International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Efficient Synonym based Multimedia Data Search Using Multi-Keyword Query

Manish M. Pardeshi[1], R. L. Paikrao[2]

[1]P. G. Student, Department of Computer Engineering, Amrutvahini C. o. E, Sangamner, Maharashtra, India.

[2]Head of Department, Department of Computer Engineering, Amrutvahini C. o. E, Sangamner, Maharashtra, India.

Abstract: ***Now a days cloud provides various featured facility and services of storage and searching of data on cloud. Efficient usage of cloud transaction is prime requirement as cloud is working on "pay-as-you-use" principle. The data of consumer which is on the cloud that are sensitive data like personnel emails, personal records, health reports, personal photos, financial data etc. As cloud is always a semi-trusted entity from user point of view hence encryption of such sensitive data before outsourcing to cloud is good idea. Encryption and then searching technique must be applied on encrypted cloud data. For encrypted data , traditional keyword search techniques or old searching methods are sometimes ineffective. So existing search approaches over encrypted cloud data , support only exact or fuzzy keyword search. Therefore, synonym-based multi-keyword ranked search over encrypted cloud data is tough but efficient with secured approach. Therefore to apply an effective searchable system with support of ranked search that is very challenging approach. To meet the challenge of effective search system, this paper proposes a practically efficient and flexible searchable scheme which supports both multi-keyword ranked search and synonym based search. Hence there must be system with an efficient approach to solve the problem of multi-keyword ranked search to achieve more accurate search result and very important search accuracy as single word search and synonym based search to support synonym queries search data on encrypted cloud. The Ranked search enables cloud customers to find the most relevant data in effective manner. Ranked search help to use network traffic efficiently as the cloud server sends back only the most relevant data. Therefore after experiment on real dataset shown proposed solution is very effective and efficient for multi keyword ranked searching in a cloud environment. Also there must be a module in system that helps to get labeled images as search result when labeled images are provided as an input to search***

Keywords: Cloud computing, consumer-centric cloud, keyword search, ranked search.

1. Introduction

Encrypted data search required new approach. Old search techniques are not useful. To achieve this multi-keyword ranked searching over encrypted data on cloud [1] provides us solution. This base paper [1]incorporates encrypted data search where we can get search result. Since data on cloud is sensitive sometime and in worst case for profit purpose there are chances of sharing this information with other resources.

Since cloud is semi-trusted entity besides it is curious also, hence user must be alert about his data on cloud. This encryption raises several issues like searching of such encrypted data because old search methods are not having efficiency. In case of transaction and view result he has to first download the data and then decrypt it , which impractica l since huge amount of raw data can be transact. Hence proposed system should have solution over encrypted data search present on cloud with the help of synonym multi keyword ranked search. Along with text file , labeled image file can be present on cloud. Hence proposed system should also support to search labeled image. In this case using OCR algorithm labels are extracted and matched while searching is performed.

2. Literature Survey

[1] is the base paper for the system proposed by Zhangjie Fu, Xingming Sun, Nigel Linge, Lu Zhou named Achiev-ing Effective Cloud Search Services:Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query launched in February 2014.This paper discussed idea regarding searching of data present on cloud in encrypted format. It is also having scope for multi-keyword search which are in ranked format. It helps us to find way for effective search for encrypted data.

In Year 2009, Georgia Koutrika presented a data cloud in which cloud search is performed on the basis of query sum-marization approach. He performed a query refinement model based on the summarization. Now based on summarization the query is presented to the web architecture and relatively the search is performed for reliable and effective cloud service [2].

A multimedia search for the cloud architecture is suggested by Wei-Ying Ma. In this work different multimedia services are suggested such as client PC, phone, TV etc. On the basic of the knowledge based search is performed to retrieve the multimedia analysis and will perform the search respective to the client request for the particular multimedia service [3].

