ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Secure Compound Keyword Search over Encrypted Data in Cloud using Enhanced Semantic Scheme

Nitin B. Pawar [1], Prof. Dinesh D. Patil [2]

P.G. Student, Department of Computer Science and Engg., HSMSSGB, COET, Bhusawal[MH], India[1]

Associate Professor & Head, Department of Computer Science and Engg., HSMSSGB, COET, Bhusawal[MH], India[2]

ABSTRACT: In cloud the stored data made secure by encrypting the data but cloud can derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted. Therefore, the security requirement for confidentiality of the encrypted data, user's query record and access patterns of data in cloud. Keyword based searching is one of the data access pattern, used for retrieving the encrypted data from cloud servers. Keyword based searching over encrypted data is necessary for accessing the outsourced sensitive data in cloud computing. In some situations, the keywords searches by user are only semantically related to the data rather than via an exact or fuzzy match. Hence, semantic-based keyword search over encrypted data in cloud becomes of most important. Even so, existing schemes dependent on global dictionary, which affects the accuracy of search results and efficiency of data updating. To deal with these limitations, we initially propose a compound concept semantic similarity (CCSS) calculation method to measure the semantic similarity between compound concepts. Next, by integrating CCSS, Locality-Sensitive Hashing (LSH) function and the preprocessing provided to secure *k*-Nearest Neighbor (SkNN) with secured bit decomposition and euclidean distance function, a enhanced semantic-based compound keyword search (ESCKS) scheme is proposed. ESCKS achieves not only semantic-based search but also multi-keyword search and ranked keyword search. Additionally, ESCKS also eliminates the predefined global library and can efficiently support data update. The experimental results on real-world dataset indicate that ESCKS introduces low overhead on computation and the search accuracy outperforms the existing schemes.

KEYWORDS: Keyword based searching, Searchable encryption, Semantic-based keyword search.

I. INTRODUCTION

Cloud storage services enable users to remotely access data in a cloud anytime and anywhere, in a pay-as-you-go manner using any device. One of the most popular services of cloud computing is data outsourcing. On the basis of cost and convenience, both public and private organizations can now outsource their large amounts of data to the cloud and enjoy the benefits of remote storage [6]. Un-encrypted data outsourcing to cloud by the owner is not much secure because the chances of leak the information from server to cyberpunks. Therefore, to preserve privacy, Cloud needs to protect a user's record when the record is a part of a data mining process [9], [10]. Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted [6]. Therefore, the security requirements problem on a cloud are threefold: (1) confidentiality of the encrypted data, (2) confidentiality of a user's query record, and (3) hiding data access patterns. Thus, the traditional keyword search cannot be directly executed on the encrypted data, which limits the utilization of data. To address this problem, proposed the idea of searchable encryption (SE) [1] that allows users to search on encrypted data through a keyword. Afterwards, various searchable encryption schemes were proposed to meet different requirements of keyword based searching such as fuzzy keyword search, multi-keyword search, ranked keyword search and semantic-based keyword search.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

In general, semantic-based keyword search is useful for users and expresses exactly user intentions [1]. In some situations, users might not be familiar with encrypted documents of cloud or might only want the semantically related results; due to this, the search keywords are usually semantically related to the document rather than via an exact or fuzzy match. For example, the predefined keyword of a document is "encrypted cloud data storage", and the keyword that a user searches is "distributed data storage". Obviously, these two words are neither an exact nor a fuzzy match, but they are semantically related. Therefore, the semantic-based keyword search is of practical importance. But at the same time, semantic similarity between keywords in a query and keywords of documents is important because it also determines the accuracy of search results. However, in existing approaches, the semantic information used to measure the semantic similarity was mined from some knowledge bases (KBs) containing noise data, which cause the semantic similarity to be inaccurate and for compound concepts composed of multiple terms, the approaches usually neglect the special constituent features of the compounds and only process them as single words, which also affects compounds' similarity accuracy [1], [4], [5].

To overcome all issues with privacy of documents and user queries, we proposed, Enhanced semantic-based compound keyword search (ESCKS) scheme. ESCKS uses a topic set in a field and Vector Space Model (VSM) to express the semantic information of keywords. The keywords and field topics can be compound concepts, we initially propose an ontology-based compound concept semantic similarity (CCSS) calculation method to measu***re*** their semantic similarity [1]. Locality-Sensitive Hashing (LSH) function is able to hash similar items to the same bucket with high probability. Hence, we construct the document index by using LSH to map multiple keyword vectors into only one vector. The query vector is generated similarly, which indicates ESCKS can support multi-keyword search. For preserving privacy of documents and queries, we adopt a secure k-Nearest Neighbor (SkNN) [19] to encrypt the index and query [1], [6]. By introducing enhanced secure k-Nearest Neighbor (SkNN) with SSED and SBD function, we improve the security of ESCKS.

