Conference Proceeding Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Compound Keyword Search over Encrypted Cloud Data by Semantic Base Method

A. Akilambigai,
Master of Computer Science in Kamban college of Arts and Science, Thiruvannamalai, India

***Abstract*—**Keyword search over encrypted information is crucial for accessing outsourced sensitive information in cloud computing. In some circumstances, the keywords that the user searches on ar solely semantically associated with the information instead of via a precise or fuzzy match. Hence, semantic-based keyword search over encrypted cloud information becomes of preponderant importance. However, existing schemes sometimes depend on a world wordbook, that not solely affects the accuracy of search results however additionally causes unskillfulness in information change. in addition, though compound keyword search is common in follow.

Index Terms*—Searchable Cryptography, Semantic-Based Keyword Search, Linguistics Similarity, Compound Thought.*

I. INTRODUCTION

In cloud computing, associate degree increasing range of non-public or enterprise users source their information to cloud storage to fancy the advantages of ”pay-on-demand” services and high computation performance. To preserve privacy, users value more highly to cipher information before outsourcing. Thus, the standard key-word search can't be directly dead on the encrypted information that limits the use of information. to deal with this downside, Song et al. [1] projected the thought of searchable cryptography (SE) that permits users to look on encrypted information through a keyword. After, varied searchable cryptography schemes were projected to satisfy totally different re-quirements like fuzzy keyword search [2]–[4], multi-keyword search [5]–[8], stratified keyword search [9]–[11], and semantic-based keyword search [12]–[17].

1. An ontology-based compound thought linguistics similar-ity (CCSS) calculation technique is projected to boost the accuracy of compound similarity activity. By shrewd the linguistics similarity between keywords and field topics mistreatment CCSS, the linguistics characteristics ar introduced into the keyword vector.
2. By integration CCSS, LSH and SkNN, we tend to propose a semantic-based compound keyword search theme (SCKS) over encrypted information. cashing in on these techniques, SCKS will at the same time support semantic-based keyword search, multi-keyword search, stratified keyword search and economical information update.
3. We analysis the protection of SCKS and any propose a security-enhanced theme, SE-SCKS. the protection analysis indicates that SE-SCKS is semantically secure underneath the adjustive model and may be employed in circumstances requiring the next security level.
4. We implement and take a look at CCSS and SCKS on a real-world dataset. The experimental results demonstrate that our approaches are correct and economical.

II. RELATED WORK

A linguistics-based keyword search theme returns results according to the semantic connexion between documents and a question . supported the co-occurrence likelihood of terms, Sun et al. [12] and Xia et al. [13] severally created a linguistics relationship library (SRL) to record the linguistics similarity between keywords. within the search section, the question keywords ar enlarged primarily based upon the SRL, and therefore the ex-tended question keywords ar employed in the search algorithmic program. Fu et al. [15] extended the question keywords by introducing m-best tree and term similarity tree, each of that ar designed supported the WordNet. Similarly, the keyword set in [14] is extended via a equivalent word wordbook. within the conceptual-graphs primarily based search theme [16], some sentences ar extracted to represent the documents and therefore the linguistics search is implemented by shrewd the connectedness score between the sentences within the document and therefore the question. in keeping with the characteristic that connected keywords sometimes have an equivalent root, the theme projected by Moataz et al. [17] extracts the keyword root by a algorithmic rule and searches with the basis rather than the keywords. Obviously, this technique cannot work once the semantically connected keywords have totally different roots.

III. PRELIMINARIES

Scheme Model

In our theme, the index of document and therefore the question ar drawn with the VSM. A document index is denoted by a vector generated with the keywords of the document, and a secure index is that the encrypted index. Similarly, a question could be a vector generated with the keywords of a probe, associate degreed a trapdoor is an encrypted question. In general, the documents are often encrypted by ancient cryptography schemes like AES. That specializes in the index and question; our theme consists of the subsequent algorithms.

- Keygen (d). This algorithmic program is dead by the information owner or a trusty authority (TA). Taking a security parameter d as input, the algorithmic program outputs a system trigonal key sk.
- BuildIndex (sk; D). This algorithmic program is dead by the information owner. supported the trigonal key sk and therefore the document D, the algorithmic program generates the secure index I(D).
- Trapdoor (sk; Q). This algorithmic program is dead by the information owner or Ta. With the keyword set Q that the user needs to look, the algorithmic program generates the corresponding trapdoor T (Q).
- Search (I(D); T (Q)). This algorithmic program is dead by the cloud server. supported the trapdoor T (Q) and every secure index I(D) keep within the server, the cloud server calculates the correlation coefficients between the question Q and every

document D and returns the stratified correlation coefficients to the user.