Another tag based summarization approach is suggested by Byron Y.L. Kuo for the web search. The presented work is suggested on the public cloud. In which the integration of the web architecture and the database extraction is integrated. The work includes the refinement of the user query based on the cloud tags. The words extracted from the query are been summarized and this summarized query is passed to the public cloud. The cloud interface enabled the extraction of new and required information [4].

Another cloud search is suggested by Daniel E. Rose based on the information retrieval. The author presented his work

on Amazon cloud service. The work is tested under different criteria such as scalability, configuration etc. The presented search reduce the barrier to allow a person or the organization to perform the content oriented search and the search is tested under the enterprises environment as well as on web search[5].

Very basic search for encrypted data in presented by Li et al [6] that gives basic search idea. But this searching was for fuzzy keywords.

In Year 2012, Cengiz Orencik presented a rank based keyword search on the data cloud. In this work the document retrieval is performed on cloud server based on the keyword analysis and the information search is performed relative to the defined information. The presented work is performed on the encrypted data that has improved the security and the reliability of the retrieval. On this basis a secure protocol is suggested called Private Information Retrieval. The system will performed the query and present final results on the basis of parametric ranking.The presented work is the efficient comp u-tation and communication of the requirement analysis[7]

Mathew J. Wilson performed a work based on web search engine based for the keyword cloud. In this work the clouds are represented by some tags called the Meta data. The Meta data defines the cloud with relative parameters in terms of th e services will be done under different parameters. The first p a-rameter considered here is the most appropriate of its security, efficiency and the reliability criteria. On the basis of this the keyword match is performed on different cloud keywords. The work includes the learning stage for the keyword extraction and the comparative analysis is performed to extract the related cloud services from the system [8].

S. Kamara, and K. Lauter proposed a paper [9] that con-sider the problem of building a secure cloud storage service on top of a public cloud infrastructure where the service provider is not completely trusted by the customer. I. H. Witten, A. Moffat, and T. C. Bell proposed a technique for indexing [10] containing text compression ideas , indexing , querying , index construction and image compression.

S. Grzonkowski, and P. M. Corcoran proposed a concept of system which analyze and use the viewing pattern of consumers to personalize the program recommendations. So-lution works for user centric approach helps to share important document and services through TCP/IP structure.

While uploading labeled images , text must be extracted from image so that it will be helpful for searching images. This can be achieved by using OCR technique [12] proposed by Ayatullah Faruk Mollah , Nabamita Majumder , Subhadip Basu and Mita Nasipuri. Optical Character recognition technique is discussed and observed that 92.74 % accuracy is achieved.

Other related techniques. Allowing range queries over encrypted data in the public key settings, where advanced privacy-preserving schemes were proposed to allow more sophisticated multi-attribute search over encrypted files while preserving the attributes privacy. Though these two schemes provide provably strong security, they are generally not efficient in our settings, as for a single search request, a fu ll scan and expensive computation over the whole encrypted scores corresponding to the keyword posting list are required. Moreover, the two schemes do not support the ordered result listing on the server side. Thus, they cannot be effectively utilized in our scheme since the user still does not know which retrieved files would be the most relevant.

3. Proposed System

Proposed system flow is as follows

A. System Flow

Cloud server, data owner and data user are the entities involved in proposed system as per figure 1. Data owner is responsible to upload the collection of document to the server. This document can be text file or labeled image. Text documents are DC are encrypted first and then save on cloud. Hence cipher text is C is preserved on it. For searchable encrypted data , owner will also generate an index I that helps for efficient searching. Keywords are first extracted from DC and then set of distinct keywords are finalized and indexing is done. Actual data hosted on cloud will contain encrypted files C and searchable index I by data owner and data saving / uploading process is done.

While searching of data , based on keywords or synonyms of the predefined keywords entered by the user (has been authorized by data owner) system generate encrypted trapdoor using which cloud refers index I and manage to return search result. Encrypted documents are the part of search result. These documents are ranked and set of K documents are return. User can set the K parameter at time of search. System will then returns top K documents.

