A Survey on Different Search Techniques Over Encrypted Data in Cloud

Amrithasree Haridas and L. Preethi

1 Introduction

Cloud computing provides the facility to the variety of applications operating over thousands of computers and servers to concurrently access the services through Internet. With the evolvement of cloud computing it has become easier for users to store, retrieve, and share their data among themselves. It offers various benefits to users as well as to service providers. It provides flexibility to work from anywhere at any time. The most extensively adopted application of cloud computing is cloud storage. A tremendous amount of information is being stored by users on cloud servers every day. This information needs protection from different kinds of cyber threats. To maintain data confidentiality and secure storage, various types of encryption algorithms are used for protecting information from unauthorized disclosure. However, searching over encrypted data was difficult to attain. Therefore, keyword based searching has been introduced where the desired file is retrieved when searched for a particular keyword. Numerous searchable encryption schemes are existing in the literature. The common factor in these existing method is that the user data must be encrypted (generating trapdoor) before sending to the cloud server. Upon receiving the search query the cloud server searches in the encrypted document (represented by an encrypted index) and returns the search result to the users. However it must be ensured that while performing the search it should not cause any information leakage. This survey makes the comparative study of some recent single/multi-keyword search techniques on large-scale encrypted data in cloud.

A. Haridas (✉) · L. Preethi
College of Engineering Trivandrum, Thiruvananthapuram, India
e-mail: amrithasreeharidas@gmail.com; preethi@cet.ac.in

M. Palesi et al. (eds.), *Second International Conference on Networks and Advances in Computational Technologies*, Transactions on Computational Science and Computational Intelligence, https://doi.org/10.1007/978-3-030-49500-8_11

1.1 Cloud Security Requirements

The infrastructure of cloud must be capable enough to implement the appropriate security measures at its premises. Although the services provided by the cloud are regularly being improved, still there is a great need for protection of data stored in the cloud. For this, the most important requirement is to build up trust between the user and service provider.

To protect cloud data, following security measures need to be implemented:

- **Authentication**: This technique helps the communicating entities to prove its identity and assures authentic communication This service also guarantees that no other unauthorized entity can masquerade itself as authorized entity to take undue advantage of ongoing communication.
- **Access control**: Authentication and identification of entity must be carried out to give access rights to the entity. It is the process of imposing the restriction to access systems and applications according to the level of security requirements.
- **Confidentiality**: The attacker is not allowed to look at frequency, length, and other attributes of traffic flowing through the network. Unauthorized exposure of information must be protected to maintain the confidentiality of sensitive cloud data.
- **Integrity**: The received data must be free from duplication, modification, and reordering. Only authorized users can make changes to it. This service assures the correctness and validity of data being transmitted through the network.
- **Availability**: To maintain it offsite backup should be done regularly, and the systems must be prevented by Denial of Service attacks. This service assures that information is available to authorized users whenever required.
- **Non-Repudiation**: This service provides the proof that the authorized sender and receiver have sent and received the information, respectively. For this, accurate and traceable records must be maintained.

1.2 Architecture of Search Over Cloud Data

Figure 1 demonstrates the architecture of the system. The different modules of the system are data user, data owner, and the cloud server. Data stored in the cloud is (Fig. 2) in encrypted format for the purpose of security. Only authenticated users can use the data in the cloud and hence security is improved. Data owner is the person who uses the cloud for storage. Also he can share the data with other peoples. Data users are the person who can access the data uploaded by data owners.

Searchable encryption permits clients to correctly retrieve the encrypted information. This method has two major disadvantages; First one is, that users who do not essentially have pre-knowledge of the encrypted cloud information have to post-process each file obtained after search, in order, to realize the one most matching

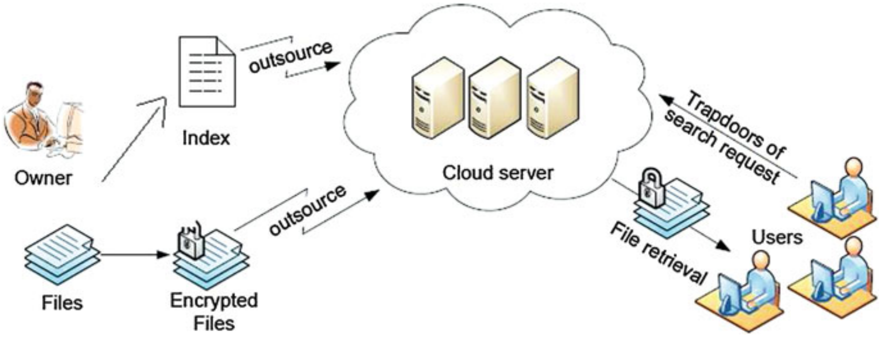

Fig. 1 Architecture of the system [1]

