ISSN No:-2456-2165

A Developed Secured Model for Searching Into Secured Encrypted Data

Mohamed Abdelhadi
Information Systems Department
Faculty of IT, University of Tripoli
Tripoli-Libya

Tiruveedula Gopi Krishna
Computer Science & Engg Department
Adama Science & Technology University
Adama-Ethiopia

Abstract:- Our research paper has presented a new developed model for secured search into a big data of document collections. The developed model has investigated the importance of secured search and also the need for its practices in the real world. We have actually, studied both side of encryption in practical techniques issues and theoretical issues to improve the integration of information retrieval and cryptographic methods used for secured searching. We have chosen 3DES encryption technique to encrypt document data collections. Our new developed secured model has provided an efficient secured searching and/or security and authenticity over all.

Keywords:- *Information Retrieval, 3DES Encryption Method.*

I. AN INTRODUCTION

We have in our research paper presented a new developed secured model for searching into data collection documents which has to be encrypted. The new developed model must work as a prototype to deal with the problem of the authenticity and security. The developed secured model consist of an extended secured layer which will work as checkpoint-layer between the user and the server. The user should build own his\her check point to ensure that the encrypted data during the sending process and at the level of securing are well done to have confidentiality and security. We have developed an extended model which can integrate the security and authenticity of the developed model as such. The work has to be divided into some components as illustrated in the design phase. Each component consist of some processes and these processes has specific tasks. This research paper was so organized as to section one includes the problem definition, section two includes model design ,and section three includes the preparation component ,section four includes inverted file building module, and section five includes pseudo code based on scenarios ,section six include the implementation of the encryption program, finally section seven includes the results and conclusion. [1].

II. RELATED WORKS

Most literature in the research area of information retrieval are available in digital libraries or via internet at all, but in the recent researches, there are not so much deals within secured searching into data document collections; [1]-Maxim Martynov, Boris Novikov, they have proposed an algorithm for query evaluation in text retrieval systems based on well-known inverted lists augmented with additional data structure and estimate expected performance gains.[2]-R.Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, they have studied mainly what so called “searchable symmetric encryption (SSE)”;while its maintaining the ability to selectively search over it.[3]-A. Swaminathan, Y. Mao, G.-M. Su, H. Gou, A. Varna, S. He, M. Wu, and D. Oard, they have introduced a new framework for confidentiality preserving rank-ordered search and retrieval over large document collections.[4]- S. E. Robertson and K. S. Jones, have studied some simple approaches to improve text retrieval.[5]- R, Brinkman, J. M. Doumen, and W. Jonker.[6]- B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan.[7]- D. Song, D. Wagner, and A.Perrig.[8]- D. Song, D. Wagner, and A. Perrig, they have all concerned in their researches about practical techniques issues for searching into encrypted data issues.[9]- .Boneh,G.Crescenzo,R.Ostrovsky,G.Persiano.[10]- E-J.Goh, they have introduced in their research works, encryption with keyword search.[11]- B. Klimt and Y. Yang,[12]- R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, they were concerned in their research works, to deal with issues of order preserving encryption for numeric data.

III. PROBLEM DEFINITION

We are concern about two important issues as a matter of fact, the first issue is the encrypted data and the second issue is data storing then how to search on stored encrypted data securely. Data can be any kind of documents, or maybe emails as such which will be under our consideration in this work. We have first applied some important definitions which has some advantages for supporting our proposed model as much as to ease understanding the implications, which can effect on the security of searching into encrypted data [3,4].

IV. OUR SIMPLIFIED SCENARIO

In the first scenario we suppose that Marta has not to trust anyone (either Alice or server), she should has to care about the security and confidentiality by herself, which means; she has to deal directly within her stored encrypted data which stored into the (server) and just her knows how to encrypt and decrypt her data. Out of this obstacles, she may does not need any more the third-party, while she has the whole authority (Marta authority) The second scenario has suggested how does Marta search into her encrypted data securely? This is not so problematic since there is already such a well-proven technique which has been used in previous scenario [3].

