ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Efficient Keyword Search Using Text mining Techniques: a Survey

Banazir B, Annes Philip
MTech student (CSE), MES College of Engineering, Kuttippuram, India
Assistant Professor (CSE), MES College of Engineering, Kuttippuram, India

Abstract*—Centralized storage of data is implemented knowledge pool application where large number of users retrieves data from the document archiques. As part of security implementation data may be encrypted with any standard cryptographically algorithms. Searching data on encrypted document pool demands the decryption of entire data which is not effective and secure. Fuzzy keyword search extract relevant documents without decrypting whole data. This existing techniques also show poor performance when multiple keyword are input for the document retrieval. When multiple attributes are needed for searching information the level of output may vary. So the replacement of specific attribute on more generic concept can be applied called Textmining techniques.Textmining techniques optimizes the result and text time will be less than the other techniques. This technique is very important in application where large amount highly secured information stored in servers are searched. Survey four techniques for searching data on encrypted document which are Public Encryption with Keyword Search, Wild Card Based Searching, Gram based techniques and Text mining techniques.

***Index Terms*—Searchable encryption, Fuzzy keyword, Texmining.**

I. INTRODUCTION

Large quantities of sensitive personal data are retained for the purpose of network forensics and cyber investigations [1]. The advantages of the availability of such data for the investigation of serious crimes and the protection of national security are considerable. However, these advantages must be counterpoised by the dangers that such data could fall into the wrong hands. The encryption of retained data is a desirable counter measure against data theft. But how, then, can the investigator, such as the police or a secret service, search the data without having to decrypt the whole document. This seems to be a hard problem, as the criteria themselves may be sensitive and thus requiring protective measures, such as encryption.

We consider a scenario in which an investigator searches for data described by multiple keywords without revealing the keywords or the search results to the server. This scenario is akin to the private searching of streaming data presented [2]. While in the data is searched as it is generated (and can thereafter be discarded), in our scenario data is first stored in encrypted form and can be searched at a later stage. To provide a high level of security we make use of asymmetric cryptography. The server only possesses the public encryption key (and cannot decrypt the retained data itself). In this way, data that is already encrypted remains secure even against a strong adversary that breaks into the server. The decryption key is stored by a security server, which will only be involved when executing search queries. Public key Encryption with Keyword Search (PEKS) [2] is the first keyword searchable encryption based on a probabilistic public key system. It is more convenient to search cipher texts for multiple users.

Figure 1 presents a classic scenario, in which senders send searchable cipher texts to the proxy server of the receiver. In traditional search able encryption schemes allow a user to securely search over encrypted data through keywords and selectively retrieve files of interest, these techniques support only exact keyword search.

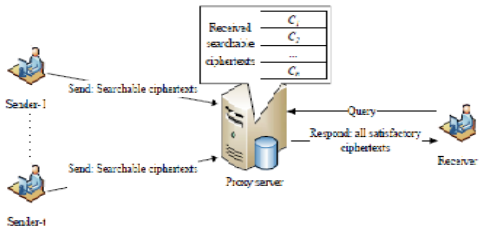
Fuzzy keyword search greatly enhances system usability by returning the matching files when users searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. In our solution, we exploit edit distance to quantify keywords similarity and develop two advanced techniques (Wildcard based, Gram based) on constructing fuzzy keyword sets, which achieve optimized storage and representation overheads. A new searching technique, SP (Sequential pattern) mining algorithm generates an optimal search result. Through rigorous security analysis, we have shown that our proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search.

II. RELATED WORK

A. Public Encryption with Keyword Search (PEKS)

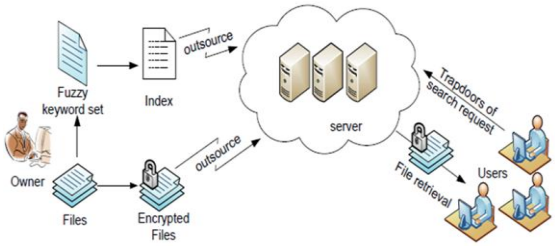
Public Encryption with Keyword search [2] can help to test the given keyword present in the document without learning anything else from the document. Data stored in untrusted server can be encrypted. Search the data by using keyword. By using PEKS reduce the processing time by retrieve only the selected files. By its disadvantage by using the application such as patient record and investigations, a small mistake on spelling on keyword cannot produce any result. Thus by going Fuzzy Keyword Searching.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013



B. Public Encryption with Fuzzy Keyword Search (PEFKS)

Fuzzy search algorithms (also known as similarity search algorithms) are a basis of spell-checkers and full-fledged search engines like Google. For example, these algorithms are used to provide the "Did you mean ..." function. Architecture as follows.



1) Wild card based techniques:

Wildcard based technique allows users to search for all files either name that contain similar qualities. The wildcard-based fuzzy set edits distance to solve the problems of performing edit distance at same position. For example, for the keyword MODERN with the preset edit distance 1, its wildcard based fuzzy keyword set can be constructed as

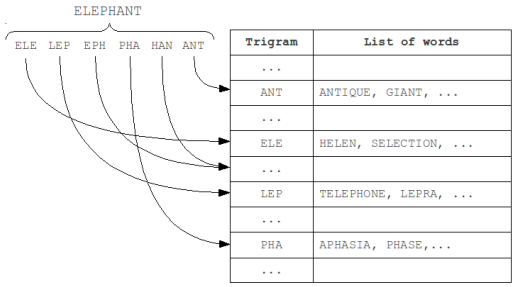
S (MODERN, 1) = MODERN, *MODERN,*ODERN, M*DERN,MO*ERN, MOD*RN, MODE*N, MODER*, MODERN*.