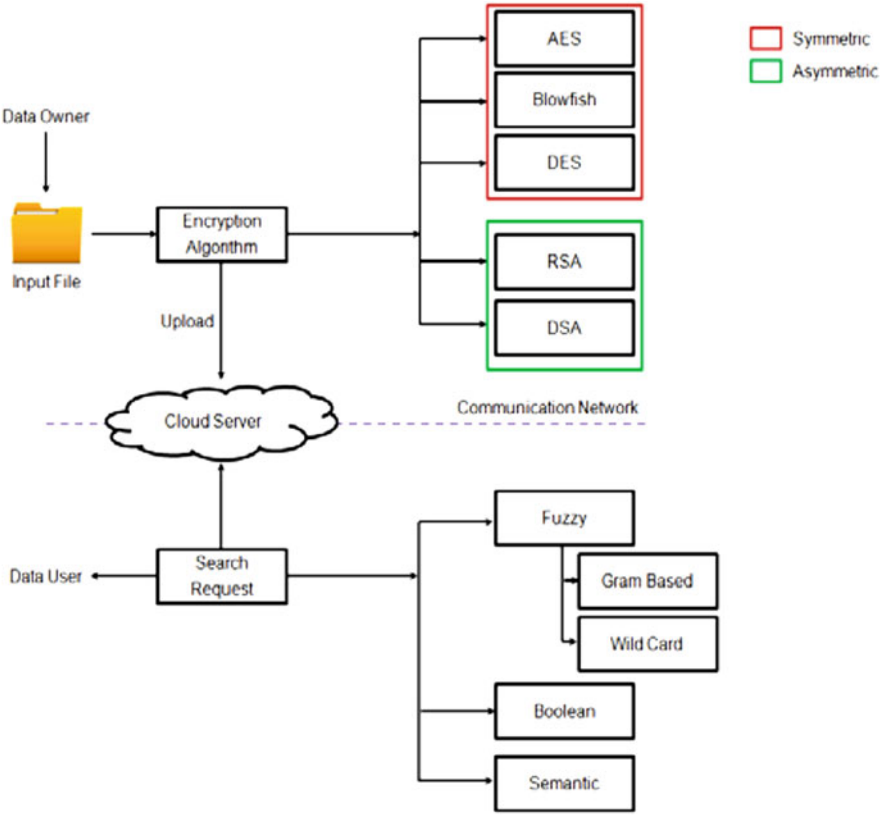

Fig. 2 Search architecture in cloud data [2]

their interest; second disadvantage is when multiple records containing the queried keyword are retrieved in the search this causes unnecessary network traffic.

2 Types of Keyword Based Searching in Encrypted Cloud Data

Song et al. introduced the concept where each word of the file is encrypted separately. But this technique resulted in higher cost as the word by word scanning of the documents is required. They suggested a sequential scan which could be executed with or without an index. When the documents in the dataset are large, then the index based scheme is preferred since it gives faster search results. But this system causes trouble in the situation where storage and updating of records are needed.

3 Searchable Encryption

With searchable encryption technique user can store and share their data in encrypted format to improve the security of data, and other users can search in this encrypted secure data. Figure 3 shows different types of searchable encryption.

