Design and Implementation of Sensitive Information Security Model Based on Term Clustering

A Model to Secure High Dimensional and Sensitive Information Based Text Documents

Sona Kaushik
Birla Institute of Technology
Mesra, Ranchi, India

Shalini Puri
Birla Institute of Technology
Mesra, Ranchi, India

Pankaj Gupta
Birla Institute of Technology
Mesra, Ranchi, India

ABSTRACT

Exchange of enormous data and information securely and frequently via Internet is very common and demanded in today's fast track scenario of world. The idea behind the proposed Sensitive Information Security Model Based on Term Clustering (SIS-TC) is to provide the security to a large volume of text documents which contain very important and sensitive information or data or both. These documents are first broken into its constituent parts, called terms, by using knowledge repository and then term clusters are made by finding out the similar terms of each category. These clusters represent the categories of Noun, Pronoun, Numeral, Punctuation etc. Only one instance of a cluster is kept and become the cluster representative. Firstly, the term frequency of each different occurred term (or word) is calculated and then all the duplicate copies of each term are removed, so that to transform it into the low dimensional data. Such reduced data set drastically decreases the total size of the complete data and space as well, and increases the performance of the system by the ratio of 65% -70%. Next, this reduced data is divided into High Risk Data (HRD) and Low Risk Data (LRD) to provide different level of security to each type. Therefore, HRD is symmetrically encrypted whereas LRD is encrypted non-symmetrically. This paper also includes the analytical experimental results based on the test data set of 8 text documents of varying sizes.

General Terms

Term clustering; knowledge repository; sensitive information; high dimensional and low dimensional data; symmetric and non-symmetric encryption.

1. INTRODUCTION

Do we really think that our high dimensional, sensitive data and information are completely secured against all the attacks while sent over on the unsecured network like Internet? Will this information still maintain 100% confidentiality and integrity at the reception? Can we be sure completely? These are some question marks which always give a birth to unreliability when we send very bulky and delicate information via an unsecured network.

As Internet has become a part of our daily lives; its harms are as high as its uses. When some data is sent over on the Internet, it must be made secured as Internet has proven itself as the most unsecured network. Sensitive information and data is always the core and important part for any enterprise, organization or an individual working around us, they require and prefer to send their data via secured networks. Such information requires a good level of security while is sent over an unsecured communication channel.

Not only the Internet provides unsecure and unreliable transmission, data can be insecure at the user interface level and in the data storage also. Therefore, the security provision is required at the three levels – at the user level, data storage and the communication channel.

Artificial Intelligence (AI) provides many methods, techniques and the base of enhanced and efficient algorithms [1] [2] in the field of text mining, intelligent systems, robotics etc. Natural Language Processing (NLP) via dictionary knowledge repository is one of the heart breaking tools of the AI which provides a strong forum in processing the language based sentences. Using NLP, a series of sequential steps on a sentence is performed to find out the constituent and small parts, called Words or Terms, based on the syntactic structure of the sentence. A sentence always follows a well – defined structure of the language predefined and uses the grammatical rules to identify all the words individually and independently.

Text mining is one of the so called related task domain of AI [3] [4]. It is used in so many fields, like Email Filtering, Text Categorization, Clustering [4], Text Classification, Spam Filtering, document dimensionality reduction and so many related fields and concerned areas. The use and combination of text mining and AI can help to identify the terms at the sentence, document and corpus levels [4] and produce the very high dimensional data. This data can further be made low dimensional to drastically improve the system performance by around 65% to 70%. If such low dimensional data is provided very high security, no doubt with the increased complexity of the system, the system is able to send the high dimensional and delicate information over on the unsecured communication channel, provided that information is first reduced and transformed into the low dimension and then made tightly secured with the hard security system.

Therefore, the idea behind the proposed model is to provide the hard level security to the text documents which are very bulky and consist of very much space dimension. Additionally, these documents contain very sensitive information. Such documents are sent over on either secure or unsecure communication channel, while maintaining good reliability, confidentiality and integrity of the data, by first reducing the dimension of the data greatly and then encrypting symmetrically and non-symmetrically the high risk and low risk data correspondingly. Such compressed and encrypted bits are sent over.

