$ is a term probability values of all terms in a cluster CR

7: For each cluster $n_r$ NR

Calculate probability of each term and write into $NR_{LM}$

End For

//$NR_{LM}$is a term probability values of all terms in a cluster $n_r$

8: For each cluster cr $CR_{LM}$

For each term t in cr

Calculate confidential score

If(score (t) >T)

Consider the term t as a confidential term and assign term t and its score to CTR

End if

// CT is a set of confidential terms and their scores of a cluster CR

End For

End for

*b) Confidential Score Calculation of a Term*

This algorithm is to identify the terms which are having high probability in confidential documents and low confidentiality in the non-confidential documents.

*Input:*

1. List of confidential clusters with probability values.
2. List of non-confidential clusters with probability values.

Score calculation threshold

*Output:*

CTR - Set of clusters,each with confidential terms and their scores.

*Steps:*

1:$CR_{LM}$← read set of confidential clusters with probability
2: $NR_{LM}$← read set of non-confidential clusters with probability
3: TR – initialize similarity threshold
4: For each cluster Cr in $CR_{LM}$

$NSR_{LM}$←Find Non-confidential clusters whose cosine similarity to CR is above TR

Term tis available in a more than one similar cluster, add probabilities and maintain one probability for a term.

For each term t in Cr

Calculate Confidential Score (t) //t available in C

// term not available in non-confidential clusters

Confidential Score (t) = 0.01

If score greater than 1

CTR [CR] ←CR[t]

CTR [CR]← score [t]

// CTR is a set of confidential cluster, each with

confidential terms and their scores

End if
End for
Return t
End for

*c) Testing Phase Algorithm*

The testing phase algorithm is to find the confidentiality of the test document in terms of score. If score greater than a predefined threshold then consider that test document is confidential otherwise non-confidential.

*Algorithm*

*Input:*

1. The test document whose confidentiality we wish to determine
2. Set of confidential term clusters
3 Confidentiality threshold to determine the confidentiality of the tested document
4.Similarity threshold, to detect the similar clusters

*Output:*

Conf-score – represents the confidentiality (yes/no) of the tested document

Steps:

1: Read test document D
2: Read confidential clusters, CTR
3: Read confidential score threshold, $T_C$
4: Read similarity threshold, Tr
5: InitializeConf_score = 0
6: For each cluster CR CTR
Cosvalue ← similarity between D and CR
If (cosvalue>Ts)
add CR to p as a similar cluster// p is a set of similar clusters
End if
End for
7: For each cluster CR$\in$p
Initialize score = 0 // score of a cluster CR
Find all terms that appear both in D and CR and assign to M
For each term t$\in M$
Score = score + score (t)
End for
Conf_Score= Conf_score+score
End for
8: If (Conf_score> $T_c$ )
Test document marked as a confidentinal
Else
Test document marked as a Non-confidential
End if

*D. Evaluation of TBCDM*

This section presents the evaluation of the Term Based Confidentiality Detection Method (TBCDM) using Reuters and Enron email dataset.

*A. Evaluation of TBCDM using Reuters News Article Dataset*

Experimental results of the proposed method using Reuters news article dataset are presented in this section[24].

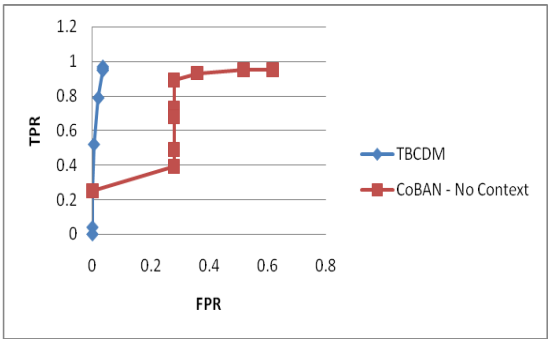


**Fig 1.Comparison of TPR and FPR between TBCDM and CoBAn-No context using Reuters News Article.**

From the above graph, the performance of the TBCDM is better than the performance of the baseline method CoBAn-No Context. Proposed method performed well for confidentiality threshold 0.05 and similarity threshold 100.

*B. Evaluation of TBCDM using Enron Email Dataset*

The following is the experimental results of the proposed method using Enron email dataset[25].


**Fig 2 Comparisons between TBCDM and CoBAn-No Context using Enron Email Dataset**

For 0.05 confidentiality threshold and 100 similarity threshold, the proposed method TBCDM and baseline method CoBAn-No Context performed well. The above graph shows, the performance of the TBCDM better than the baseline method for all threshold values.