V. MODEL DESIGN SCHEME

In our design scheme we have showed how we designed such a prototype model which consist of some components of predefined structured-object-oriented. We have first defined our model; then explained how should the proposed model functionally looks like .We have define the model as it consist of some processes which are logically related to each other in such away as structured-objected-oriented in a system or model.[7,8]. The proposed model consists of six main components such as:

1- Preparation component (predefined module).
2- Encryption (predefined module).
3- Decryption component (predefined module).
4- Searching component (predefined module).
5- Checkpoint component (predefined module).
6- Tracing component (predefined module).

Fig. 1:- Block diagram of the new developed Secured model

VI. DESIGN OF PREPARATION MODULE

We have started with designing the first component which is defining the preparation process. This process should have characteristics of the document preparation for storing the data as text or what is so called metadata. As it has already been decided using the information retrieval techniques for implementing the data structure algorithm which will be used for storing and sorting, merging, and finally searching the data into the targeted documents available to the IR systems [1].

VII. BUILDING INDEXED FILES MODULE

1- Set up a stop words filter. 2- Parse the text into a list of words and store them as keywords in the index file. Set up a postings file to record the relationship between indexed files and document itself.

A. ***First step:*** we have done stopwords removal; stemming all words before organizing the indexed file where it should be saved. First example will shows how we made all steps that explained in the document preparation process.

B. **Second step:** is to use a posting file to organize the inverted file as illustrated down in the table1,and table2 below explained by example two: suppose we have these three different original documents to preprocess them by using inverted file technique ,which are represent 1-The use of computers in library and information science education. 2-Education of librarians in faculties of education, library and information science, and communications compared. 3-Computers in science: a case study of computers and computer use by scientists. We use the same documents in example one to deal within the data stored into the inverted file by using a posting file to create pointer file as to ease the searching into the documents stored.

Step #1 Extract Terms		Step #2 Sort Terms		Step #3a Remove stop-words stemming		Step #3b Remove Duplicates (Pointers to Postings List)	
Term	Doc #	Term	Doc #	Term	Doc #	Term	Ptr
use	1	case	3	case	3	case	1
computers	1	communication	2	communicate	2	communicate	2
library	1	compared	2	compare	2	Compare	3
information	1	computers	3	comput	3	compute	4
science	1	computers	1	comput	1;3		5
education	1	computers	3	educate	1; 2	educate	6
education	2	education	2	faculty	2	faculty	7
librarian	2	education	1	info	1;2	inform	8
faculties	2	education	2	library	2	library	9
education	2	faculties	2	library	1:2		10
library	2	information	1	science	1;2; 3	science	11
information	2	information	2	science	3		12
science	2	librarians	2	study	3	study	13
communications	2	library	2	use	1;3	use	14
compared	3	library	1				
computers	3	science	3				
science	3	science	2				
case	3	science	1				
study	3	scientists	3				
computers	3	study	3				
use	3	use	1				

Table 1:- Documents Preprocessing Building (Inverted File)

Inverted File		Postings File#1		Postings File#2	
Term	Pointer	Pointer	(Doc#, WDFreq)	Pointer	(Doc#, WDFreq, Pos)
case	1	1	(3,1)	1	(3,1,5)
communications	2	2	(2,1)	2	(2,1,13)
compared	3	3	(2,1)	3	(2,1,14)
computer	4	4	(3,1)	4	(3,1,10)
computers	5	5	(1,1)(3,2)	5	(1,1,4)(3,2,1,8)
education	6	6	(1,2)(2,2)	6	(1,2,10)(2,2,1,7)
faculties	7	7	(2,1)	7	(2,1,5)
information	8	8	(1,1)(2,1)	8	(1,1,8)(2,1,10)
librarians	9	9	(2,1)	9	(2,1,3)
library	10	10	(1,1)(2,1)	10	(1,1,6)(2,1,8)
science	11	11	(1,1)(2,1)(3,1)	11	(1,1,9)(2,1,11)(3,1,3)
scientists	12	12	(3,1)	12	(3,1,13)
study	13	13	(3,1)	13	(3,1,6)
use	14	14	(1,1)(3,1)	14	(1,1,2)(3,1,11)