Section II discusses the background and related work about the security of sensitive information, term clustering, and a discussion of confidentiality and integrity. The detailed design of the proposed model is discussed further in section III. Section IV discusses the SIS-TC implementation in detail. In section V, the analytical experimental results are described based on the data sets of 8 different text documents. Finally, section VI concludes the paper and suggests future work.

2. INTRODUCTION

Do we really think that our high dimensional, sensitive data and information are completely secured against all the attacks while sent over on the unsecured network like Internet? Will this information still maintain 100% confidentiality and integrity at the reception? Can we be sure completely? These are some question marks which always give a birth to unreliability when we send very bulky and delicate information via an unsecured network.

As Internet has become a part of our daily lives; its harms are as high as its uses. When some data is sent over on the Internet, it must be made secured as Internet has proven itself as the most unsecured network. Sensitive information and data is always the core and important part for any enterprise, organization or an individual working around us, they require and prefer to send their data via secured networks. Such information requires a good level of security while is sent over an unsecured communication channel.

Not only the Internet provides unsecure and unreliable transmission, data can be insecure at the user interface level and in the data storage also. Therefore, the security provision is required at the three levels – at the user level, data storage and the communication channel.

Artificial Intelligence (AI) provides many methods, techniques and the base of enhanced and efficient algorithms [1] [2] in the field of text mining, intelligent systems, robotics etc. Natural Language Processing (NLP) via dictionary knowledge repository is one of the heart breaking tools of the AI which provides a strong forum in processing the language based sentences. Using NLP, a series of sequential steps on a sentence is performed to find out the constituent and small parts, called Words or Terms, based on the syntactic structure of the sentence. A sentence always follows a well – defined structure of the language predefined and uses the grammatical rules to identify all the words individually and independently.

Text mining is one of the so called related task domain of AI [3] [4]. It is used in so many fields, like Email Filtering, Text Categorization, Clustering [4], Text Classification, Spam Filtering, document dimensionality reduction and so many related fields and concerned areas. The use and combination of text mining and AI can help to identify the terms at the sentence, document and corpus levels [4] and produce the very high dimensional data. This data can further be made low dimensional to drastically improve the system performance by around 65% to 70%. If such low dimensional data is provided very high security, no doubt with the increased complexity of the system, the system is able to send the high dimensional and delicate information over on the unsecured communication channel, provided that information is first reduced and transformed into the low dimension and then made tightly secured with the hard security system.

Therefore, the idea behind the proposed model is to provide the hard level security to the text documents which are very bulky and consist of very much space dimension. Additionally, these documents contain very sensitive information. Such documents are sent over on either secure or unsecure communication channel, while maintaining good reliability, confidentiality and integrity of the data, by first reducing the dimension of the data greatly and then encrypting symmetrically and non-symmetrically the high risk and low risk data correspondingly. Such compressed and encrypted bits are sent over.

Section II discusses the background and related work about the security of sensitive information, term clustering, and a discussion of confidentiality and integrity. The detailed design of the proposed model is discussed further in section III. Section IV discusses the SIS-TC implementation in detail. In section V, the analytical experimental results are described based on the data sets of 8 different text documents. Finally, section VI concludes the paper and suggests future work.

3. BACKGROUND AND RELATED WORK

According to Philip Zimmermann, “If privacy is outlawed, only outlaws will have privacy”, very truly said by him. Keeping it in our minds, we always try to have best security provision to send huge data over on the unsecured communication channel. In this direction, a lot of research work has been done and still being done to fight against the different attacks, intrusions, loop holes, etc. The effort here within is to provide hard level security of the huge sensitive information based text documents. The next subsections discuss the related research work done on the concert.

3.1 Securing Sensitive Information

To secure sensitive information, many sensitive information models [5] - [12] have been developed. The security provided in different research models is at the user interface level, at the communication channel or at the data storage [5]. Some models provide security to sensitive information at the user interface level. For that, either they develop a security mechanism for the information usage at memory level [6] or the operating system [8] – [10] on the client side. Secondly, some information models have been designed to provide the security mechanism upon the unsecured communication channel [12]. Such channel is much used by the normal user. Thirdly, the security provisions are provided for the data stored at the enterprise, and organization’s web server. Lastly, the research work has also been performed to provide security mechanism to the data storage. All the three cases consider the sensitive data or information which if hit or hurt once, can result in total harm.