The proposed system works under following 4 sections.
1)Setup
2)GenIndex
3)GenQuery
4)Search

1) Setup: In this phase, the system is initialized. The data owner is only person who can generate the secret key SK and picks a random key sk .
2) GenIndex: The data owner calls procedure
2) buildindex(DC) to generate index for File F. Built index function uses stemmer, stopword, frequent word, ocr algorithms to generate index. User encrypt the index as well as File using SK. And upload the document and index to the cloud.
3) GenQuery: Data user generates trapdoor after getting the access to the data from data owner . with t keywords of interest in W , the query vector Q is generated. The word is query are encrypted and query is uploaded to the cloud.
4) Search: With the given interest of keyword appli-cation searches for a desired document using BST algorithm. And return the top K result to the user. While searching for the data it uses Latent semantic analysis technique to search for synonym keywords.

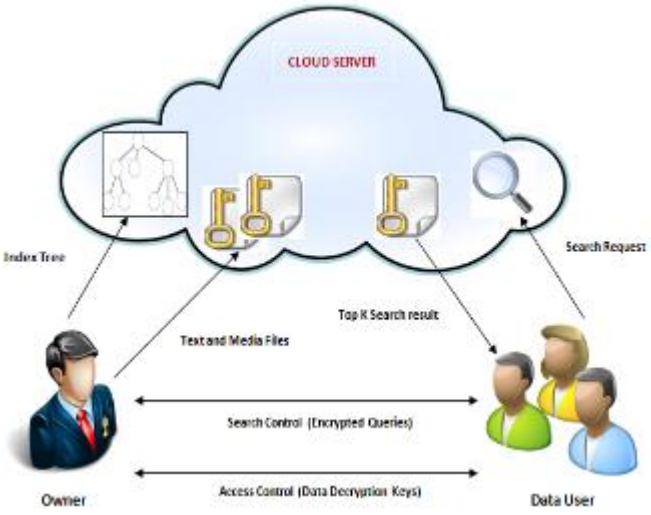

Figure 1: System Diagram

B. Algorithm Used

1) **Stemmer:** used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form e.g. word running is converted into run. This algorithm is used for index generation. **Input :** When user upload particular document then its inflected words are input **Steps :** Step 1 : Get rid of plurals and ed or ing suffixes Step 2 : Turns terminal y to i when there is another vowel in the stem Step 3 : Maps double suffixes to single ones ization , -ational etc. Step 4 : Deals with suffixes , -full , -ness etc. Step 5 : Takes off ant , -ence , etc. Step 6 : Removes a final e **Output :** Normalize word forms eg. Destructiveness =¿ destruct.
2) **Stop word:**
In this algorithm stopword like: punctuation marks, words like a,an,the are removed from the given text. **Input :** Stemmed words and sentences **Steps :** Step 1 : Get word list Step 2 : Calculate word count of frequently occurred words Step 3 : Create own stop word list Step 4 : Match these frequent words from stop word list Step 5 : Remove most frequent words **Output :** Filtration of words like the , and , a , to , of , was , it , in , that , he etc.
3) **Frequent word calculation:** Most frequent words above the threshold are calculated in this algorithm **Input :** Raw data with stemmed and without stop word **Steps :** Step 1 : Frequent words are buffered in hash table and word and its count is calculated Step 2 : Rank K is decide by the user and such top K results are shown to user from all calculated word counts **Output :** Dataset with word and its word count
4) **OCR:** For image tagging words are extracted from labeled image. **Input :** Labeled Image
5) **Steps :** Step 1 : Get labeled image Step 2 : Rough Preprocessing of image Step 3 : Search and recognition of the first character Step 4 : If feasible accuracy achieved
6) then goto step 5 else goto step 1 Step 5 : Position Evaluation of next character Step 6 : Again preprocessing of image Step 7 : Search and recognition of first character Step 8 : If feasible accuracy is not achieved goto step 5 else goto step 9 Step 9 : All characters are recognize and stop
Output : Text extracted by labeled images