Table 2:- Posting List File As Output List Of Words

VIII. PSEUDO CODE

Search and retrieval must be initiated by the content owner, Marta; which can be done in the following steps:

- When searching for a particular word w in the data collections of Marta; first of all, she has to perform stemming and stop words removal to obtain the stemmed word w(s). The word-key is then derived from the master key and used to encrypt the stemmed-word wS to obtain (w(e)s).
- Hash value of (w(e)s) should be calculated and sent to server. Using the received hash value (k=h(w(e) s)), the server searches the protected term frequency table TF(e) C to identifies the (row) corresponding to the query word w.
- At last, she get conceal the query content from the server. After the server identifies the targeted(row TF(e)C(k,)) from the encrypted term frequency table (TF(e) C), that particular (row TF(e) C (k, .)) is sent back to Marta who then decrypts and decodes to obtain the plaintext term frequencies {TF(k, j), ∀j} and search about the particular words.[3].

IX. IMPLEMENTATION

As well as, we have determined the problem definitions and the scenarios which would be used to implement the secured model; we have chosen a programming language to write the code of testing prototype. As a matter of fact, due to the need of such very high quality package of software, we have decided to use Microsoft Visual Studio.Net 2008 which include many different programming tools. We have used first C# programming to create the inverted files which will consist of two files, indexed file and posting file as well. These two files has an a specific tasks , for example , the inverted file will have all the properties of whole stemmed words which has been prepared during the first documents preprocessing as shown in the section IV-A and IV-B. The second process is to encrypt and compress the inverted table which has been already prepared from the user Marta). When the user (Marta) has the process of encryption\compression finished, she must setup a password for sending the whole (C (E (inverted table)) to the server to store it as data base object. The third process is to setup the server to interact with the virtual server at the local IP address=127.0.0.0.1, and also setup the ports for sending\receiving the following [port 3000] for User Accounts, [port 3001] for Query search. The final process is when the user (Marta) wants to retrieve any encrypted document to search about some words, so she must send a query in an encrypted\compressed message to the Server, the server will receive the message then (decompress\decrypt) it, and retrieve back the data base object to the client (Marta) as is it. The client should then decompress\decrypt the inverted table, then search over the stemmed words into posted file.

X. RESULTS

As usual in the most practical issues we have programed such a prototype model to test how the model secure and reliable is. We have chosen some test documents as sample, we have selected 4 different document sizes, such as (10 kb, 20 Kb, 100Kb, and 500 Kb). The preparing time of documents for initialize the inverted file was somehow proportional to documents sizes that means the largest document size the largest time it took. For 10 Kb, it takes about 2 sec, for 100Kb is about 40 min. There are also some facts that must be mentioned, i.e., the document sizes has been reduced twice, once at the stemming which remove the duplicate words from the documents and also at (encryption/compression process). For our tests we found that the search time for any words, it depends on the document size, that means if there are few occurrences of terms then the search will be longer than if there are many terms occurrences. In the sample test program we investigate the criteria of collision and search time; we found in all stemmed inverted documents no collision, since there are no one stemmed word has the same pointer shared to any other. The sample test results has showed the following outputs:

QueryDate	SearchTime	DocumentSize	Collision
1/28/2008 10:1...	00:00:00.7343750	8926	0
1/28/2008 10:1...	00:00:02.8437500	8926	0
1/28/2008 10:2...	00:00:00.7500000	10378	0
1/29/2008 7:38 PM	00:00:16.5625000	143311	0

Table 3:- The output of some queries processed in testing phase.