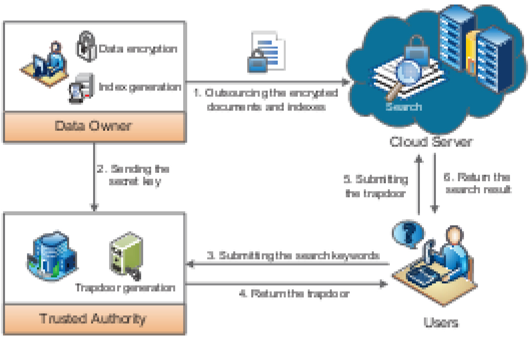

Data encryption
Index generation
Data Owner
1. Outsourcing the encrypted documents and indexes
Search
Cloud Server
2. Sending the secret key
5. Submitting the trapdoor
6. Return the search result
3. Submitting the search keywords
4. Return the trapdoor
Trapdoor generation
Trusted Authority
Users

Fig. 1. The model of keyword search

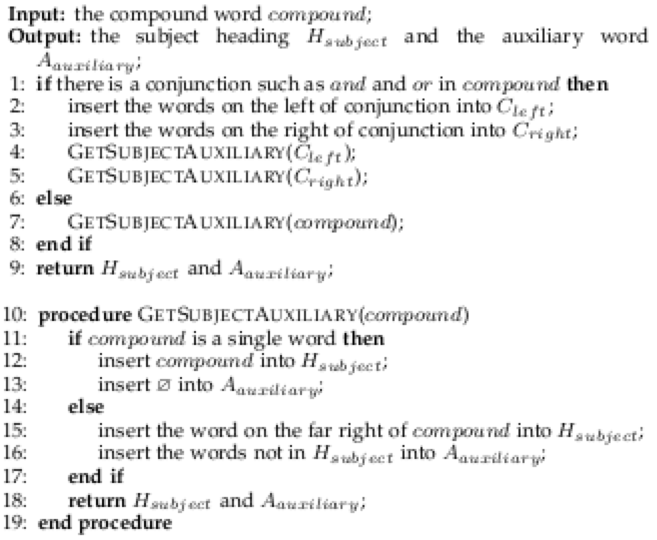
The running method of our theme is represented in Fig.1.First, the information owner publishes the encrypted documents and secure indexes to the cloud server. to scale back the computation burden, the information owner is allowed to source the generation of the trapdoor to Ta by giving the personal key thereto.

IV. COMPOUND THOUGHT LINGUISTICS SIMILARITY CALCULATION TECHNIQUE

Features of Compound ideas

Depending upon their linguistics constituents, the compound ideas are often divided into 2 types: endocentric structure and exocentric structure. The endocentric structure is true once one or additional constituents of a compound will play the central role and function a determinable subject heading. as an example, the topic heading of "information retrieval" is "retrieval". The exocentric structure is true once there's no subject heading in a very compound. as an example, "pick pocket" are often expressed as "a one who picks a pocket" which suggests neither "pick" nor "pocket" however rather "a person". we tend to found that the majority compounds of English have associate degree endocentric structure, whereas compounds with exocentric structure ar normally employed in informal things [45]. Hence, during this paper, we tend to concentrate on the endocentric structure compounds.

Algorithm 1: Subject headings and Auxiliary words Recognition (SAR)



```
Input: the compound word $compound$;
Output: the subject heading $H_{subject}$ and the auxiliary word $A_{auxiliary}$;
if there is a conjunction such as $and$ and $or$ in $compound$ then
    insert the words on the left of conjunction into $C_{left}$;
    insert the words on the right of conjunction into $C_{right}$;
    GETSUBJECTAUXILIARY($C_{left}$);
    GETSUBJECTAUXILIARY($C_{right}$);
else
    GETSUBJECTAUXILIARY($compound$);
end if
return $H_{subject}$ and $A_{auxiliary}$;

procedure GETSUBJECTAUXILIARY($compound$)
    if $compound$ is a single word then
        insert $compound$ into $H_{subject}$;
        insert $\varnothing$ into $A_{auxiliary}$;
    else
        insert the word on the far right of $compound$ into $H_{subject}$;
        insert the words not in $H_{subject}$ into $A_{auxiliary}$;
    end if
    return $H_{subject}$ and $A_{auxiliary}$;
end procedure
```

Fig. 2. algorithmic program of subject headings and auxiliary words recognition

Linguistics Similarity Calculation for Compound

To measure the linguistics similarity of compound ideas, we tend to propose a unique approach that considers the thought constituent options {and several and a number of alternative and several other} other factors influencing similarity. The compound ideas are rotten into SaA by mistreatment algorithmic program one. The compartmentalization options, local density, path length and depth of ontology also influence the similarity between concepts; therefore, we respectively calculate the impact factors associated with them and then comprehensively consider all of the impact factors to im-prove the accuracy.