5) **Keygen: Input :** Selection of key generation option **Steps :** Step 1 : Public key is generated Step 2 : Private key is generated **Output :** Public key for encryption and private keys for sharing is generated
6) **RSA:** RSA is used to encrypt and decrypt data stored on cloud **Input :** Raw text **Steps :** Step 1
: Choose two prime numbers, Prime1 and Prime2 to get the ProductOfPrime1Prime2 variable Step 2

: Find the Totient of ProductOfPrime1Prime2 (ProductOfPrime1Prime2) = (Prime1 -1) * (Prime2 1) Totient = (Prime1 -1) * (Prime2 1)

Step 3 : Get a list of possible integers that result in 1 mod Totient EncryptPrime * DecryptPrime = 1 mod Totient (Totient * AnyInteger) + 1 = 1 mod Totient Step 4 : Choose a 1 mod Totient value with exactly two prime factors: EncryptPrime and DecryptPrime
Step 5 : Actual Encryption CipherText = Plain-
^
TextEncryptPrime mod ProductOfPrime1Prime2 Step 6 : Actual Decryption
^
PlainText = CipherTextDecryptPrime mod ProductOfPrime1Prime2
Output : For encryption we get cipher text and for decryption we get plain text

7) **BST :** Binary search tree. To find appropriate docu-ment from I : searchable index tree
Input : Keyword for search **Steps :** Here k is the key that is searched for and x is the start node. BST-Search(x, k) Step 1 : y x Step 2: while y != nil do Step 3: if key[y] = k then return y Step 4: else if key[y] ¡ k then y right[y] Step 5: else y left[y] Step

6: return (NOT FOUND)
Output : Expected Search result

C. Mathematical Model

S = { I, O, P, U } U = { DO, DU }
Where, DO = Data owner DU = Data User

I={ DC, UAD, SK, PK,W }
where, DC = Uploaded Document UAD= User Authentication SK = Secret key W = set of n keyword to search

O={ EDC, DDC, In, TPK }
where, EDC = Encrypted Document DDC = Decrypted Document In = Index Tree TPK = Top K document

F={ UA, KG, ENC, DEC, GI ,SE, GQ, SS }
where, UA = User Authentication KG = Key Generation ENC = Encryption of Document using RSA DEC = Decryption of document usng RSA
GI = Generation of index tree

SE = search GQ = Query generation for W keywords SS = Synonym search

D. Set Theory

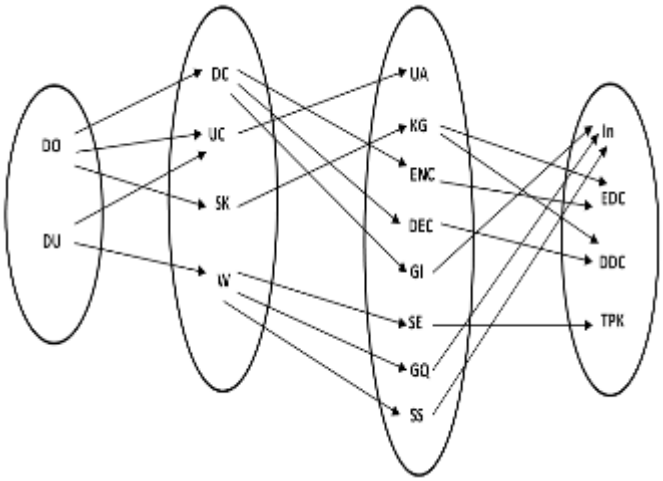

Figure 2: Set Diagram

E. Experimental Setup

For the betterment of search result over encrypted data on cloud , we have worked on synonym query technique. For this we have used wordnet API that provides us nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Also we have purchased cloud for 2 months to get real world system experience. We have used dataset Reuters News stories. From this we put 1000 training document and 400 test document on live cloud. For accuracy of result , dictionary of 2000 words is created and placed on cloud which is used for searching purpose.