## IV. Conclusion.

The present paper described a new confidentiality detection method called "TBCDM" for data leakage prevention. TBCDM detects the confidential content of the document and classifies test documents in to confidential and non-confidential. This paper presents the detailed methodology of the TBCDM .The performance of the proposed method has been measured with different similarity threshold values and different confidential score threshold values.TBCDM detects the confidentiality of the test documents with high true positive rate and low false positive rate. The performance of the proposed method is compared with the base line method (**CoBAn-No Context)**, by using Reuter's news article and Enron email data sets.

## REFERENCES


1. Michael Hart , Pratyusa Manadhata and Rob Johnson, "Text Classification for Data Loss Prevention", Lecturer notes in computer science ,11th International symposium on pririvacy enhancing technologies ,PETS,2011, PP 18-27.
2. Yuri Shapira, Bracha Shapira and Asaf Shabtai, "Content based data leakage detection usingextended fingerprinting, 2006.
3. Shabtai, A., Elovici, Y., Rokach, L," A Survey Of Data Leakage Detection And Prevention Solutions", Springer Briefs in Computer Science, Springer, 2012.
4. Cathey Rebacca, Ma L and Goharian N, "Misuse Detection for Information Retrieval System". Proceedings, 12th ACM conference on Information and Knowledge Management, 2003.
5. Ma L and Goharian N, "Query Length Impact On Misuse Detection In Information Retrieval Systems", proceedings, ACM symposium on Applied Computing, pp 1070-1075.
6. Mun H, Han K, Yeun C and Kim K, "Yet another Intrusion Detection System against Insider Attacks", symposium on Cryptography and Information security.
7. Karma, a Terzi, Evimaria and Bertino E,"Detecting Anomalous Access Patterns in Relational Databases", International Journal on Very Large Databases, 2008, 17(5), pp 1063-1077.
8. Sunu, M Michalis P, Hung N and ShambhuU,"A Data Centric Approach to Insider Attack Detection in Database Systems." Technical Report,2009.
9. Simon D Byers "Information Leakage Caused By Hidden Data In Published Documents" IEEE Security And Privacy, Volume 2, Issue 2,March 2004, Pp 23-27
10. Abbadi, I.M., and Alawneh, M," Preventing Insider Information Leakage For Enterprises", Proceedings, International Conference on Emerging Security Information, Systems and Technologies, pp 99–106, 2008.
11. De Capitani , Vimercati S, Foresti S, Jajodia S, Paraboschi S, and Samarati, P," Encryption Policies For Regulating Access To Outsourced Data", ACM Transactions on Database Systems, 35(2), 12:2–12:46, 2010.
12. Bowen B.M, Hershkop S, Keromytis A.D, and Stolfo S.J," Baiting Inside Attackers Using Decoy Documents," Proceedings, 5th International ICST Conference (SecureComm'09,), 2009.
13. Čenys, Rainys D., Radvilavičius L, and Gorani, N," Implementation of Honeytoken Module in DBMS Oracle 9ir2 Enterprise Edition for Internal Malicious Activity Detection", 2005.
14. Maya Bercovitch, Meir Renford, Lior Hasson Asaf "HoneyGen: An automated honeytokens generator", : 2011 IEEE International Conference on Intelligence and Security Informatics, ISI 2011, Beijing, China, 10-12 July, 2011
15. Cohen W.W, "Learning rules that classify e-mail", Proceedings, AAAI Symposium on Machine Learning in Information Access, pp 18–25, 1996.
16. Michael Hart , Pratyusa Manadhata and Rob Johnson, "Text Classification for Data Loss Prevention", Lecturer notes in computer science ,11th International symposium on pririvacy enhancing technologies ,PETS,2011, PP 18-27.
17. Zilberman, Polina, Shabtai, Asaf, Rokach Lior "Analyzing Group Communication for Preventing Data Leakage via Email", .IEEE, 2011.
18. Katz Gilad, Elovici Yuval, Shapira Bracha, "CoBAn: A Context Based Model for Data Leakage Prevention". Elsevier, Journal of Information Science, 262(2014), pp 107-128.
19. Caputo D.D, Stephens G.D, and Maloof M.A,"Detecting insider theft of trade secrets", IEEE Security and Privacy, 7(6), pp 14–21, 2009.
20. Salton G, Buckley C, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, Vol. 24, No. 5, pp 513-523, 1988.
21. A Anguera, Jon Barreiro and J. A Lara,"Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry", Computational and structural Biotechnological journal, Elsevier, Vol 14, pp 185-199, 2016.
22. Steinbach M, Karypis G, and Vipin Kumar," A Comparison of Document Clustering Techniques", A Comparison of Document Clustering Techniques. Technical Report #00-034.
23. Gregory Glen Odom and La Grand Prairie, "Method and Apparatus for Secured Transmission of Confidential Data Over an Unsecured Network", United States Patent, Oct. 30, 1997.
24. **https://archive.ics.uci.edu/ml/datasets/reuters -** 21578 + text + categorization + collection.
25. http://www.cs.cmu.edu/enron