3.1.1 Sensitive Information

The sensitive information contains very important and delicate data. This data varies among one organization’s need to other organization’s need. As such, there are many applications and areas which contain sensitive information;

they are - Military and Navy Based Information Systems, Terrorism Based information, Crime Branch Investigation, Satellite Related Concerns, Astronautical Based Information Systems, Nuclear Power Plant, Financial Accounts, Bank Transactions and Management, Health Care Systems, and many more. These areas and their information systems always contain delicate information and data which when sent over on the unsecured network require the provision of hard level security. For an example, in financial accounts, the precision value up to 6th decimal place in floating point data value is very sensitive and need to be kept secured. These are some of the cases showing how much the sensitive information can be and of which level?

3.1.2 How to Provide Security to Sensitive Information?

Sensitive information requires the best level of security. In this direction, many models [5] – [12] and algorithms have been developed. As such, an information system can have three security levels- Hard Level Security (HLS), Soft Level Security (SLS) and Mixed Level Security (MLS). HLS based systems are the tight security based information systems that can never be penetrated, SLS based information security mechanism cannot be as rigid as HLS, and lastly MLS provides the mixed security features based upon the HLS and SLS as well.

Therefore, sensitive information based security models must have the features of HLS to reliably secure the information system without any penetration or loophole.

3.2 A Game of Large and Small Sized Data

The heavy volume of text documents can be made compressed in the form of low dimensional data sets. To achieve such objective, clustering provides the most suitable and appropriate solution of term reduction.

3.2.1 High Dimensional Information

Many text documents contain bulky information, called the high dimensional information [13]. In text mining, these documents are processed to be categorized into the different categories based either on the supervised learning or unsupervised learning paradigm [4]. For the security provision, this large volume of textual information is preprocessed at the sentence, document and the corpus levels using NLP tools, so that the basic constituent parts; terms or words can be extracted. This huge and large set of data contains all nouns, pronouns, adverbs, adjectives, prepositions, delimiters, punctuations, and spaces [14].

3.2.2 Term Clustering

Term (or word) clustering refers to make the clusters or groups of the similar terms occurred in a category set. When clustering is used in the system, it provides the feature of term reduction so that to drastically making it very low – dimensional. This transformed (or compressed) form of data is effectively and efficiently managed.

3.3 Confidentiality and Integrity

Data confidentiality and integrity are the primary issues of the security. When large volume of sensitive information is sent over the non-secured channel, it is required that the information must be protected completely, kept confidential from others [15] and maintaining the integrity of the data. The information sent by the client must be received exactly by the receiver as it was sent.

Therefore, we have made an effort to provide hard level security to the large set of text documents having sensitive information, so that the data is sent and received reliably, while maintaining its confidentiality and integrity of best level. The model is very complex as it considers many parameters related to the basic details of NLP, Text Mining and Information System Security.

4. DESIGN OF SENSITIVE INFORMATION MODEL BASED ON TERM CLUSTERING

The proposed detailed design is the extension of the work [14].In this section the design of the complete system and its required details are discussed to explain each component minutely.

4.1 The Framework Of Proposed Methodology

Our proposed model, i.e., The Sensitive Information Security (SIS) Model consists of two layers; Scrambling Layer and Transformation Layer to convert the text documents into data chunks and then to provide them high security to send over on the insecure network. Each layer plays its own important role in securing the data, so the proposed model works efficiently and effectively with high performance and accuracy.

Scrambling Layer processes the text documents and converts them into the form various clusters and then realizes the data into the High Risk Scrambled Data and Low Risk Scrambled Data. This layer transforms the information into data. This used the concepts of Natural language Processing.

Transformation Layer encodes the content of each Refined Term Cluster into a non-readable format. This is basically encryption, but before encrypting the data, two things are considered. Firstly, the contents are not directly encrypted, but are changed to the byte array. Secondly, the encryption scheme applied is symmetric or non-symmetric; is based on the criticality of the content. For high risk data, which contains the most important information, like nouns or verbs, is encrypted using non-symmetric encryption. The rest of the data is encrypted using symmetric single key schemes for encryption.