1.1 Our contributions

The main contributions of our work are summarized as follows:
(1) An ontology-based compound concept semantic similarity (CCSS) calculation method is proposed to improve the accuracy of compound similarity measurement. By calculating the semantic similarity between keywords and field topics using CCSS, the semantic characteristics are introduced into the keyword vector.
(2) We enhanced S*k*NN by adding pre-processing steps of secure squared euclidean distance (SSED) and secure bit-decomposition (SBD) to protect confidentiality of the encrypted data, user's query record, and hiding data access patterns. By integrating enhanced SkNN with CCSS, LSH, we propose a enhanced semantic-based compound keyword search scheme (ESCKS) over encrypted data. Benefiting from these techniques, ESCKS can simultaneously support semantic based keyword search, multi-keyword search, ranked keyword search and efficient data update.
(3) We improve the security of ESCKS by introducing secure bit-decomposition and euclidean distance function preprocessed steps in SkNN algorithm.
(4) We implement and test CCSS and ESCKS on a real-world dataset. The experimental results demonstrate that our approaches are accurate and efficient.

In our proposed system we focus on solving the classification problem over encrypted data to protect privacy of the documents, user's queries, and hides access patterns of data. In particular, we proposed a secure k-NN classifier over encrypted data in the Cloud. We also focus on by enhancing SkNN algorithm to proposed secure enhanced semantic-based compound keyword search (ESCKS) scheme to increase the accuracy of search result and efficient data updating. To the best of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted data, and then integrate with CCSS and LSH to propose secured ESCKS.

The rest of paper is organized as follows; Section 2 summarizes the related work on classification problem, searchable encryption and similarity measures. Section 3 presents preliminaries used in this paper. In Section 4, we propose our work on enhanced SkNN with preprocessing steps, ontology-based compound concept semantic similarity (CCSS)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

calculation method. Based on CCSS and SkNN, describe our enhanced semantic-based compound keyword search (ESCKS) scheme. In Section 5 evaluates our approaches through experiments and final section concludes the paper.

II. RELATED WORK

Traditional searchable encryption has mostly studied in the context of cryptography [6], [23], [24].The most of works, are focused on improvements of efficiency and formalizations of security definition. Initially, In documents construction of searchable encryption each word is encrypted independently using two-layered encryption construction. Index table entry consists of the trapdoor of a keyword and file identifiers encrypted set of whose relative data files contain the keyword [9], [37].

The K-nearest neighbors algorithm (KNN) [6], in pattern recognition, is a non-parametric method used for classification and regression. The input consists of the k closest training examples in both cases of the feature space. The output depends on whether KNN is used for classification or regression: In KNN classification, the output is a class membership [11], [12]. Through, majority vote of its neighbors an object is classified and being assigned to the class most common among its *k* nearest neighbors. Wildcard-based Fuzzy Set Construction (WFSC) [13], [14]**,** this scheme to enable fuzzy keyword search over encrypted cloud data. The drawback of this system is that with the increase in the edit distance value, the search quality is improved but there is huge increase in modified keyword set. Comparatively better storage usage of DFSC than WFSC. In DFSC unrelated keywords can still be pulled into the modified key set from the dictionary as edit distance value increases, due to this DFSC inherited the search quality degradation problem and it continues to decreases the search coverage of DFSC, it makes DFSC less accurate than WFSC. The previously investigated schemes has major weakness is, they are unable to consider user real intent of search. To effective data search and overcome the issue of confidentiality of encrypted data and users queries, Abirami Swetha.L [6] proposed protocol is to develop a privacy preserving k-NN classifier over encrypted data under the semi –honest model. Specifically, focus on the classification problem only. A semantic-based keyword search scheme returns results according to the semantic relatedness between documents and a query. Based on the co-occurrence probability of terms, Sun et al. [31] and Xia et al. [32] respectively constructed a semantic relationship library (SRL) to record the semantic similarity between keywords. In the search phase, the query keywords are expanded based upon the SRL, and the extended query keywords are used in the search algorithm.

Fu et al. [33] extended the query keywords by introducing *m*-best tree and term similarity tree, both of which are built based on the WordNet. Similarly, the keyword set in is extended via a synonym thesaurus. In the conceptual graphs based search scheme [17], some sentences are extracted to represent the documents and the semantic search is enforced by calculating the relevance score between the sentences in the document and the query. In general, the schemes aiming at multi-keyword search can also achieve ranked keyword search. Orencik et al. [34] utilized the MinHash function and Term Frequency-Inverse Document Frequency (TF-IDF) to construct a document index. However, the TF-IDF should be recalculated when documents or keywords are added or removed, which causes a data updating difficulty. Yu et al. [28] proposed a two round search scheme employing homomorphic encryption to support multi-keyword search. This scheme also adopts TF-IDF; thus, it cannot efficiently support data updating. In the scheme proposed by C. Wang [18] and N. Cao [35], each document is associated with a binary vector of which the dimensionality is equal to the number of keywords in the global keyword set. Hence, the global keyword set should be static because its adding or deleting will cause the reconstruction of the document index.