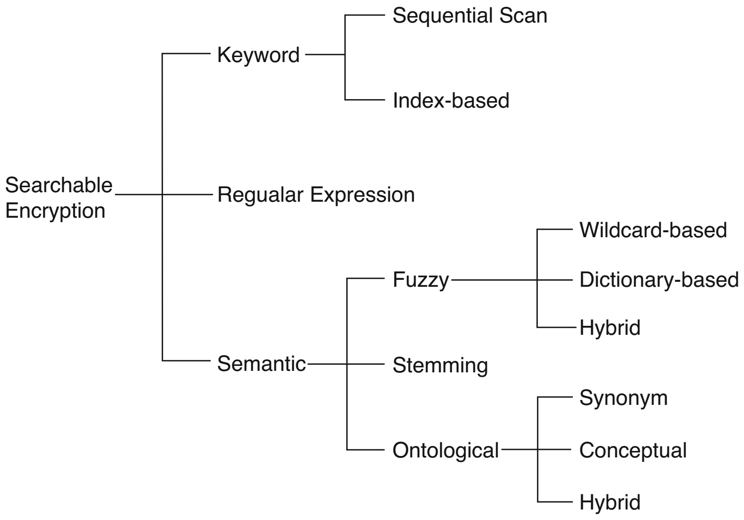

Fig. 3 Classification of the current searchable encryption system

Various encryption techniques can be used to achieve this, such as AES, ECC, etc. This can work on large-scale data.

3.1 Keyword Based Approaches

Ranked Single Keyword Search

C. Wang [3] proposed an efficient solution for supporting ranked keyword search problems. In this technique, single random keyword is the input to the cloud server and the cloud server generates the most related file that matches with the input keyword. Ranked keyword search generates the results rank wise instead of just providing matched results. It will reduce the cost of searching and also provide the most related results to improve the user experience.

Data owner outsources their n data files to the cloud. Before outsourcing the files to the cloud server, they encrypt their file for the security purpose. Even though they are encrypted they can also retain their ability to search for effective data usage. Before outsourcing their information, create an index i by using a set of k distinct keywords extracted from the file collection D, and store both the index I and the encrypted file collection D on the cloud server. After this, on receiving search query T_w from data user, the server is actually responsible for finding the index i and provides results without revealing the actual content of sensitive data. For data user, to search for a keyword w authorization is performed with trapdoor generation T_w and submitted to the cloud server.

Multi-Keyword Search

To make searching system more practical, system can support multi keyword search in place of single keyword search. Zhihua Xia et al. [4] proposed a scheme, where input search query may contain more than one keyword, this improves the accuracy of search query. Multiple keywords have the capability to explain the search query accurately.

It also supports dynamic functions such as inclusion and removal. In particular, the widely used TF X IDF model and the vector space model have been integrated into index building and generation of queries. It specifies an index structure based on tree structure and suggests “Greedy Depth-First Search” algorithm for ranked multi-keyword search. And for encrypting query vectors it uses a secure kNN algorithm, then accurately calculates the correct score between the encrypted query vectors and the index.

Ranked Multi-Keyword Search

Ranked keyword search generated the results rank wise instead of just providing matched results. Cloud server then generates the most related file that matched with input keywords. In this technique, multiple random keywords are input to cloud server. N. Cao [5] tried to improve the previous works in this field by adding the ranking functionality together with search over encrypted data in cloud having multi-keywords and thus bettering the user search experience.

Here is used a privacy-preserving, similarity based text retrieval scheme where the search results are hidden from the unauthorized entities. Also, the server is unable to reconstruct the term composition of documents and queries performed. They employed similarity measure of "coordinate matching" organized as multi-keyword semantics. It uses "inner product similarity" to quantitatively evaluate the similarity measure. But there were two shortcomings of this scheme. First of all, it requires the reorganization of static dictionary each time with the entry of a new keyword. As the size of the collection of records grows exponentially the time for search also increases exponentially. The system is MRSE based on secure inner product computation. But in this technique the synonym searching was not taken into consideration. And therefore the searching time was increased intensely.

Fuzzy Keyword Search

To make the search process more user interactive, this were introduced over encrypted data in cloud. Initially, the concept of search using fuzzy keyword in encrypted cloud data was given by J. Li [6]. It states that the search system used in this method can give the accurate result even if the keyword is slightly misspelled by the user. This technique attempted to make the search procedure user interactive. On the other hand, in traditional techniques, no result is found when there are minor errors in spelling of keywords entered, and hence it makes the user's task very complicated. To handle this problem, J. Li [6] implemented fuzzy keyword searching. It also focused on preserving the privacy of keywords. If user spell incorrectly then integrate edit distance with wildcard-based technique to build fuzzy keyword sets, to address minor misspelling issues and format inconsistency and by using this method it calculate the closest matching keyword. To diminish the difficulty in storage and to handle the issues in representation, they developed keyword dictionary. They demonstrated that their work was proficient in maintaining the privacy and security employing detailed security analysis. It also showed the utility of this technique.

A Conjunctive Keyword Search

C. Wang [7] proposed a method where, a query request has multiple keywords and for each keyword trapdoor is generated. The final result is the intersection of results of each keyword in search request.

