Composite Keyword Based Search Over Data on Remote Information

G. Senthil Kumar¹, T.S. Shiny angel¹, K. Vijayakumar¹, Akhilesh Chamoli¹, Sarang Kejriwal¹ and N. Snehalatha¹
¹ Department of Software Engineering, SRM Institute of Science and Technology, Chennai 603203, India.
Email: senthig1@srmist.edu.in

Abstract: Majority of data owners have moved their services to the cloud servers. Members of any organization can access data from the cloud by using keyword-based search. Being based on Boolean keyword matching system or semantic search they are unable to provide accurate search results as well as unable to understand misspelled keywords. With the existing searching mechanisms like keyword search in place, the search results are often slow and they result in missing of information and produce in accurate results. In this paper, a K-Gram based approach is proposed to improve the search efficiency. Using this approach, the words are divided into groups and then they are identified by the corresponding Indexes. This proposed approach can improvise secure and efficient retrieval of information.

Keywords: Cloud data; Encryption; top-k similarity search; file retrieval, semantic search

1. Introduction
Cloud computing infrastructure can be good technology and it can accelerate the progress of large-scale information storage, process and execution [1]. Once, data owners source their personal information on public cloud servers, they are not in their assurance management domain, but security and personalization are a major concern. To prevent data leaks, tactile records must be encrypted before being uploaded to the cloud server, which creates a large-scale organization to assist inexperienced keyword-based queries and rank-match results on encrypted data. Most current works that do not have appropriate rating schemes for key rating questions [2]. In the current multi-keyword hierarchical search process, the keyword dictionary is stable and increases the number of keywords only once.
What’s more, it does not take into account user behavior and keyword access frequency. For a query matching result that contains large quantities of documents, the out-order ranking is reduced. These data annoy the patron to hunt a set that is too high to satisfy its desires. In this paper, we justify the equivalent multi-keyphrase query concept referred to as MKQE to address similar errors. The MKQE keyword dictionary can significantly reduce security overhead within the boom. It takes into account keyword weight and user access history when generating query results. Therefore, documents with high access frequency environments that are close to the user’s access history will rank higher in the similar output.
2. Related Work
D. X. Song et al [1] proposed a method which have variety of essential blessings like they will be incontrovertibly secure: they supply obvious secrecy for secret writing, inside the texture that the untrusted server cannot examine a number of hassle concerning the plaintext whereas simplest given the cipher text; they supply question isolation for searches, owing to this that the untrusted server cannot studies some part additional concerning the plaintext than the search quit stop result; they supply managed looking, so as that the untrusted server cannot seek for Associate in Nursing discretionary phrase while not the purchaser’s authorization; more they helpful resource hidden queries, so as that the person might in addition what is more raise the untrusted server to seem for word while not revealing the phrase to the server. The algorithms that we tend to square measure providing square measure simple, speedy (for a record of amount n, the secret writing and look for algorithms simplest wish O(n) flow cipher and block cipher operations), and introduce nearly no house and voice communication overhead, and eventually square measure realistic to use recently. Fuzzy multikey search proposed by wang et al [2]. The basic method of defensive records is to encrypt it before outsourcing, but retrieving the required documents from the encrypted cloud becomes a hassle of viewing encrypted data. various schemes had been projected to address this hassle of wanting over encrypted cloud data, however no longer one in every of the existent schemes offer finest client look for experience cherish plaintext look for. The theme mainly improves the index generation time and the use of a binary tree based on dynamic indexes to reduce the time spent in the analysis of reward schemes
Searching efficiently based on similarity which is proposed by Kuzu et al [3]. Cloud computing has several appealing countenances, owing to those countenance; it’s become masses easier to carry great quantity of information thereon. Therefore, keeping the knowledge on cloud facilitates character to urge admission to the info from any region on each occasion its miles required. For this reason, owing to the big facts, a very huge mission emerges i.e. protection and privateness. secret writing permits person to save lots of you their statistics from embezzled get right of access and for this reason it improves powerfulness, but furthermore convolutes a number of the essential talents. On this have a glance at, we tend to suggest a cherished theme for similarity look for and linguistics square measure seeking over encrypted records.
Wang et al [4] proposed a privacy search based mechanism and the statistics made by means that of people and corporations that require to be hold on and applied are unexpectedly increasing, statistics house owners are prompted to source their native advanced facts management systems into the cloud for its exceptional flexibility and financial money savings. However, as sensitive cloud information may additionally furthermore get to be encrypted prior outsourcing, that obsoletes the normal records utilization institution based on plaintext key-word are seeking out, to allow privateness-assured utilization mechanisms for outsourced cloud info is thus of dominant importance. Considering the large amount of on-call for info purchasers and large quantity of outsourced info documents in cloud, the effort is principally tough, because it is kind of exhausting to satisfy furthermore the smart needs of universal overall performance, system usability, and immoderate-degree shopper trying stories. In this paper, we have a tendency to examine the attempt of secure and inexperienced equality seeking outsourced cloud statistics. Similarity ask for is an important and powerful device broadly speaking applied in plaintext facts retrieval, but has no longer been pretty explored at intervals the encrypted records region. Our mechanism style 1st exploits a suppressing approach to assemble storage-inexperienced similarity key-phrase set from a given record series, with edit distance as a result of the similarity metric. primarily based altogether totally on that, we have a tendency to then construct a private trie-traverse trying index, and show it effectively achieves the represented similarity are seeking out capability with normal try to seek out time quality.
Orencik et al [5] investigated search on encrypted data. Cloud computing generation become additional more well-known every year, as several agencies unremarkably have a tendency to source their statistics utilizing robust and speedy offerings of clouds whereas decreasing the fee of hardware possession. The proposed have
a tendency to suggest associate in Nursing economical privateness-maintaining ask for technique over encrypted cloud information that creates use of min hash capabilities. most of the add literature will best guide one operate are finding out in queries that reduces the effectiveness. we have a tendency to furthermore integrate a strong ranking practicality this can be primarily based altogether undoubtedly altogether on fundamental measure frequency inverse file frequency (tf-idf) values of key-word report pairs. Our assessment demonstrates that the planned theme is proven to be privateness-maintaining, inexperience and powerful.