However, all of these methods primarily aim at single concepts. For compound concepts, the methods usually neglect the special constituent features of compounds and only process them as single concepts, which decrease the accuracy of similarity measurement and issue of security of data in cloud remains as a question.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

III. PRELIMINARIES OF PROPOSED WORK

3.1 Keyword Search Model

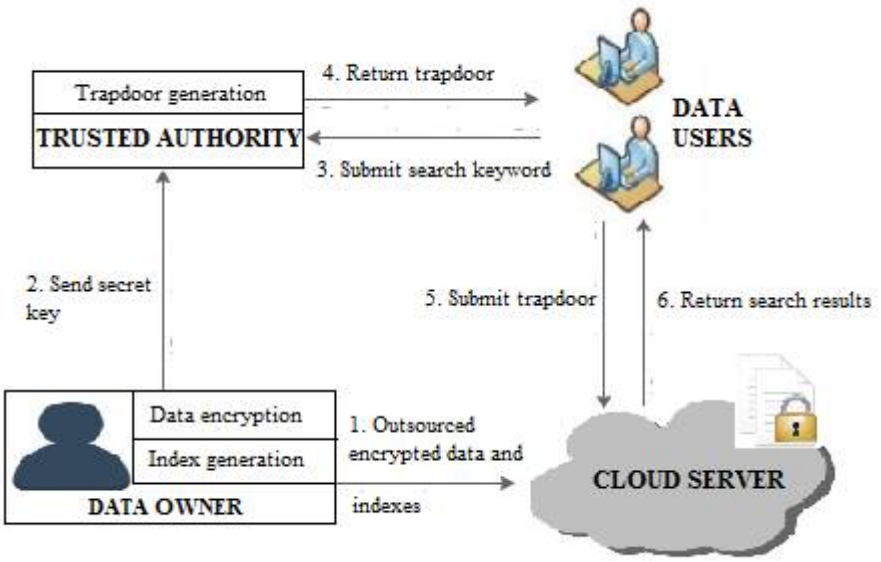


Fig. 1.The model of keyword search

A document index is denoted by a vector generated with the keywords of the document, and a secure index is the encrypted index. A query is a vector generated with the keywords of a search, and a trapdoor is a query which is encrypted. In general, the document can be encrypted by traditional encryption schemes such as AES [1], [7].

Keygen : - Execution is done by the data owner or a trusted authority. Taking a security parameter d as input, it outputs a system symmetric key.
BuildIndex: - It is executed by the data owner. Based on the symmetric key sk and the document D, the algorithm generates the secure index.
Trapdoor: - This is performed by the data owner. With the keyword set Q that the user wants to search, the algorithm generates the corresponding trapdoor.
Search: - It is performed by the cloud server. Based on the trapdoor and each secure index stored in the server, the cloud server determines the correlation coefficients between the query Q and each document D and returns the ranked correlation coefficients to the user.
In Fig.1, The data owner publishes the encrypted documents [1], [2], [7]. He also secures the indexes to the cloud server. To reduce the computation burden, the data owner is allowed to outsource the generation of the trapdoor to TA by giving the private key to it. In this case, whenever a user wants to search over the encrypted documents, he/she submits the keywords to TA which generates the corresponding trapdoor and returns it to the user. Then, the user sends the trapdoor to this cloud server. Finally, the cloud server executes the search algorithm with the trapdoor on all secure indexes and returns the relevant documents to the user. TA is usually an internal server. If there is no such trustable server in the system, the trapdoor can be initiated by the data owner. Also, some existing schemes assume that the authorization between the data owner and users is carried out effectively.

3.2. Locality-Sensitive Hashing (LSH)

A method for efficient approximate nearest neighbor search algorithms, is provided by one of the main applications of LSH. Consider an LSH family. The algorithm has two main parameters are shown in Fig.2: the width parameter *k* and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

the number of hash tables *L* [1], [3], [7]. The algorithm then constructs *L* hash tables, each corresponding to a different randomly chosen hash function *g*. In the preprocessing step we hash all *n* points from the data set *S* into each of the *L* hash tables. Given that the resulting hash tables have only *n* non-zero entries, one can reduce the amount of memory used per each hash table to using standard hash functions [4], [5], [6], [23], [29].

Algorithm steps of LSH

Given a query point q, the algorithm iterates over the L hash functions g. For each g considered, it retrieves the data points that are hashed into the same bucket as q. The process is stopped as soon as a point within distance cR from q is found.

Given the parameters k and L, the algorithm has the following performance guarantees:

- preprocessing time: $O(nLkt)$, where t is the time to evaluate a function $h \in \mathcal{F}$ on an input point p;
- space: $O(nL)$, plus the space for storing data points;
- query time: $O(L(kt + dnP_2^k))$;
- the algorithm succeeds in finding a point within distance cR from q (if there exists a point within distance R) with probability at least $1 - (1 - P_1^k)^L$;

For a fixed approximation ratio $c = 1 + \epsilon$ and probabilities P_1 and P_2, one can set $k = \dfrac{\log n}{\log 1/P_2}$ and $L = n^\rho$, where $\rho = \dfrac{\log P_1}{\log P_2}$. Then one obtains the following performance guarantees:

- preprocessing time: $O(n^{1+\rho}kt)$;
- space: $O(n^{1+\rho})$, plus the space for storing data points;
- query time: $O(n^\rho(kt + d))$;

Fig.2. Algorithm for LSH

IV. PROPOSED WORK

4.1 Enhanced SkNN algorithm

We proposed enhanced secure kNN (SkNN) algorithm by adding two preprocessing steps SSED and SBD [6] before kNN algorithm as below,

Secure squared Euclidean distance (SSED) - In this protocol, P1 with input (Epk(X), Epk(Y))and P2 with sk securely compute the encryption of squared Euclidean distance between vectors X and Y. Here X And Y are m dimensional vectors whereEpk(X)=(Epk(x1),….,Epk(xm))andEpk(Y)=(Epk(y1),…,Epk(ym)).The output Epk(|X-Y|2) will be known only to P1.

Secure bit-decomposition (SBD) - Here P1 with input Epk (z) and P2 securely compute the encryptions of the individual bits of z, where 0 <=z < 2l. The output [z]= (Epk (z1),….,Epk (Zl) is known only to P1. Herez1 and zl are the most and least significant bits of integer z, respectively.

SkNN Classification Algorithm - SkNN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. Here we take input from SSED and then SBD of preprocessing steps to kNN to made it enhanced SkNN. The kNN algorithm is among the simplest of all machine learning algorithms. The neighbors are taken from a set of objects for which the class (for kNN classification) or the object property value (for kNN regression) is known and steps are.

STEP 1: BEGIN

STEP 2: Input: D = {(x1, c1), . . . , (xN , cN)}

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

STEP 3: x = (x1. . . xn) new instance to be classified
STEP 4: FOR each labelled instance (xi, ci) calculate d (xi, x)
STEP 5: Order d (xi , x) from lowest to highest, (i = 1. . . N)
STEP 6: Select the K nearest instances to x: Dkx
STEP 7: Assign to x the most frequent class in Dkx
STEP 8: END.

4.2 Compound concept semantic similarity calculation method

On the basis of their semantic constituents, the compound concepts can be divided into two types: endocentric structure and exocentric structure. The endocentric structure is true when one or more constituents of a compound can play the central role and serve as a definable subject heading. To measure the semantic similarity of compound concepts, we propose a novel approach that considers the concept constituent features and several other factors influencing similarity. For a compound concept, the subject heading generally has greater effect than the auxiliary word because the former expresses the major meaning, whereas the latter is primarily used for modifying the former [1], [6].

For the approaches basing on ontology paths, all possible taxonomical links (i.e., paths) between concepts are calculated, but only the shortest one is kept. To improve the accuracy, other available taxonomical features should be considered. Moreover, the local density of semantic nets affects the similarity between concepts. Although Li et al. [30] considered the local density; they computed it from a corpus that has a problem with data sparseness. Therefore, a corpus-independent local density method is needed. Because all concepts in the ontology are connected by links, we also describe the effect of path length according to [36]. Concepts at different layers of the hierarchy have different similarities. Therefore, the depth of concept in the ontology should be considered. Finally, we introduce a similarity computational approach to support compound concepts, i.e., CCSS.

4.3 Enhanced semantic-based compound keyword search (ESCKS) scheme

In ESCKS scheme, each document contains more than one keyword. If the document index was generated directly using all of the keywords' vectors, it should be expressed as a matrix with high dimensions, such as to improve efficiency, LSH is introduced in the generation of the document index to map multiple keyword vectors into only one vector, i.e., the document index. The value of each element in the index is the sum of keyword frequency appearing in the document. To preserve privacy and enhancing existing SCKS [1], the document index is encrypted with S*k*NN by adding preprocessing steps of SSED and SBD, which is finally outsourced to the cloud. Benefiting from the features of enhanced S*k*NN, ESCKS is able to achieve ranked keyword search and can return the *Top k* results. The trapdoor is generated similarly, except that each element of the query vector is 0 or 1 rather than the frequency of the keywords. LSH is able to map multiple keyword vectors in a query to one vector; for this reason, ESCKS also can support multi-keyword search.

Generating semantic-based keyword vector - Since each element of the keyword vector corresponds to a topic in a specific field, the topics should be determined in advance by the field experts or generated by the Latent Dirichlet Allocation (LDA) [38] method. Suppose the number of topics in a field is *R* and the topic set is $Y = \{Y1, Y2, \cdots, YR\}$. The generation process of the keyword vector is shown in Fig. 3. For a keyword *w*, the algorithm *KeywordVectorGen* first initializes an *R*-dimension vector $\rightarrow wf$ and sets the value of each element to 0. Next, it calculates the semantic similarity *si* between *w* and each field topic *Yi* using CCSS. Finally, the semantic-based keyword vector is output as $\rightarrow wf = (s1, s2, \cdots, sR)$ [1], [15], [33].