To be precise, it is the statistical measure approach, that is, it calculates the relevant scores. It generates secure search results from information and builds up a one-to-many order-preserving mapping techniques to appropriately protect these sensitive score information. Order-preserving encryption (OPE) is a practical method to support fast ranked search. OPE is used to encrypt relevance scores in the inverted index, and so often the data privacy cannot be assured in applications. The OPE has improved by Wang et al. [7], in their secure keyword search scheme to "One-to-many OPE," where they tried to build a probability encryption scheme and hide the distribution of plaintext. The server side rankings will have an effective design without losing data. These methods, however, cause substantial overhead in communicating due to sharing the secret and increasing computing costs due to the bilinear mapping.

3.2 Semantic Based Approaches

MRSE System Supporting Synonym Query

The search results are purely based on user authentication and only authorized user can input the search request. It increases the flexibility of search when the user forgets the exact keyword, and the user can search by using some of its synonyms. Zhangjie Fu [8] suggested a synonym based multi-keyword search system in an encrypted data in cloud. This is the first method suggested based on semantics.

Apart from other methods the main difference in this method is that, here the keyword set is extended by adding some of its synonym also, for that keywords need to be extracted from the file collection before outsourcing it to the cloud. Here uses a better text feature weighting method, which adds a new component module to indicate the distinguishability of the term on the basis of the original TFIDF (term frequency-inverse document frequency) method (term frequency–inverse document frequency) method. The new element C_d has been added to the equation of TFIDF,

$$\text{Weightingfactor} = \text{TF}X\text{IDF}XC_d \tag{1}$$

To accomplish an efficient meaningful search for outsourced data, the keyword set should be extended by adding synonyms. If a keyword has more than 2 synonyms, then all synonyms are added into the keyword set. So using this improved method will extract keywords from outsourced text files. All keywords separated from a single text form a keyword set at the end. The redundant keywords are deleted to reduce the burden of storage.

Semantic Search Using Stemming Algorithm

T. Moataz [9] incorporated a stemming algorithm with an efficient searchable encryption technique. In the case of keyword based approach each keyword is mapped to all the documents having this word. In a semantic search not only the

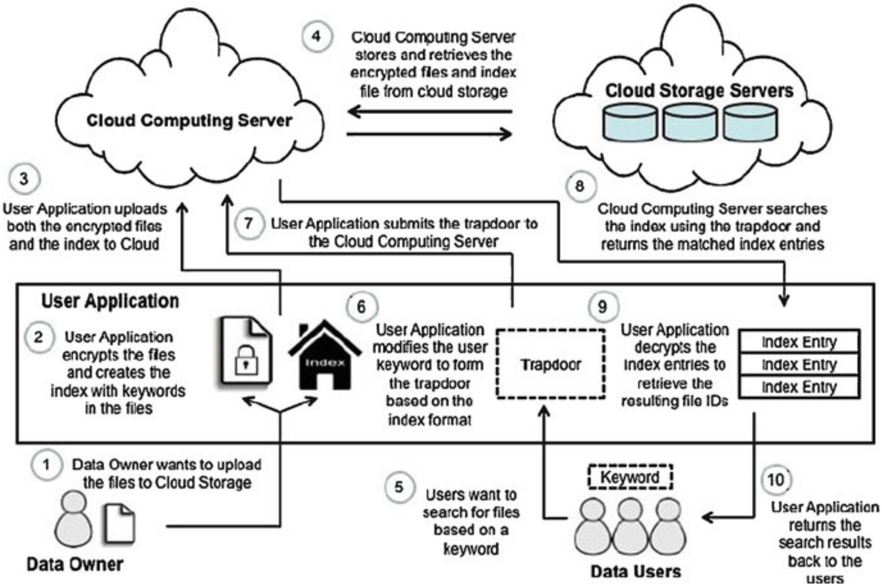
4 Cloud Computing Server stores and retrieves the encrypted files and index file from cloud storage
Cloud Computing Server
Cloud Storage Servers
3 User Application uploads both the encrypted files and the index to Cloud
7 User Application submits the trapdoor to the Cloud Computing Server
8 Cloud Computing Server searches the index using the trapdoor and returns the matched index entries
User Application
2 User Application encrypts the files and creates the index with keywords in the files
Index
6 User Application modifies the user keyword to form the trapdoor based on the index format
Trapdoor
9 User Application decrypts the index entries to retrieve the resulting file IDs
Index Entry
Index Entry
Index Entry
1 Data Owner wants to upload the files to Cloud Storage
Data Owner
5 Users want to search for files based on a keyword
Keyword
Data Users
10 User Application returns the search results back to the users