V. COMPARTMENTALIZATION OPTIONS AND NATIVE DENSITY OF METAPHYSICS IDEAS

For the approaches basing on metaphysics methods, all attainable compartmentalization links (i.e., paths) between ideas are calculated, however solely the shortest one is unbroken. to boost the accuracy, other available taxonomical features should be considered. Moreover, the local density of semantic nets affects the similarity between concepts. Definition 3. Impact factor of local density can be written as follows

$$F_{density}(c_1,c_2)=\sqrt{1-\frac{||\Phi(c_1)|-|\Phi(c_2)||}{|\Phi(c_1)|+|\Phi(c_2)|+1}} \qquad (7)$$

where P ∈ [0, 1], and therefore the vary of ϵ is [−1/2,1/2]. P = one means the SaA recognition is precisely correct, and P = zero means the SaA recognition is totally incorrect. density, path length and depth of metaphysics additionally influence the similarity between concepts; so, we tend to severally calculate the impact factors related to them and so comprehensively take into account all of the impact factors to improve the accuracy.

VI. SEMANTIC-BASEDCOMPOUNDKEYWORD SEARCH SCHEME

Overview

In our scheme, VSM and the topic set in a field are used to construct the semantic vector for each keyword. More specifically, in the keyword vector, each element corre- spends to a field topic, and its value is the similarity between the topic and the keyword, which is obtained? using CCSS. Because the topics are almost invariable, the Dimensionality of the keyword vector will not change with the adding or deleting of the keywords or documents, which is helpful in supporting data update.

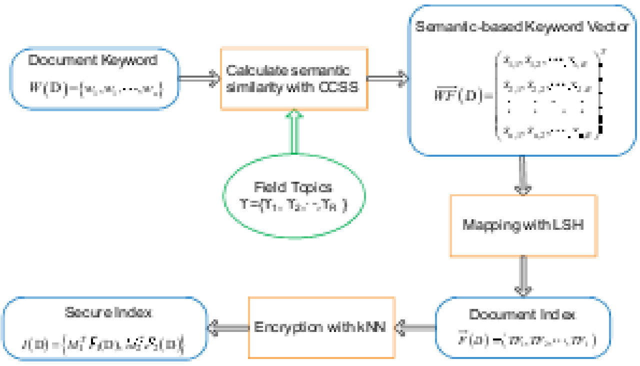
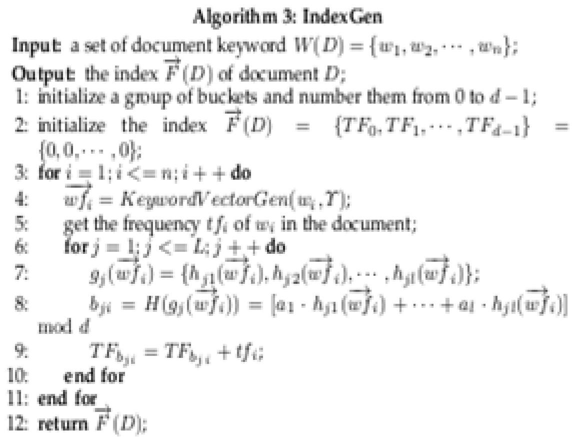
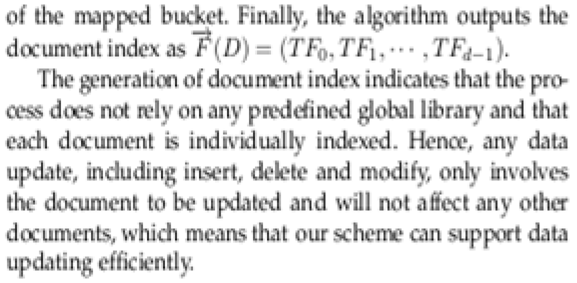


Fig. 3. Proposed secure index generation process



```
Algorithm 3: IndexGen
Input: a set of document keyword $W(D) = \{w_1, w_2, \cdots, w_n\}$;
Output: the index $\overrightarrow{F}(D)$ of document $D$;
initialize a group of buckets and number them from 0 to $d-1$;
initialize the index $\overrightarrow{F}(D) = \{TF_0, TF_1, \cdots, TF_{d-1}\} = \{0, 0, \cdots, 0\}$;
for $i = 1; i <= n; i++$ do
    $\overrightarrow{wf}_i = KeywordVectorGen(w_i, T)$;
    get the frequency $tf_i$ of $w_i$ in the document;
    for $j = 1; j <= L; j++$ do
        $g_j(\overrightarrow{wf}_i) = \{h_{j1}(\overrightarrow{wf}_i), h_{j2}(\overrightarrow{wf}_i), \cdots, h_{jl}(\overrightarrow{wf}_i)\}$;
        $b_{ji} = H(g_j(\overrightarrow{wf}_i)) = [a_1 \cdot h_{j1}(\overrightarrow{wf}_i) + \cdots + a_l \cdot h_{jl}(\overrightarrow{wf}_i)] \bmod d$
        $TF_{b_{ji}} = TF_{b_{ji}} + tf_i$;
    end for
end for
return $\overrightarrow{F}(D)$;
```

Fig. 5. Algorithm of index generation



of the mapped bucket. Finally, the algorithm outputs the document index as $\overrightarrow{F}(D) = (TF_0, TF_1, \cdots, TF_{d-1})$.