Li et al [6] proposed a methodology to calculate linguistic similarity between words in information retrieval, information acquisition, and in various fields. Popular studies mainly target the same ideas that are made up of the same words. For compound ideas made up of multiple words, they usually ignore the specific structural choices of the combinations and fully process them into single ideas, which can affect the final word accuracy. Throughout this paper, we got to propose a fully specialized ontology-based compound thinking linguistics analogous computational approach, referred to as CCSS, which exploits thought-making alternatives. In CCSS, compound headings and auxiliary words (SaA) are decomposed, and therefore the relationship between these devices is a habit for proof of similarity. Shadow detection, except for mistakes, is corrected. In addition, an acute knowledge of metaphysics of receiving enchanting functions, neighborhood density and intensity was taken into account. In-depth experimental evaluation shows that our approach outperforms the current approach. Zhou et al [7] assessed results using statistics content material cloth (IC) and it could be a crucial size of assessing the linguistics similarity among terms or word senses in phrase information. The traditional technique of effort IC of phrase senses is to combine ability of their hierarchical type from Associate in nursing metaphysics like word web with actual utilization in matter content as derived from an outsized corpus. In this paper, a brand cover version of IC is obtainable, this can be based totally on hierarchical form on my own. The version considers not solely the hyponyms of every word expertise however furthermore its intensity at intervals the shape. The IC rate could be a quite a little tons abundant less powerful to calculate based totally on our version, and while used because the thought of a similarity approach it yields judgments that correlate bigger rigorously with human assessments than others, that the utilization of IC rate nonheritable solely considering the hyponyms and IC fee got via means of exploitation corpus assessment.