Fig. 4 An overview of semantic searches using encrypted data in cloud computing

documents containing the keyword itself but also the documents with words related to the keyword have to be mapped. In the search request the user inputs the query and the keywords are extracted from the query. Then the root of this keyword is found and these roots are sent instead of the keywords in the search request. Related keywords have the same root keyword, so that it only stores the root of these keywords in the index. By combining a searchable encryption algorithm with a stemming algorithm an efficient semantic search technique over encrypted cloud data can be achieved. As a result, the user will retrieve all the encrypted documents that contain all related terms. Figure 4 shows this method.

Semantic Search Using Online Ontological Network

Z. Jason Woodworth [10] proposed a semantic search using online ontological network. The basic idea of this methodology is to consider the frequency of occurrence of keywords in document. It ignores the importance of the terms in the text or the request. For that reason, change every instance of a particular word in each document into a similar token, and subsequently apply a similar change when that word shows up in the search query. For normal text recovery, Okapi BM25 algorithm will be used. This is a TF and IDF model. Algorithm is never considering the true meaning of each term in the document. This feature makes this algorithm very applicable in this method. This technique has 2 phases: upload and search.

Upload Process

The purpose of this process is to parse the file to the indexable format and encrypt it before sending it to the cloud. Term frequency of each keywords for the text is the product of TF and IDF collected. For each word it finds its hashed value and writes them to a temporary index file and sends it to the cloud with encrypted documents. A subset of words (usually called keywords) is usually taken from the document to represent the semantics of that document.

Search Process

This process has mainly two components: One is Query modification and the other is index searching and ranking. Query modification is performed on the client side and searching is performed on the cloud processing server. The process of query modification involves splitting, semantic expansion, and weighting. The clients query cannot be semantically identified by this methodology alone. To achieve this it uses more advanced ontological networks. For example, it utilizes the contents from Wikipedia and performs keyword extraction on them to get related words and expressions. These related terms are merged into modified query set Q. As a result, users can retrieve the documents containing ideas related to the query.

3.3 Comparison of Different Searchable Encryption Techniques

There are several methods for performing a search over encrypted data in cloud, using keyword as a factor for searching. Single keyword and multi-keyword searches are possible. To enhance the efficiency and fastness of searching multi-keyword technique can be used. A comparative study on different search algorithms in the cloud is described in Tables 1 and 2.

Table 1 Comparative analysis of various semantic based searching techniques

Methods	Keyword search (Single/Multiple)	Process used
Multi-keyword ranked search supporting synonym query	Multi-keyword	Input the synonyms of the extracted keyword
Semantic search using stemming algorithm	Multi-keyword	Stemming algorithm finds the root of queried keyword in the search request. And the e root of queried keyword is used for search instead of keywords in the request
Semantic search using online ontological network	Multi-keyword	Use online ontological network

Table 2 Comparative analysis of various semantic based searching techniques

Sr.No	Method	Process used	Advantage	Disadvantage
1	Searchable Encryption Scheme	Symmetric Public Key Encryption	Secure search employed over encrypted data on cloud	Costly in terms of computation
2	Boolean Keyword Searchable Encryption Scheme	Structural and Boolean keyword search using Boolean operators AND, OR, and NOT	Comfortable enough to express small, easy information needs	Excess network traffic. Efficient document ranking is not supported
3	Single keyword searchable encryption scheme	Encrypted searchable index	5 keyword frequency utilization to rank results	Not comfortable enough to precise complex information needs
4	Ranked keyword searchable encryption scheme	Relevance score is employed to make a secure searchable index. Order-preserving mapping function	Enhances system usability by returning the matching files in a ranked order concerning to certain relevance criteria. Eliminate excess network traffic	Compromise the privacy
5	Fuzzy keyword searchable encryption scheme	Wildcard-based technique	Eliminates the requirement for enumerating all the fuzzy keywords	Supports only Boolean keyword search. Huge storage complexity
6	Plaintext Fuzzy keyword searchable encryption scheme	Plaintext searching, string matching algorithm	To find relevant information it allows user to search using try and- see approach	Statistics and dictionary attacks and fails to attain the search privacy
7	Conjunctive Keyword Searchable Encryption Scheme	Decisional Diffie–Hellman (DDH) and hardness assumption	Solution to Boolean keyword search problem	Privacy overhead
8	Multi-keyword Searchable Encryption Scheme	Provides secure index structure, generates secret trapdoors	Documents confidentiality and privacy of index, trapdoor, trapdoor unlinkability	CSPs that keep the data for users may access users sensitive information without authorization