F. Results

To verify some basic ideas we observe output for search result time and index generation time.

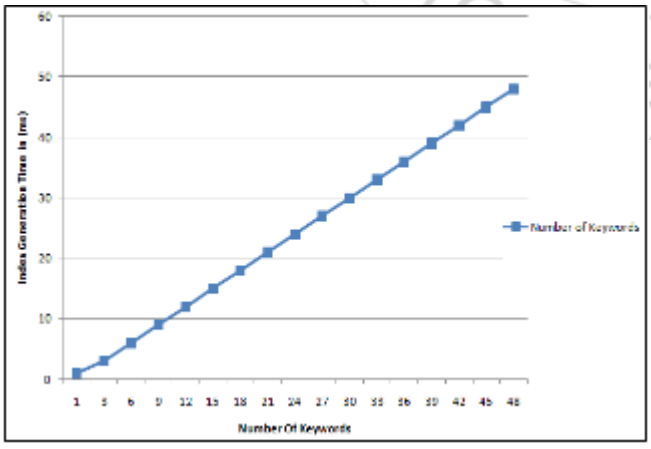

Figure A: Number of keyword Vs Index Generation Time

From fig A it is clear that for same size of dataset , time required to generate index is directly proportional to the number of keywords as expected. Time is shown in milliseconds.

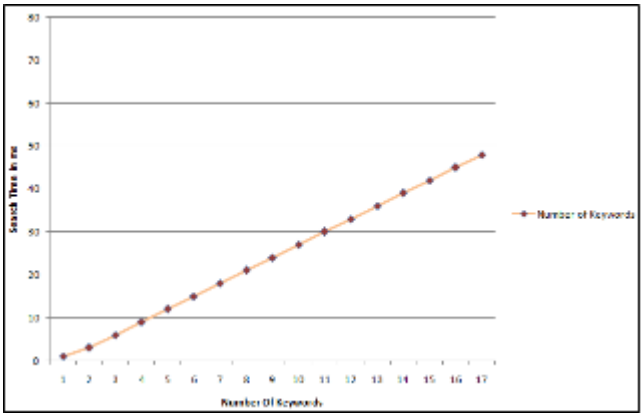

Figure B: Number of keyword Vs Document Search Time (in ms)

From fig B it is also clear that as keywords increases , then for same size of dataset time required is directly proportional to the number of search keywords given by the user. While searching number of keywords are mapped and for betterment ranking of search result is done. Hence it is obvious that time should increased with number of keywords.

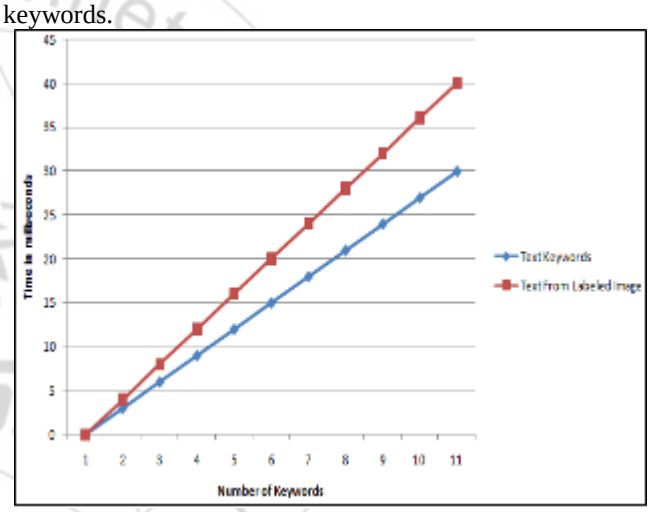

Figure C: Number of keyword Vs Document Search Time

(in ms) for text as input to search and labeled image as input to search

In our system we can give labeled image as input. Labeled are extracted through OCR algorithm hence extraction time is added in search time. Hence in comparison with normal text as input , time required is greater for labeled image as input.

It is prime requirement that labeled image should contain labels present in built dictionary so mapping can be done.