In the model extended in this paper makes its distinction from its original in number of ways. Firstly, the use of Natural Language is replaced by dictionary’s knowledge repository. And the most important of all is that in this extension model of SIS, we are sending over the data by first converting the data from high dimension to low dimensional data. This improves the efficiency of the system drastically.

4.2 Detailed Design

The generalized design view of the design of SIS-TC is shown in figure 1. The two layers involved; scrambling and transformation, are further detailed in various block.

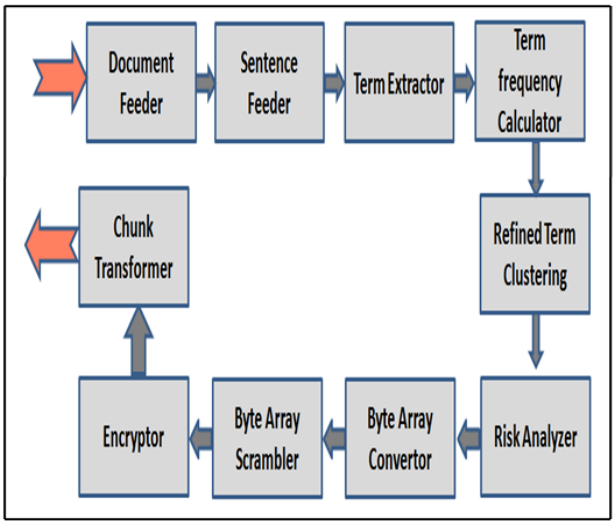

Fig 1: Detailed Design of SIS-TC

Group of various documents are fed to the document feeder which feeds further the information one document at a time. Term Extractor categorizes the data into various predefined term clusters. This involves the use of dictionary API attached to the block. Further is the term frequency calculator. It calculates the frequency of each term incorporated in the clusters and thus helps in the conversion of high dimensional data into low dimensional data. Next risk analyzer analyses and bifurcate the data into High risk scramble data and low risk scramble data. The idea is to give high security to high risk data and comparative low security to low risk data, thus improving the efficiency and decreasing the complexity of the system. This bifurcated data is provided security in transformation layer by converting the data into byte arrays and then scrambling the arrays. Further it is encrypted using standard encryption algorithms.

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

5. SYSTEM IMPLEMENTATION

Addition of new feature of converting the data into low dimensional sets makes the model strengthener. The implementation approach is discussed further and shown in figure 2. Initially, text documents are fed to the Document Feeder. Its functionality is to pick one file at one time and feeds the sentence feeder as an input of single file.

An input reader reads the file content and using the token of dot operator extracts the sentences out of the file. Further it analyzes the sentence and extracts the words out of it. Each word acts as an input to search it in the dictionary repository. If the word is found in the repository, its type is stored in variable type and appropriately inserted into the respective Cluster Set. Any word not found in the repository is treated as ‘noun’. The category sets used are noun, verb, proper noun, conjunction, preposition, number, preposition and determiner.

On completion of the process discussed yet, for each cluster the frequency of each term in the cluster is examined. If the count of similar term in a cluster exceeds the threshold value, it is bifurcated in two and named as cluster01 and cluster02. Within each cluster if the count of a single term is exceeding the threshold value, it is further allocated a separate cluster naming them by term name; say, verb_is or verb_the cluster.

All the clusters related to nouns or verbs are tagged as high risk items and rest are tagged as low risk items. Then all these clusters are converted to byte arrays and scrambled accordingly using some predefined scrambling logic for bytes. Now these scrambled byte arrays, are if belongs to the high risk items, then encrypted using Public key encryption schemes and rest are encrypted using non-symmetric information to reduce the system’s complexity with compromising the security.

Finally, the encrypted data for each cluster is splitten to various data chunks from the chunk transformer black box. These data chunks are then sent over to the destination. At destination, the reverse process is followed. The chunks of data can be reassembled by using the addresses of the sentence and words within the sentence. These addresses were appended at the sender’s end ahead of each word in the format [sentence number, position]. For the documents the address is appended along with the encrypted chunk in the format [document_id].