4 Performance Analysis

Here compare the time used for searching with different size of dataset in different methods. The first figure (Fig. 5) shows the search time in case of synonym query and the second figure (Fig. 6) shows search in case of synonym query by using ontological network. The final figure (Fig. 7) shows search in case of by using stemming algorithm.

5 Conclusion

This survey makes the comparative study of some recent single/multi-keyword search techniques on large-scale encrypted data in cloud. The traditional methods of keyword searching were limited to the exact keyword search. But recently, many researchers have implemented fuzzy keyword searching in which the encrypted file is retrieved when the keyword matches exactly or when it is slightly misspelled and preserving the privacy of keywords at the same time. Further suggested a synonym based multi-keyword search system in an encrypted cloud data. Then the user can do searching by similar meaning words. Such techniques maintain the privacy and security of data during search. It might be possible that user forgets the exact

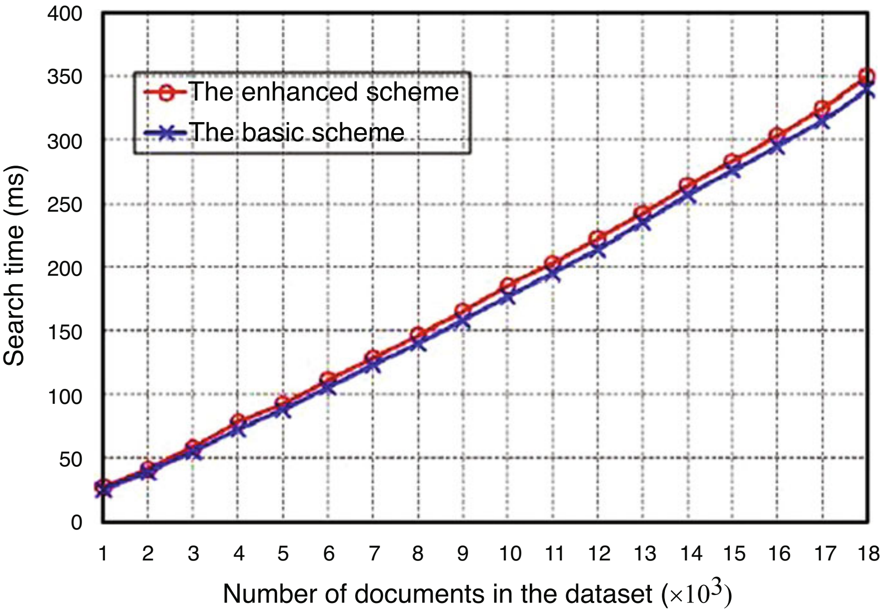

Fig. 5 Time taken to search for MRSE system supporting synonym query [5]

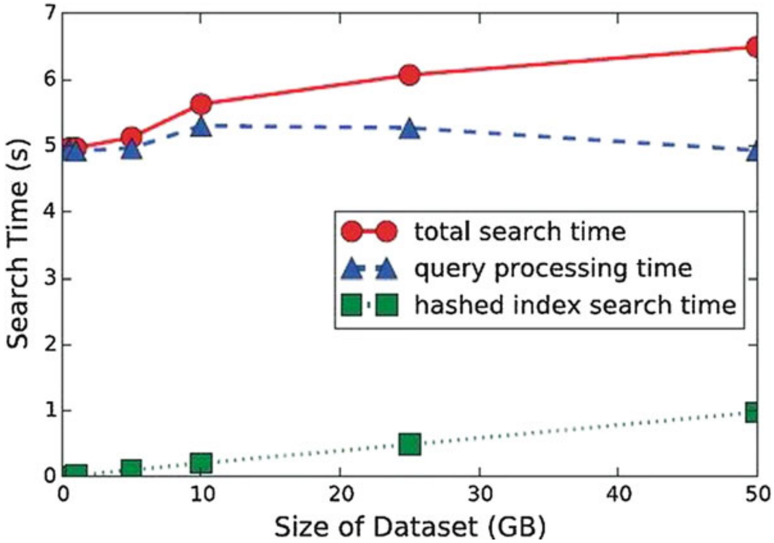

Fig. 6 Time taken to search for semantic search using online ontological network [10]