Algorithm for *KeywordVectorGen* -

Input: a keyword w and a set of field topics $\Upsilon = \{\Upsilon_1, \Upsilon_2, \cdots, \Upsilon_R\}$;
Output: the semantic vector $\overrightarrow{wf}$ of keyword w;
1: initialize the semantic vector $\overrightarrow{wf} = (0, 0, \cdots, 0)_R$;
2: **for** $i = 1; i <= R; i++$ **do**
3: $s_i = Sim_{CCSS}(w, \Upsilon_i)$
4: **end for**
5: set $\overrightarrow{wf} = (s_1, s_2, \cdots, s_R)$;
6: **return** $\overrightarrow{wf}$;

Fig. 3.Algorithm of keyword vector generation

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Generating document index - After the semantic-based vectors for keywords in the document are generated, the LSH function is introduced to map the keyword vectors to a group of buckets. As shown in Fig.4, First we initialize a group of buckets and number them from 0 to $d-1$, where d is the total number of buckets. Then, each keyword vector is mapped using the functions in an LSH family, which produces a set of hash values. Next, a universal hash function is used to map the set of LSH values to a bucket number. Based on the characteristics of LSH, the semantically related keywords will be mapped to the same bucket with a high probability. Finally, the document index, which is also a vector, is generated by the group of buckets, and the value of each element is the frequency sum of keywords mapped to that bucket.

Algorithm for *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;
1: initialize a group of buckets and number them from 0 to $d-1$;
2: initialize the index $\overrightarrow{F}(D) = \{TF_0, TF_1, \cdots, TF_{d-1}\} = \{0, 0, \cdots, 0\}$;
3: **for** $i = 1; i <= n; i++$ **do**
4: $\overrightarrow{wf_i} = KeywordVectorGen(w_i, \Upsilon)$;
5: get the frequency tf_i of w_i in the document;
6: **for** $j = 1; j <= L; j++$ **do**
7: $g_j(\overrightarrow{wf_i}) = \{h_{j1}(\overrightarrow{wf_i}), h_{j2}(\overrightarrow{wf_i}), \cdots, h_{jl}(\overrightarrow{wf_i})\}$;
8: $b_{ji} = H(g_j(\overrightarrow{wf_i})) = [a_1 \cdot h_{j1}(\overrightarrow{wf_i}) + \cdots + a_l \cdot h_{jl}(\overrightarrow{wf_i})] \bmod d$
9: $TF_{b_{ji}} = TF_{b_{ji}} + tf_i$;
10: **end for**
11: **end for**
12: **return** $\overrightarrow{F}(D)$;

Fig. 4.Algorithm of index generation

Generating query vector - The generation of the query vector is similar to the process for the document index. Suppose the search keyword set is $Q = \{qw1, qw2, \cdots, qwm\}$, where m is the number of keywords in the query. Q can be considered the keyword set of a virtual document QW. Then, the query can be generated by Algorithm 3, except the value assignment for each bucket (i.e., Line 9); i.e., in the generation of a query vector, if the value of bucket bji is 0, it is set to 1; otherwise, no substitution is made. Hence, the values in a query vector are only 0 or 1. Finally, we obtain the query vector $\overrightarrow{-}Q = (qTF0, qTF1, \cdots, qTFd-1)$.

ESCKS scheme - Finally ESCKS scheme, to preserve the privacy of data and users, both the index and query are encrypted with enhanced SkNN. The ESCKS scheme includes mainly 4 algorithms are, KeyGen (d), BuildIndex (sk,D), Trapdoor (sk,Q), Search (I(D), T (Q))

4.4 Security Enhanced – ESCKS

The adversary is allowed to submit queries adaptively, i.e., submitting the next query after receiving the outcomes of previous queries. Thus, the adversary can decide the next query depending upon the previous outcomes. However, the Enhanced S*k*NN encryption has been proved vulnerable under linear analysis [1], [6]. When the server obtains d query vectors and the corresponding trapdoors, it can enforce linear analysis to recover the index of documents. Hence, enhanced SkNN directly improves security of ESCKS is vulnerable under the adaptive model.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

V. PERFORMANCE EVALUATION

5.1 Experiments and analysis of CCSS

5.1.1 Dataset

In this experiment, we used the dataset is the abstract set of 3700 documents that contains more than 100 noun concepts, with a subset of minimum 150 compound concepts. The following experiments are performed based on this dataset.

5.1.2 Experimental results and analysis

We evaluate the performance of our method through experiments. The experiment compares CCSS with some other approaches against human assessments. There are 4 approaches in the experiment, including Li [30], Pirro [39], existing approach CCSS of SCKS [1] and our proposed approach CCSS of ESCKS. The parameters for the other approaches are chosen according to the best correlation values reported in the authors' experiments. Then we compute the Pearson correlation coefficient for each approach. As shown in Fig.5 compared with other approaches, CCSS of both SCKS and ESCKS provides the highest accuracy because it synthesizes the concept constituent features, taxonomical features, local density, path length and depth.