Question process queries [8] that preserves the facts privacy of the man of affairs and also the question privacy of the patron may be a spanking new analysis hassle. It advises to improve interest in remote information as it drives further teams to move their statistics and querying offerings. Usually, gift analysis, which contains those on data outsourcing, address the statistics privateness and question privateness one when the opposite and can't be accomplished to the present problem. The framework is scalable to huge datasets with the helpful resource of investment Associate in Nursing index-based wholly approach.

Cloud-centric cloud computing example has emerged in recent years as a result of the phenomenon of intelligent virtual gadgets combined with a growing cloud computing generation, supported by Z. Fu et al.[9].An expansion of cloud offerings square measure introduced to the shoppers with the premise that a robust and economical cloud request supplier is achieved. For customers, they have to find the utmost applicable product or statistics, that is pretty applicable at intervals the “pay-as-you use” cloud computing paradigm. This paper proposes an effective method to solve the pitfalls of multi-key-phrase-hierarchical class measurement by searching for encrypted cloud data that serves equivalent word queries. The discourse contribution of this paper is summarized in 2 parts: a multi-key-phrase-hierarchical class measure, which requests large accuracy research effects and a similar word-based search, to aid in the use of similar words to a supportive support-based query. Large experiments were performed on real-international datasets. Verify Tactic indicates that the approximate solution for a multi-key-phrase hierarchy in a cloud environment is very effective and inexperienced.
Chai et al [10] verified symmetric encryption by outsourcing statistics to cloud servers, at the equal time as growing organization availability and decreasing client’s burden of handling statistics, perpetually brings in new issues aboard part statistics privacy, due to the very fact the server is also sincere-but-curious. To mediate the conflicts amongst statistics usability and statistics privateness within the kind of situation, studies of searchable coding is of growing hobby. prompted by exploitation the reality that a cloud server, except its hobby, is egocentric in order that you'll be able to keep its computation and/or down load information measure, a observation was done at intervals for the presence of a half dedicated but with curiosity, To combat within the direction of this handiest somebody ever, a verifiable south southeast (VSSE) theme is projected to supply verifiable search ability in larger to the statistics privacy, each of which can be additionally showed through manner of our rigorous safety assessment. Besides, we have a tendency to address the practicality/normal overall performance as a vital demand of a searchable coding theme. To illustrate the slight weight of our theme, we have a tendency to completed and examined the projected VSSE on a laptop (serving because of the very fact the server). The experimental outcomes with a bit of fine fortune endorse that the projected theme satisfies all of our format goals.

3. Composite Keyword Based Search Framework

![Composite Keyword Based Search Framework](image)

Figure 1 Composite Keyword based framework

The proposed Composite keyword-based framework is shown in Figure 1. It consists of 3 main phases as follows,
1. **Login/New User**
   The login and new user module enable the user to access and allows first user to sign up by registering.

2. **Upload File:**
   The admin of the organization can upload the necessary files for the user to search those later and download it.

3. **Search**
   In the proposed framework the searching is performed by three step process

   3.3.1 **Frequent Search.**
   The words are searched based on letters entered by the user. Results will be shown according to input value.

   3.3.2 **Similarity Search.**
   Based on the word net dictionary the similar words are searched and user can choose expected word from dropdown lists which is shown in Figure 2.

   ![Similarity search](image)

   **Figure 2 Similarity search**

   3.3.3 **Linear search:**
   In this process, Searches are performed based on keywords. The matching words with same key are shown in figure 3.
3.3.4 Mail alert method

The file is searched then file is downloaded from user’s mail. The user gets the mail and the file to be downloaded.

3.3.5 File Downloads method:
File can be downloaded by user when he gets the email alert and by user personal email, he can get the file.

4 K-gram algorithm

Rabin-Karp[11][12] is the algorithm which is based on pattern matching. It matches the hash value of input string and accordingly finds the similar outputs. In this Rehashing is also done in substring.