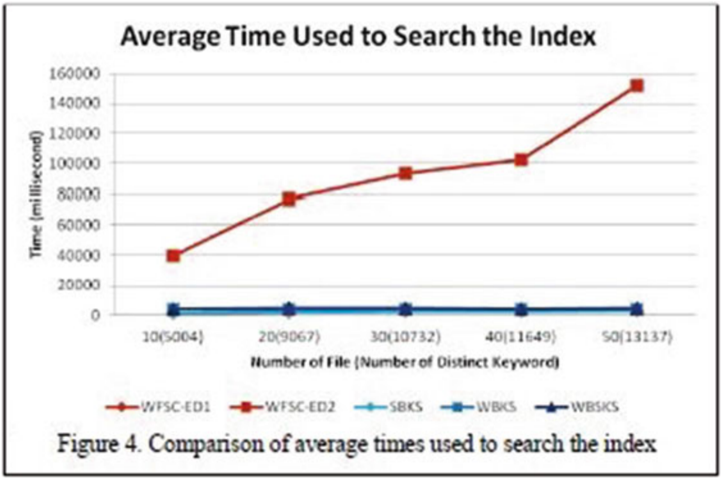

Fig. 7 Time taken to search for semantic search using stemming algorithm [9]

keyword. In this comparative analysis, compare them on the basis of various criteria such as, key idea of approach, their advantages and disadvantage. And this survey also identifies the limitations of existing system.

References

1. Mistry, S., Tandel, P.: A survey on context based search over encrypted cloud data techniques. In: Proceedings of the International Journal of Modern Trends in Engineering and Research (2015)
2. Ingale, S., Phulpagar, B.D.: A survey on different keyword-based search techniques over encrypted data. In: Proceedings of the (IJCSIT) International Journal of Computer Science and Information Technologies (2016)
3. Wang, C., Cao, N., Li, J., Ren, K., Lou, W.: Secure ranked keyword search over encrypted cloud data. In: Proceedings of the IEEE 30th International Conference on Distributed Computing Systems (ICDCS 10) (2010). https://doi.org/10.1109/ICDCS.2010.34
4. Xia, Z., Wang, X., Sun, X., Wang, Q.: A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. IEEE Trans. Paral. Distri. Syst. **27**, 340–352 (2015). https://doi.org/10.1109/TPDS.2015.2401003
5. Cao, N., Wang, C., Li, M., Ren, K., Lou, W.: Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Trans. Paral. Distri. Syst. **25**, 222–233 (2014). https://doi.org/10.1109/TPDS.2013.45
6. Li, J., Wang, Q., Wang, C., Cao, N., Ren, K., Lou, W.: Fuzzy keyword search over encrypted data in cloud computing. In: Proceedings of the IEEE INFOCOM, San Diego, CA (2010). https://doi.org/10.1109/INFCOM.2010.5462196
7. Wang, C., Cao, N., Ren, K., Lou, W.: Enabling secure and efficient ranked keyword search over outsourced cloud data. IEEE Trans. Paral. Distri. Syst. **23**, 1467–1479 (2012). https://doi.org/10.1109/TPDS.2011.282
8. Fu, Z., Sun, X., Xia, Z., Zhou, L., Shu, J.: Multi keyword ranked search supporting synonym query over encrypted data in cloud computing. In: Proceedings of the IEEE 32nd International Performance Computing and Communications Conference (IPCCC2013), San Diego, CA (2013). https://doi.org/10.1109/PCCC.2013.6742783
9. Moataz, T., Shikfa, A., Cuppens-Boulahia, N., Cuppens, F.: Semantic search over encrypted data. In: Proceedings of the 20th International Conference on Telecommunications (ICT) (2013). https://doi.org/10.1109/ICTEL.2013.6632121
10. Woodworth, J.Z., Salehi, M.A., Raghavan, V.: S3C-An Architecture for space-efficient semantic search over encrypted data in the cloud. In: Proceedings of the IEEE International Conference on Big Data (Big Data) (2016). https://doi.org/10.1109/BigData.2016.7841040