4. Conclusion

Proposed system is decentralized system in which distributed nodes work together for data security on cloud by implementing encryption facility , also these nodes manages multi user tasks like sharing , writing data , reading data etc. Due to this decentralized approach keys are managed at different node hence cloud is not having keys for decryption hence data security is assured. Also KDC is not having data hence only encryption keys are not useful to it. Three types of user like owner , writer and reader has respective access

control to the data. Hence this system is also manages hierarchical scenarios as far as users role is concern.

References

[1] Zhangjie Fu,Xingming Sun Nigel Linge, Lu Zhou,"Achieving Effective Cloud Search Services:Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query ", IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, February 2014.
[2] Georgia Koutrika,Saint Petersburg, Russia "Data Clouds: Summarizing Keyword Search Results over Structured Data", EDBT 2009, March 2426,ACM, pp. 391-402,2009.
[3] A multimedia search for the cloud architecture is suggested by Wei-Ying Ma
[4] Byron Y-L. Kuo (2007), "Tag Clouds for Summarizing Web Search Results" ,WWW 2007, May 812, 2007, Banff, Alberta,Canada.pp.1203.
[5] Daniel E. Rose (2012), "CloudSearch and the Democratization of Information Retrieval Wei-Ying Ma (2009),Rethinking Multimedia Search in the Clients + Cloud Era, LS-MMRM09, October 23, 2009,Beijing, China. Pp. 1-1.
[6] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing", Proceedings of IEEE INFOCOM10 Mini-Conference, San Diego, CA, USA, pp. 1-5, Mar.2010.
[7] Cengiz Orencik (2012), Efficient and Secure Ranked Multi-Keyword Search on Encrypted Cloud Data , PAIS 2012, March 30, 2012, Berlin,Germany. ACM, p 186-195.
[8] Mathew J. Wilson (2012), "Keyword Clouds: Having Very Little Effect on Sensemaking in Web Search Engines" , CHI 2012, May 510, 2012, Austin,Texas, USA,ACM,p2069-2074.
[9] S. Kamara, and K. Lauter, "Cryptographic cloud storage", ,FC 2010 Workshops, LNCS 6054, PP. 136-149, Jan. 2010.
[10] I. H. Witten, A. Moffat, and T. C. Bell, "Managing gigabytes: Compressing and indexing documents and images", Morgan Kaufmann Publishing:San Francisco, May 1999, PP. 36-56
[11] S. Grzonkowski, and P. M. Corcoran, "Sharing cloud services: user authentication for social enhancement of home networking", IEEE Trans.Consumer Electron., vol. 57, no. 3, pp. 1424-1432, 2011.
[12] Ayatullah Faruk Mollah , Nabamita Majumder , Subhadip Basu and Mita Nasipuri , "Design of an Optical Character Recognition System for Camerabased Handheld Devices", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.
[13] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou, "Privacy preserving multi-keyword text search in the cloud supporting similarity based ranking", ASIACCS 2013, Hangzhou, China, May 2013, pp. 71-82,2013.
[14] Mehmet Kuzu, Mohammad Saiful Islam, "Efficient Similarity Search over Encrypted Data Department of Computer Science", The University of Texas at Dallas Richardson, TX 75080, USA.
[15] D. A. Grossman, and O. Frieder," Information retrieval: algorithms and heuristics", 2nd ed., Springer Publisher: Berlin, 2004, pp. 18-20.

References

Manish M. Pardeshi completed B.E.(Computer) from Gokhale Education society's college of Engineering, Nashik and Pursuing Master degree (Computer Science and Engineering) from UoP, Amrutvahini College Engineering, (AVCOE), Sangamner.

Prof. Rahul L. Paikrao is a associate professor and head of computer engineering department at Amrutvahini College of Engineering (AVCOE), Sangamner (Computer Science and Engineering), Pursuing PhD in Cloud computing security from UoP, He has 10 years of teaching experience.