6. ANALYTICAL EXPERIMENTAL RESULTS

The SIS-TC model has been analyzed by using the test data sets of 8 IEEE Research Papers; i.e. text documents. Such analysis is done to check the efficiency and effectiveness of the system in terms of the performance issue. These data sets are experimented and evaluated for dimensionality reduction and creating a comparative graph among different 8 terms (clustered) appeared in each of the text document (represented by means of Number of Pages) along with the associated term frequency.

6.1 Description of Test Sample Data

Table 1 shows test data set of 8 IEEE text Documents. The data set shows the total number of occurrences of each term of 8 different Term Clusters (TC) in each of the 8 Text Documents (TD). A TD is represented in terms of the number of pages of that text document. For an example, the size of TC1 in TD1 is 34.

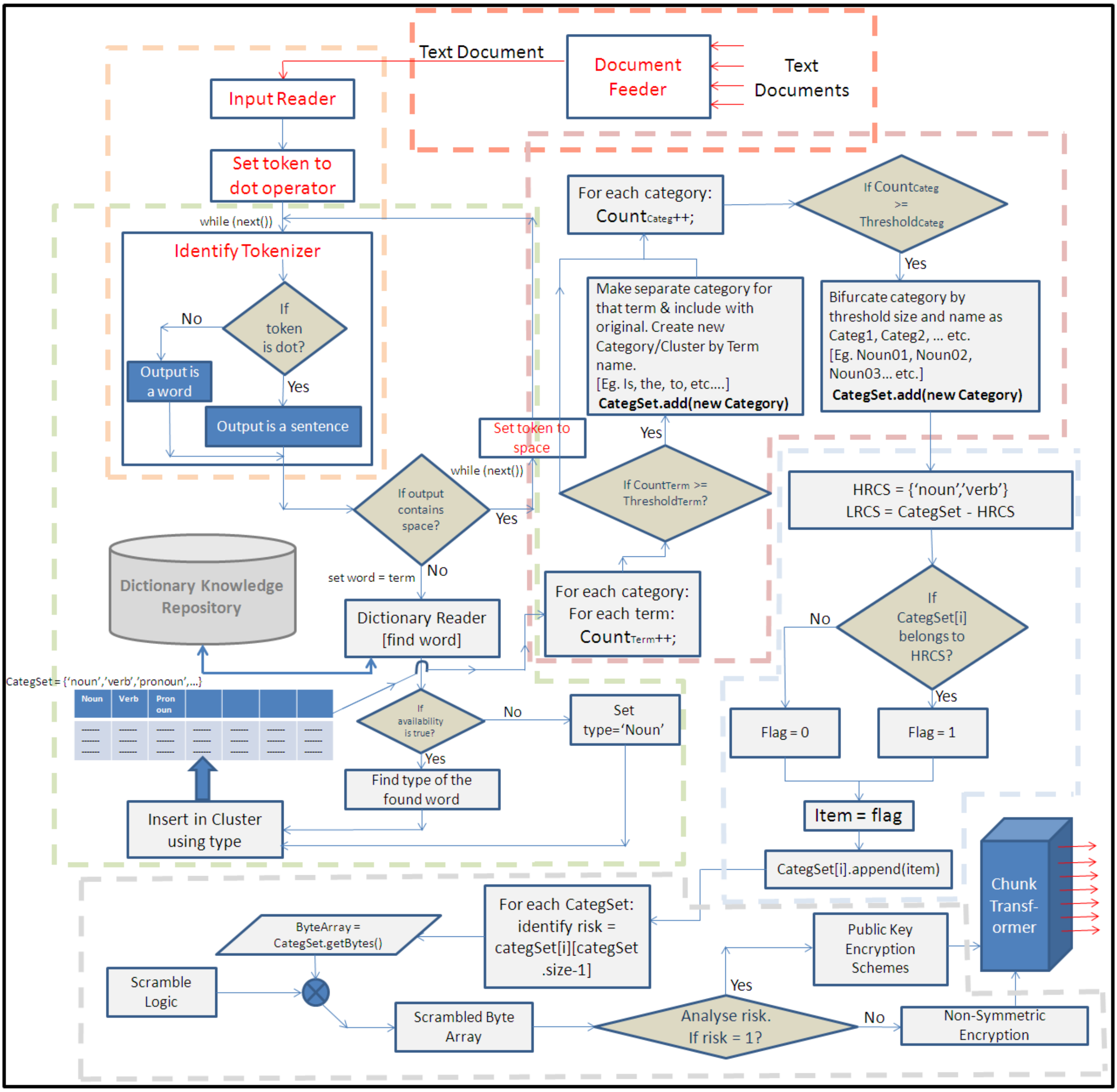
Text Document
Document Feeder
Text Documents
Input Reader
Set token to dot operator
while (next())
Identify Tokenizer
If token is dot?
No
Yes
Output is a word
Output is a sentence
Set token to space
while (next())
If output contains space?
Yes
No
set word = term
Dictionary Knowledge Repository
Dictionary Reader [find word]
CategSet = {'noun','verb','pronoun',...}
Noun
Verb
Pron oun
If availability is true?
No
Yes
Set type='Noun'
Find type of the found word
Insert in Cluster using type
For each category: Count_Categ++;
If Count_Categ >= Threshold_Categ
Yes
Make separate category for that term & include with original. Create new Category/Cluster by Term name. [Eg. Is, the, to, etc....] CategSet.add(new Category)
Bifurcate category by threshold size and name as Categ1, Categ2, ... etc. [Eg. Noun01, Noun02, Noun03... etc.] CategSet.add(new Category)
Yes
If Count_Term >= Threshold_Term?
For each category: For each term: Count_Term++;
HRCS = {'noun','verb'} LRCS = CategSet - HRCS
If CategSet[i] belongs to HRCS?
No
Yes
Flag = 0
Flag = 1
Item = flag
CategSet[i].append(item)
Chunk Transf-ormer
For each CategSet: identify risk = categSet[i][categSet .size-1]
ByteArray = CategSet.getBytes()
Scramble Logic
Scrambled Byte Array
Analyse risk. If risk = 1?
Yes
No
Public Key Encryption Schemes
Non-Symmetric Encryption