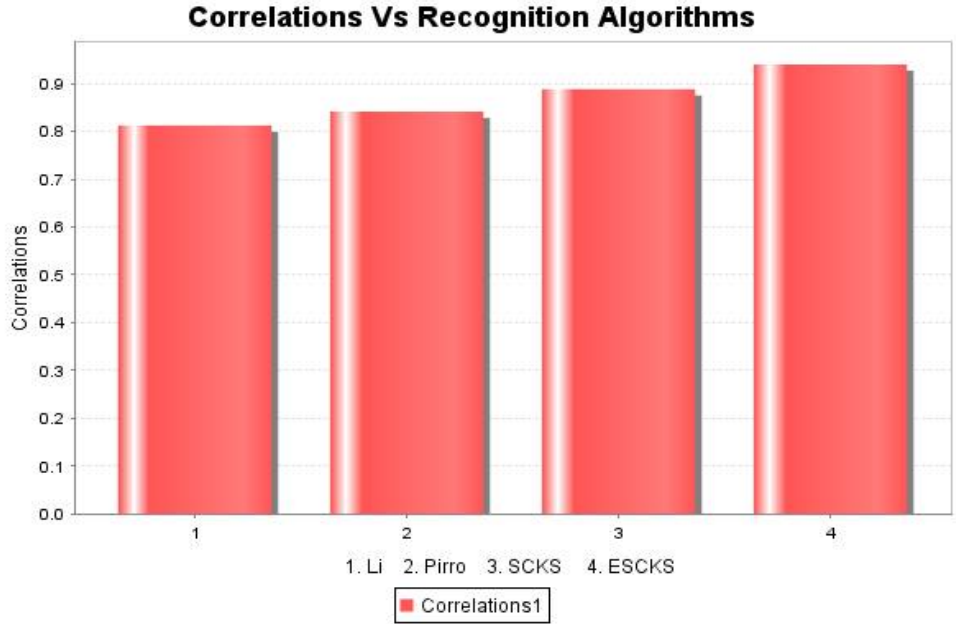


Fig.5. Pearson correlation coefficient of each approach against human assessments.

5.2 Experiments and analysis of ESCKS

5.2.1 Indicators and dataset

In this experiment, we compare ESCKS with the existing SCKS, which achieves search over encrypted data based on LSH and CCSS methods. The comparison focuses on the following three factors.

• Accuracy, denoted as *P*. In this factor, the order of the outputs is discarded and the weight of each result is the same.
• Recall rate, denoted as *R*.
• Average accuracy, denoted as *AP*, which indicates the average accuracy at many recall rates.

In the following experiments, the parameters of SCKS and ESCKS are set as *l*, *L*, *d* and *n*, where *l* and *L* are the parameters in LSH functions, *d* is the dimensionality of document index, and *n* is the dimensionality of keyword vector. From the parameters, we can determine that the index dimensionality SCKS and ESCKS, which causes the index

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

generation and keyword search of SCKS and ESCKS to be more efficient. The dataset in the experiments is the abstract set of 3700 documents.

5.2.2 Accuracy and recall rate comparison

5.2.2.1 Exact keyword search

For exact search, as shown in Table 1, the accuracy P, average accuracy AP and recall rate R of ESCKS are all greater search results Because ESCKS uses CCSS to measure the similarity between keywords like SCKS, which greatly improves the search accuracy of the compound concepts with the use of SkNN, enhanced security to users data and hide data access patterns.

Table 1
Experimental results of exact keyword search

Parameter	**SCKS**	**ESCKS**
Recall(*R*)	0.88	0.94
Accuracy(*P*)	0.76	0.82
Avg. Accuracy(*AP*)	0.53	0.63

5.2.2.2 Semantic-based keyword search

The semantic correlation is determined by a correlation threshold i.e., a document is considered semantically related to a query when the correlation coefficient between them equals or is greater than the threshold. The experimental results of semantic search are shown in Table 2 and indicate that the accuracy *P*, average accuracy *AP* and recall rate *R* of ESCKS are all improved and outperforms with existing scheme.

Table 2
Experimental results of Semantic-based keyword search

Threshold	Accuracy		Recall		Average Accuracy	
	SCKS	ESCKS	SCKS	ESCKS	SCKS	ESCKS
0.5	0.786	0.79	0.98	0.93	0.545	0.59
0.6	0.7818	0.86	0.903	0.94	0.538	0.56
0.7	0.7852	0.82	0.902	0.93	0.543	0.6

5.2.3 Resultant graph for generating secure index

Following graphs shows ESCKS scheme compare with existing scheme on the basis of time required for generating secure index with number of keywords in document and number of document in dataset