Input: The input file
Output: The stream which is encoded

1. inputFile = fetch the user file
2. countOfGrams = In the input file count the grams
3. do
4.  store_5_gram = fetch1st5 grams from input
5.  indexOfGrams = find (store_5_gram, dict_of_5_gram)
6.  if indexOfGrams>=0 then
7.    force_4_gram_compress(store_4_gram)
8.    output +=compress(indexOfGrams,5)
9.    delete 1st5 grams from input
10.   countOfGrams-= 5 (while countOfGrams>=5)
11. end
12. else
13.    store_4_gram += fetch 1st gram of the input File
14.    remove 1st gram of the input File
15. countOfGrams:= 1
16. if number of grams of store_4_gram= 4 then
17. 4_gram_compress(store_4_gram)
18. end
19. end
20. end
21. if countOfGrams> 0 then
22. 4_gram_compress(input File)
23. end

![User Interface](image)

Figure 4 User Interface

The user interface is shown in Figure 4. Being a larger capacity, a model can store many instances with very sanctionative, small experiments to quickly rescale expeditiously.

5 Conclusion
A Composite Keyword based search on remote information allows the private organization to secure their files in the private cloud and the user can search the file. In this paper, we have proposed an approach for searching efficiently and at the same time assure facts privateness. Then, with a purpose to enhance the search efficiency, we design the K-gram search scheme, which searches the filename if there is one letter or two letter missing in whole name. The file can be downloaded from email when the user searches and want that particular file. This enables to keep the safety of the files inside the cloud.

References
[1] D. X. Song, D. Wagner, and A. Perrig, “Practical techniques for searches on encrypted data,” in IEEE Symposium on Security and Privacy, 2000, pp. 44–55.
[2] B. Wang, S. Yu, W. Lou, and Y. T. Hou, “Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud,” in IEEE International Conference on Computer Communications, 2014, pp. 2112–2120.

[3] M. Kuzu, M. S. Islam, and M. Kantarcioglu, “Efficient similarity search over encrypted data,” in IEEE 28th International Conference on Data Engineering (ICDE), 2012, pp. 1156–1167.

[4] C. Wang, K. Ren, S. Yu, and K. M. R. Urs, “Achieving usable and privacy-assured similarity search over outsourced cloud data,” in 2012 Proceedings of IEEE INFOCOM, 2012, pp. 451–459.

[5] C. Orencik, M. Kantarcioglu, and E. Savas, “A practical and secure multi-keyword search method over encrypted cloud data,” in IEEE Sixth International Conference on Cloud Computing (CLOUD), 2013, pp. 390–397.

[6] M. Li, B. Lang, and J. Wang, “Compound concept semantic similarity calculation based on ontology and concept constitution features,” in Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on, 2015, pp. 226–233.

[7] Z. Zhou, Y. Wang, and J. Gu, “A new model of information content for semantic similarity in wordnet,” in Future Generation Communication and Networking Symposia, 2008. FGCNS’08. Second International Conference on, vol. 3. IEEE, 2008, pp. 85–89.

[8] H. Hu, J. Xu, C. Ren, and B. Choi, “Processing private queries over untrusted data cloud through privacy homomorphism,” in 2011 IEEE 27th International Conference on Data Engineering, 2011, pp. 601–612.

[9] Z. Fu, X. Sun, N. Linge, and L. Zhou, “Achieving Effective Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query,” IEEE Transaction Consumer Electronics, vol. 60, no.1, pp. 164-172, 2014.

[10] Q. Chai and G. Gong, “Verifiable Symmetric Searchable Encryption for Semi- Honest-but-Curious Cloud Servers,” IEEE International Conference on Communications (ICC’12), pp. 917-922, 2012.

[11] S. K. Shivaji and P. S., "Plagiarism Detection by using Karp-Rabin and String-Matching Algorithm Together," International Journal of Computer Applications, vol. 116, no. 23, pp. 37-41, 2015.

[12] M. Cebrián, M. Alfonseca and A. Ortega, “Towards the Validation of Plagiarism Detection Tools by Means of Grammar Evolution,” IEEE Transactions on Evolutionary Computation, vol. 13, no. 3, pp. 71-77, 2009.