The generation of document index indicates that the process does not rely on any predefined global library and that each document is individually indexed. Hence, any data update, including insert, delete and modify, only involves the document to be updated and will not affect any other documents, which means that our scheme can support data updating efficiently.

VII. SECURITY-ENHANCED SCKS THEME

In the adjustive model, the resister is allowed to submit queries adaptively, i.e., submitting subsequent question once receiving the outcomes of previous queries. Thus, the ad-versary will decide subsequent question relying upon the previous outcomes. However, the SkNN cryptography has been tried vulnerable underneath linear analysis [47]. once the server obtains d question vectors and therefore the corresponding trapdoors, it will enforce linear analysis to recover the index of documents. Hence, SCKS is vulnerable underneath the adaptive model. during this section, we tend to propose a security-enhanced SCKS (SE-SCKS) theme that's secure underneath the adjustive model.

Constructions of SE-SCKS

Inspired by the strategy mentioned in [2], we tend to introduce a pseudo-random operate within the generation of document indexes and trapdoors to boost the protection of SCKS. In SE-SCKS, most processes ar an equivalent as SCKS aside from the subsequent steps.

1. Generating a hash key pool. within the KeyGen algorithmic program, additionally to the key key sk = , generate a hash key pool HK = {keyi|keyi ← θ, one ≤ i ≤ l × L} with another given parameter θ, wherever L is that the range of LSH families, and l is that the range of LSH functions in a very family.
2. Introducing the pseudo-random operate. In algorithmic program three, select a pseudo-random operate f : ∗ × θ → ∗ additionally to the L freelance p-stable LSH operate families. Then, the LSH functions employed in the algorithmic program ar replaced by new functions

$$G = \{hf_i | hf_i = f_{key_i} \circ h_i, h_i \in LSH, 1 \leq i \leq l \times L\}.$$

Because the pseudo-random operate is powerfully collision-resistant, mistreatment this operate won't have an effect on the search results. In follow, HMAC-SHA1 [48] are often used because the pseudo-random operate as a result of its collision rate is very low.

Security Analysis of SE-SCKS

The security definition within the adjustive model is comparable to the one within the non-adaptive model except that the resister is allowed to settle on the history adaptively. Specifically, the resister at first submits the document assortment and receives the corresponding index before he chooses the primary question. Then, he can receive the question's trapdoor before he chooses subsequent query, and so on. Intuitively, the resisterin the adjustive model is in a position to perform additional subtle attacks than within the

CONCLUSION

Focusing on the keyword search over encrypted cloud information, we tend to propose a semantic-based compound keyword search (SCKS) theme during this paper. To accurately extract the linguistics info of keywords, we tend to 1st propose associate degree ontology-based compound thought linguistics similarity calculation technique (CCSS), that greatly improves the accuracy of similarity activity between compound ideas by comprehensively considering the compound options and a range of data sources in metaphysics. Then, the SCKS theme is built by integration CCSS with LSH and SkNN. additionally to a semantic-based keyword search, SCKS can do multi-keyword search and stratified keyword search at an equivalent time. as a result of every document is indexed separately, the update of 1 document won't have an effect on alternative documents, which suggests that SCKS will support dynamic information with efficiency. to boost the protection of SCKS, we tend to propose a security-enhanced SCKS (SE-SCKS) by introducing a pseudo-random operate. Thorough security

Time value of SCKS. (a) Secure index generation for one document with totally different range of keywords. (b) Secure index generation for different-sized datasets. (c) Trapdoor generation with totally different range of keywords. (d) look for totally different keywords in a very dataset containing a thousand documents. (e) look for five keywords in different-sized datasets analysis of each SCKS and SE-SCKS is given, and therefore the experiments on real-world dataset demonstrate that the projected approaches introduce low overhead on computation which the search accuracy outperforms the prevailing schemes.

References

[1] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted information," in IEEE conference 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 information within the cloud," in IEEE International Conference on laptop Communications, 2014, pp. 2112–2120.

[3] M. Kuzu, M. S. Islam, and M. Kantarcioglu, "Efficient similarity search over encrypted information," in IEEE twenty eighth International Conference on information 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 information," in 2012 Proceedings of IEEE INFOCOM, 2012, pp. 451–459.

[5] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword stratified search over encrypted cloud information," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 222–233, 2014.