Edit Distance operation can be perform in 3 ways:

- Substitution: Changing one character to another in a word.
- Deletion: Deleting one character from a word
- Insertion: Inserting a single character into a word.

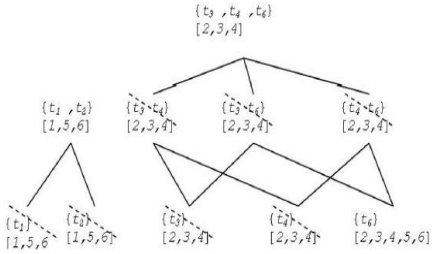
2) Gram based techniques:

This algorithm is based on the principle that if the word A matches with the word B considering several errors, then they will most likely have at least one common substring. For It works based on the grams. Gram of a string can be considered as a substring example



C. Text mining Technique:

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. Text mining Technique is used for fuzzy keyword search to find the multi keyword search. Here using SP(sequential pattern) mining algorithm. For example, pattern t3, t4, t6 in the taxonomy are usually more specific since they may be used only in positive documents. The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.



Here t1, t2, t3,t4 and t5 are terms in the document. The t1 is present in the 1,5, 6 sentences.The repetition is avoided and optimize the result. In the above case we only get the sentences is 1,5,6 and 2,3,4.

ALGORITHM 2 SPMining (SP,v)

Input: a set of frequent sequential patterns, SP; a set of pairs of term and its weight, v

Output: an updated set of pairs of term and its weight

Method:

1: foreach pattern P in SP do begin

2: sum supp += $supp_a$ (P)

3: end for

4: foreach pattern P in SP do begin

5: weight$'$ = $supp_a$ (P)/sum supp/len (P)

6: foreach term t in P do begin

7: if (t, weight)$\in v$ then

8: (t, weight) = (t, weight + weight$'$)

9: else

10: v$\leftarrow v \cup (t,$ weight$'$)

11: end if

12: end for

13: end for

III. OBSERVATION AND ANALYSIS

The table shows the performance of various schemes. The workload and communication cost are denoted by the number of keyword searchable cipher texts, n is the total number of keyword searchable cipher texts stored in the proxy sever, t is the number of keyword searchable cipher texts satisfied the query of the receiver.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Scheme	Time cost of the proxy server	Communication Cost	Time cost of receiver
PEKS	n	t	0
PEFKS	n	2t	2t
Text mining	n	t	2t

IV. CONCLUSION AND FUTURE WORK

Analyzed about different search schemes [2]-[9] which is based Public Encryption with Keyword Search, Public Encryption with Fuzzy Keyword Search (wild card based and gram based.) and Text mining techniques. Here Text mining techniques is efficient and privacy preserving. Text mining technique realizing the goal of fuzzy keyword search. Future work is on security mechanisms that support search semantics that takes into consideration conjunction of the complex natural language semantics to produce highly relevant search results and search ranking that sorts the searching results according to the relevance criteria.

REFERENCES

[1] Jan Camenisch, Markulf Kohlweiss, Alfredo Rial, and Caroline Sheedy. Blind and anonymous identity-based encryption and authorized private searches on public key encrypted data. In Proceedings of the 12th International Conference on Practice and Theory in Public Key Cryptography: PKC '09, Irvine, pages 196{214, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] Peng Xu and Hai Jin. Public-key encryption with fuzzy keyword search: A provably secure scheme under keyword guessing attack. Cryptology ePrint Archive, Report 2010/626, 2010. http://eprint.iacr.org/.

[3] Jin Li, Qian Wang, Cong Wang, Ning Cao, Kui Ren, and Wenjing Lou. Enabling efficient fuzzy keyword search over encrypted data in cloud computing. Cryptology ePrint Archive, Report 2009/593, 2009. http://eprint.iacr.org/.

[4] Mihir Bellare, Alexandra Boldyreva, and Adam O'Neill. Deterministic and efficiently searchable encryption. Cryptology ePrint Archive, Report 2006/186, 2006. http://eprint.iacr.org/.

[5] Jin Li, Qian Wang, Cong Wang, Ning Cao, Kui Ren, and Wenjing Lou. Fuzzy keyword search over encrypted data in cloud computing. In Proceedings of the 29th conference on Information communications, INFOCOM'10, pages 441{445, Piscataway, NJ, USA, 2010. IEEE Press.

[6] Changyu Dong, Giovanni Russello, and Naranker Dulay. Shared and searchable encrypted data for untrusted servers. In Proceedings of the 22nd annual IFIP WG 11.3 working conference on Data and Applications Security, pages 127{143, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] M. Chuah and W. Hu. Privacy-aware bed tree based solution for fuzzy multi-keyword search over encrypted data. In Proceedings of the 2011, 31st International Conference on Distributed Computing Systems Workshops, ICDCSW '11, pages 273{281, Washington, DC, USA, 2011.IEEE Computer Society.

[8] Saeed Sedghi, Peter Van Liesdonk, Svetla Nikova, Pieter Hartel, and Willem Jonker. Searching keywords with wildcards on encrypted data. In Proceedings of the 7th international conference on Security and cryptography for net-works, SCN'10, pages 138{153, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou. Privacy-preserving multi-keyword ranked search over encrypted cloud data. In INFOCOM, 2011 Proceedings IEEE, pages 829{837}, 2011.