Fig 2: Implementation of SIS-TC

Table 1. Test Data Set Of 8 Text Documents

	Text Docs (No. of Pages) →	TD1	TD2	TD3	TD4	TD5	TD6	TD7	TD8
↓ Term Clusters		4	5	6	7	8	10	14	28
TC1	is	34	49	54	51	40	26	101	101
TC2	the	101	101	101	101	101	101	101	101
TC3	a	24	45	29	34	47	49	101	101
TC4	to	25	31	35	51	67	46	86	101
TC5	of	50	101	101	101	101	63	101	101
TC6	for	17	34	25	26	35	48	79	93
TC7	in	29	68	59	18	55	42	101	101
TC8	,	61	101	97	101	101	101	101	101

6.2 An Analysis of Sample Test Data Sets

Figure 3 shows a comparative chart among 8 text documents by the means of different term clusters and their term frequencies. This chart is based upon the test data set of the table 1. Therefore, the chart shows that how many times each term occurs in a particular TD.

7. CONCLUSION AND FUTURE SCOPE

This paper discussed the detailed design of the Sensitive Information Security Model Based on Term Clustering (SIS-TC) which provides a strong framework to send reliably a set of text documents having sensitive and delicate information over on the secure or unsecured network. Although this system seems to be very complex and time consuming, yet it provides the hard level security to the data.

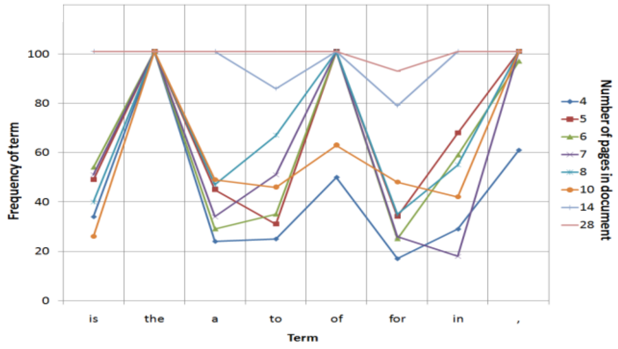

Fig 3: A Comparative Chart Showing Different Term Clusters, Term Frequency of 8 Text Documents

The transformation of the high dimensional data into the low dimensional increases the system performance greatly. Different security mechanisms – symmetric and non-symmetric provided to HRD and LRD help to focus more on the data which must be kept completely confidential by using the hard leveled security architecture. As this system maintains the data integrity at the best level, so, at the receiver side, the exact text documents are received as it was sent by the sender, thereby maintaining the confidentiality and integrity of this reliable system.