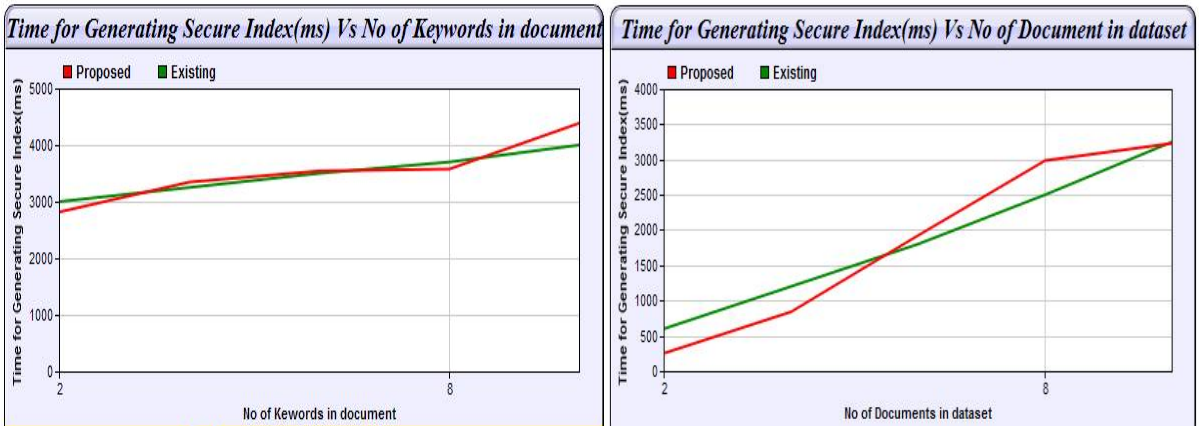
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019



(a) (b)

Fig.6. Generating secure index graphs (a) Time for generating secure index Vs number of keywords in document. (b) Time for generating secure index Vs number of documents in dataset

5.2.4 Result analysis and discussion

In exact keyword search and semantic-based search, the accuracy *P*, average accuracy *AP* and recall rate *R* of ESCKS are improved. But in existing semantic-based search schemes need to predefine a global library and the accuracy, average accuracy and recall rate all depend on the quality of the library. However, the libraries and datasets used in existing schemes are unavailable. Hence, results compare with using our dataset. Compared with the schemes focusing on accuracy of search result, and ESCKS is much more secure because we improved SkNN algorithm with SSED and SBD function.

VI. CONCLUSION

The focus of work on the secure keyword search over encrypted cloud data, we proposed a enhanced semantic-based compound keyword search (ESCKS) scheme in this paper. To accurately extract the semantic information of keywords, we first proposed an ontology-based compound concept semantic similarity calculation method (CCSS), which greatly improves the accuracy of similarity measurement between compound concepts by comprehensively considering the compound features and a variety of information sources in ontology. Then, the ESCKS scheme is constructed by integrating CCSS with LSH and enhanced S*k*NN by adding SDB and Euclidean distance function in preprocessing steps. Our work mainly focus on enhancing S*k*NN with SDB and Euclidean distance function, to protect the privacy of documents, user queries and hide data access patterns in cloud and we achieve it by avoiding classification problem. By enhanced S*k*NN, we improve the security of ESCKS. In addition to a semantic-based keyword search, ESCKS supports multi-keyword search and ranked keyword search at the same time. Because each document is indexed individually, the update of one document will not affect other documents, which means that ESCKS can support dynamic data efficiently. The experiments on real-world dataset demonstrate that the proposed approaches introduce low overhead on computation and search accuracy outperforms the existing schemes.

VII. ACKNOWLEDGEMENT

I would like to thank our honorable principal sir, Dr. R. P. Singh and my special thanks to my guide and head of the department, Prof. Dinesh. D. Patil Sir & sincere thanks to all the respected teaching faculties of Department of Computer Science & Engineering of Hindi Seva Mandal's, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal. My special gratitude to all the authors of reference paper that are referred by me.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