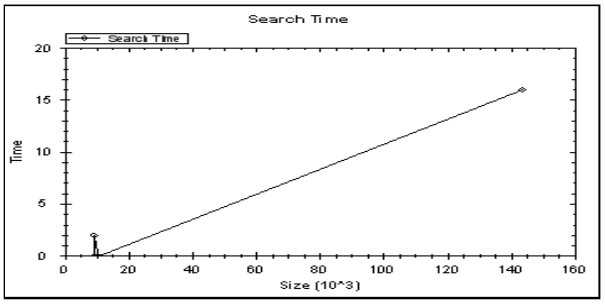

Fig. 2:- Collision rate plot diagram

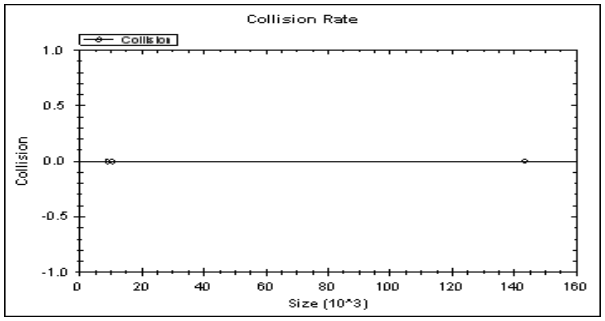

Fig. 3:- Search time plot diagram

XI. CONCLUSION

We have in this work, introduced a developed secured model for search into encrypted document collections. We have explored some techniques to securely search into the documents and retrieve the documents from an encrypted collection based on the encrypted search queries. The proposed secured model has improved the confidentiality of the query as well as the content of retrieved documents. In addition to our research work we will focus in future studies on the issue of encrypted data collection in general as well as data collection into multimedia documents aspect.

REFERENCES

[1]. Maxim Martynov, Boris Novikov, University of St.-Petersburg, Russia , :An Indexing Algorithm for Text Retrieval", Proceedings of the International Workshop on Advances in Databases and Information Systems (ADBIS'96).Moscow, September 10–13, Moscow, September 10–13, 1996.
[2]. R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. of the ACM Comp. and Comm. Security (CCS), Oct. 2006..
[3]. Ashwin Swaminathan, Yinian Mao, Guan-Ming Su, Hongmei Gou, Avinash L. Varna, Shan He,Min Wu, Douglas W. Oard. "Confidentialitypreserving rank-ordered search", Proceedings of the 2007 ACM workshop on Storage security and survivability - StorageSS '07, 2007.
[4]. S. E. Robertson and K. S. Jones, "Simple Proven Approaches to Text Retrieval," Technical Report TR356, Cambridge Univ. Computer Laboratory, 1997.
[5]. R. Brinkman, J. M. Doumen, and W. Jonker, "Using Secret Sharing for Searching in Encrypted Data," Workshop on Secure Data Management in a Connected World, LNCS 3178, pp.18-27, Aug. 2004.
[6]. B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private Information Retrieval," J. ACM, vol. 45, no. 6, pp. 965–982, 1998.
[7]. D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," IEEE Sym. on Research in Security and Privacy, pp. 44-55, May 2000.
[8]. D. Boneh, G. Crescenzo, R. Ostrovsky, G. Persiano, "Public-key Encryption with Keyword Search," Proceedings of Eurocrypt, 2004.
[9]. E-J. Goh, "Secure Indexes," Cryptology ePrint Archive, Report 2003/216,2003.
[10]. B. Klimt and Y. Yang, "Introducing the Enron Corpus," Conf. On Email and Anti-Spam (CEAS), Mountain View, CA, 2004.
[11]. R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order Preserving Encryption for Numeric Data," Proc. of SIGMOD, Jun. 2004.
[12]. D. K. Harman, "Common Evaluation Measures," in Appendix, Proceedings of Text Retrieval Conference, 2005. Available online at http://trec.nist.gov/.
[13]. Practical Cryptography, Bruce Schneider, Niels Ferguson, Boook; 2003; Wiley.Com.
[14]. Information Retrieval; Gonzalo Navarro; Book; 2004.
[15]. Microsoft Visual Studio. Net package 2008.