The analytical experiments performed on the sample test data set of the 8 text documents show that the system implementation drastically reduces the dimension of the data with the ratio around 6:1 for the corresponding term clusters respectively and finally the data security is provided.

Therefore, this model increases the overall performance of the system in terms of the space and also produces accurate results with high efficiency and effectiveness.

This model can further be extended by including one more layer of clustering of data. This clustering will make the different groups of the similar term clusters to reduce the data dimensions more. On the other side of the future scope of the paper, as this model is more concerned about the textual based information security provision only, so this proposed work will be extended for the inclusion of image information based documents.

8. REFERENCES

[1] Eliane Rich, Kevin Knight and Shivashankar B Nair, Artificial Intelligence, 3rd ed., Mc Graw Hill, 2010.

[2] N. P. Padhy, Artificial Intelligence and Intelligent Systems, 5th ed., Oxford University Press, 2009.

[3] Jiawei Han, and Miche Line Kamber, Data Mining: Concepts and Techniques, 2nd ed., Elsevier, 2006.

[4] Shady Shahata, Fakhri Karray, and Mahamed Kamel, “Enhancing Text Clustering using Concept-based Mining Model,” IEEE Proc. of the Sixth International Conference on Data Mining, 2006.

[5] Xianping Wu, Phu Dung Le and Balasubramaniam Srinivasan, “Security Architeture for Sensitive Information Systems,” Convergence and Hybrid Information Technologies, pp. 239–266, March 2010.

[6] Weidong Shi, Joshua B. Fryman, Guofei Gu, Hsien-Hsin S. Lee, Youtao Zhang, and Jun Yang, “InfoShield: A Security Architecture for Protecting Information Usage in Memory,” IEEE Twelfth International Symposium on High-Performance Computer Architecture, 2006.

[7] Shi-Hua Wang, and Xiao-Yongli, “A Security Model To Protect Sensitive Information Flows Based On Trusted Computing Technologies,” IEEE Proc. of the Seventh International Conference on Machine Learning and Cybernetics, July 2008.

[8] Zhao Yong, Liu Ji Qiang, Han Zhen, and Shen ChangXiang, “An Operating System Trusted Security Model For Important Sensitive Information System,” IEEE First International Symposium on Data, Privacy and E-Commerce, 2007.

[9] Khaled Hussain, Sharon Rajan, Naveen Addulla, and Ghada Moussa, “No-capture Hardware Feature for Securing Sensitive Information,” IEEE Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), 2005.

[10] John Black, Martin Cochran, and Ryan Gardner, “Lessons Learned: A Security Analysis of the Internet Chess Club,” IEEE Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC),2005.

[11] Kun Wang, Ruidan Su, Zengxin Li, Zhen Cai, and Li Hua Zhou, “Study of Secure Complicated Information System Architecture Model,” IEEE Proceedings of the First International Conference on Semantics, Knowledge, and Grid (SKG), 2006.

[12] J. L. Mejia-Nogales, S. Vidal-Beltran, and Y. J. L. Lopez-Bonilla, “Design and Implementation of a Secure Access System to InformationResources for IEEE 802.11 Wireless Networks,” IEEE Proceedings of Conference of the Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2006.

[13] Benjamin C. M. Fung, Thomas Trojer, Patrick C. K. Hung, Li Xiong, Khalil Al-Hussaeni, and Rachida Dssouli, “Service-Oriented Architecture for High-Dimensional Private Data Mashup,” IEEE Transactions On Services Computing (Under Publication), 2011.

[14] Sona Kaushik, and Shalini Puri, “Sensitive Information on Move,” International Journal of Scientific and Engineering Research, vol. 2, December 2011.

[15] Shihab A. Hameed, Habib Yuchoh, and Wajdi F. Al-khateeb “A Model For Ensuring Data Confidentiality,” IEEE Fourth International Conference on Mechatronics (ICOM), May 2011.