REFERENCES

[1] Bo Lang, (*Member, IEEE*), Jinmiao Wang, Ming Li, and Yanxi Liu, "Semantic-based Compound Keyword Search over Encrypted Cloud Data", IEEE TRANSACTIONS ON SERVICES COMPUTING, Vol. No.10, Page No. 1 – 14,2018.
[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati,"Managing and accessing data in the cloud: Privacy risks and approaches," in Proc. 7th Int. Conf. Risk Security Internet Syst., 2012, pp. 1–9.
[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage," in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148.
[4] M. Batet, D. S´anchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine." *Journal of Biomedical Informatics*, vol. 44, no. 44, pp. 118–25, 2011.
[5] 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.
[6] Abirami Swetha.L,Gokila.R,Mr.Vijaya Ragavan.P, "EFFECTIVE DATA SEARCH FOR ENCRYPTED RELATIONAL DATA IN CLOUD USING K-NEAREST NEIGHBOR ALGORITHM", International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), Volume 20, Issue 3, 2016.
[7] Harsh Gupta, Kartik Ahirrao, Noopur Sonaje, S.N. More, "Compound Keyword Search of Encrypted Cloud Data by using Semantic Scheme", International Research Journal of Engineering and Technology (IRJET), Vol. No.05,Issue 12, 2018.
[8] Avani Konda, Sai Praneeth Gudimetla, Balaji T, Gopi Krishna Subramanyam, Usha Kiruthika, "Synonymous Keyword Search Over Encrypted Data in Cloud", International Journal of MC Square Scientific Research Vol.9, No.2, 2017.
[9] Jin Li, Jingwei Li, Xiaofeng Chen, Chunfu Jia, Wenjing Lou (Senior Member, IEEE), "Identity-Based Encryption with Outsourced Revocation in Cloud Computing", IEEE TRANSACTIONS ON COMPUTERS, Vol. No.64, Paper no.02, Page No.425-437, 2015.
[10] Mikhail Strizhov, Indrajit Ray, "Multi-keyword Similarity Search Over Encrypted Cloud Data", IFIP International Information Security Conference, pp 52-65, 2014.
[11] Y. Elmehdwi, B. K. Samanthula, W. Jiang, "Secure k-nearest neighbor query over encrypted data in outsourced environments", *2014 IEEE 30th International Conference on Data Engineering*, pp. 664–675, 2014.
[12] B. Yao, F. Li, X. Xiao, "Secure nearest neighbor revisited", IEEE 29th International Conference on Data Engineering (ICDE), pp. 733–744, 2013.
[13] Neelam S. Khan, Dr. C. Ramakrishna, "A Survey on Secure Ranked Keyword Search over Outsourced Encrypted Cloud Data", International Conference on Computer Networks and Information Technology (ICCNIT), 2014
[14] B. Wang, S. Yu, W. Lou, Y. T. Hou, "Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud", *IEEE International Conference on Computer Communications*, pp. 2112–2120, 2014.
[15] M. Kuzu, M. S. Islam, M. Kantarcioglu, "Efficient similarity search over encrypted data", IEEE 28th International Conference on Data Engineering (ICDE), pp. 1156–1167, 2012.
[16] R. Li, Z. Xu, W. Kang, K. C. Yow, and C. Z. Xu, "Efficient multikeyword ranked query over encrypted data in cloud computing", *Future Generation Computer Systems*, vol. 30, no. 1, pp. 179–190,2014.
[17] Z. Fu, F. Huang, X. Sun, A. Vasilakos, C. N. Yang, "Enabling semantic search based on conceptual graphs over encrypted outsource data", *IEEE Transactions on Services Computing*, vol. PP,no. 99, pp. 1–1, 2016.
[18] C. Wang, N. Cao, J. Li, K. Ren, W. Lou, "Secure ranked keyword search over encrypted cloud data", *IEEE 30th International Conference on Distributed Computing Systems (ICDCS)*, pp.253–262,2010.
[19] W. K. Wong, D. W.-l. Cheung, B. Kao, N. Mamoulis, "Secure kNN computation on encrypted databases", Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 139–152, 2009.
[20] H. Hu, J. Xu, C. Ren, B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism", Proc. IEEE 27th Int. Conf. Data Eng., pp. 601–612, 2011.
[21] D. S´anchez and M. Batet, "A semantic similarity method based on information content exploiting multiple ontologies," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1393–1399, 2013.
[22] S. Kamara, C. Papamanthou, and T. Roeder, "Dynamic searchable symmetric encryption," in *ACM Conference on Computer and Communications Security*, 2012, pp. 965–976.
[23] S. Kamara and P. Charalampos, "Parallel and dynamic searchable symmetric encryption," in *Financial Cryptography and Data Security*, 2013, pp. 258–274.
[24] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.
[25] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web." in *Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 803–812.
[26] B. Hore, S. Mehrotra,M. Canim, and M. Kantarcioglu, "Secure multidimensional range queries over outsourced data,"VLDB J., vol. 21, no. 3, pp. 333–358, 2012.
[27] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Inf. Syst., vol. 29, no. 4,pp. 343–364, 2004.
[28] J. Yu, P. Lu, Y. Zhu, G. Xue, and M. Li, "Toward secure multikeyword top-k retrieval over encrypted cloud data," IEEE Transactions on Dependable and Secure Computing, vol. 10, no. 4, pp. 239–250,2013.
[29] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2004, pp. 563–574.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

[30] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge & Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.
[31] X. Sun, Y. Zhu, Z. Xia, and L. Chen, "Privacy- preserving keyword based semantic search over encrypted cloud data," International Journal of Security & Its Applications, vol. 8, no. 3, pp. 9–20, 2014.
[32] Z. Xia, Y. Zhu, X. Sun, and L. Chen, "Secure semantic expansion based search over encrypted cloud data supporting similarity ranking," J. Cloud Comput., vol. 3, no. 1, pp. 8:1–8:11, 2014.
[33] Z. Fu, J. Shu, X. Sun, and N. Linge, "Smart cloud search services: verifiable keyword-based semantic search over encrypted cloud data," IEEE Transactions on Consumer Electronics, vol. 60, no. 4, pp. 762–770, Nov 2014.
[34] 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.
[35] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 222–233, 2014.
[36] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," Wordnet: An Electronic Lexical Database, vol. 49, no. 2, pp. 265–283, 1998.
[37] Nitin B. Pawar, Dinesh D. Patil, "A Review on Correspondence Techniques for Effective Keyword Search over Encrypted Data in Cloud", in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 7, Issue 3,2019, pg.2026-2031.
[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003
[39] G. Pirr´o, "A semantic similarity metric combining features and intrinsic information content," *IEEE Transactions on Data & Knowledge Engineering*, vol. 68, no. 11, pp. 1289–1308, 2009.