International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 4808–4815

Efficient Indexing Method for Multi-Keyword Ranked Search for Encrypted Cloud Data Using E-TFIDF

[1]D. Pradeepa, [2]Dr. P. Sumathi.
[1]Research Scholar, PG & Research Department of Computer Science, Government Arts College, Coimbatore INDIA.
[2]Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore INDIA.

Abstract - In Cloud Storage, data owners may also share their outsourced data with a large number of users who may need to best retrieve positive, reliable data documents that they are inquisitive about during a given session. One of the most common ways to do this is by looking for keywords. Such a keyword search method helps users to selectively retrieve documents of interest. Similar works on searchable encryption are primarily focused on the frequency of the occurrence of the keyword in the study. Documents that contain extra keyword-related records with less frequency of the prevalence of the keyword in the report may be available. In this paper, implement a method that recognizes such documents at the same time as returning the documents to the applicant's level. To maintain the security of the data, outsourced data is typically stored in a cloud server encryption form, making it incredibly difficult to search for unique encrypted documents that match those keywords on the cloud server for customers. To resolve this problem, use the Commonplace Enhanced-Term Frequency—Inverse Document Frequency (E-TFIDF) rule in this paper to measure the relevance scores of documents that fit the search request. The weighting of the results of the proposed E-TFIDF obtained the weight value significance of each article from the highest to the lowest weight. The overall success of the proposed approach is the product of the use of precision and recall.

Keywords: - Retrieval, Cloud Storage, Index, Keyword, Cloud User, Cloud Owner, files, document.

1 INTRODUCTION

Cloud computing is emerging as a promising pattern for outsourcing data and an amazing data offering. As cloud computing is becoming more widespread, a growing amount of confidential data is being outsourced to the cloud, consisting of emails, private fitness statistics, company finance records, and government documents, and so on. Since data owners and cloud servers no longer rely on the same domain, our outsourced unencrypted statistics could be an opportunity. However, the fact that data owners and cloud servers are not in the same trusted domain can additionally put outsourced data at risk, as the cloud server may also not be fully relied on in any such domain cloud environment because of some of Motives; the cloud server may additionally leak data to unauthorized entities or be hacked.

It follows that sensitive information usually need to be encrypted before outsourcing to privacy records and countering unsolicited access. However, the encryption of records makes the use of high-performance statistics a very difficult project since a large number of outsourced data documents may exist. [7].

In order to satisfy practical search criteria, the following three features should be assisted by the search for encrypted documents. Second, searchable encryption schemes need to help search for multi-keywords; searching for single-keywords is far from desirable by simply returning very restricted and deceptive search effects. Second, in order to quickly choose the maximum available results, the quest user

ISSN: 2233-7857 IJFGCN

will usually decide on cloud servers to retrieve search results in a relevance-based order [14] ranked by the relevance of the hunting request to the documents. Showing a ranked search to customers can also discard useless network site visitors by sending the most appropriate cloud-based effects back to looking customers. Third, as far as search efficiency is concerned, since the variety of files in the database can be fairly broad, searchable encryption schemes should be green in order to respond quickly to hunting requests with minimal delays.

The standard approach to keep records confidential is to encrypt specific information. On the other hand, this is one of the very difficult processes for information use. Search techniques primarily according to the ciphertext can certification statistical privacy, but search algorithms are time-consuming and space-consuming, so it is difficult for cloud statistics retrieval [9]. To resolve this difficulty, researchers have projected a set of searchable encryption techniques according to the presumption of cryptography. These encryption techniques moreover no longer have high-accuracy retrieval results that are worth a bunch of moment in time and space in the clouds. Its miles therefore make it possible to propose an effective and functional search technique.

In this paper, introduce an improved space vector model called E-TFIDF. The proposed E-TFIDF method constructs a keyword list and converts documents and keywords into "factors" in a multidimensional space that can be represented by vectors.

The paper relaxation is structured as follows. In Section 2, the evaluation of the different methods relevant to the statement of difficulties used in the proposed scheme. Explaining the stairs for the proposed E-TFIDF technique to compile a record for retrieval in section 3. In part 4, the performance assessment of the proposed work is defined in comparison with the current strategies. Finally, the planned study is outlined in section 5.

2. RELATED WORK

Data pre-processing is the method for inserting a brand new document into a record retrieval system. Document pre-processing is a complicated technique that results in the representation of each report using a collection set of index words. Various methods are used to pre-process the file for retrieval.

Strzalkowski, T (1995) proposed Data retrieval machine in which superior natural language processing techniques are used to improve the efficacy of term-based record retrieval. The backbone of our device is a conventional statistical engine that creates inverted index files from pre-processed documents, then searches and ranks documents in response to user queries. Natural language processing is used to (a) pre-order documents if you want to extract content-sporting words (b) find inter-term dependencies and create a specific database-space conceptual hierarchy, and (c) apply the natural language of the person to powerful search queries.

Handa, R., Rama Krishna, C., & Aggarwal, N. (2018) proposed with the advent of cloud storage, data owners tend to outsource their sensitive data to the cloud because they have storage facilities at a reduced cost of administration and security. As these personal data are withdrawn from the user's premises, the protection of statistics becomes a major concern, as cloud service provider (CSP) and stop users are uniquely in agreement with domain names. In order to provide data protection, it is preferred that the discontinued user should encrypt the data prior to outsourcing to the cloud. Encryption presents security, however, makes data usage a difficult task, i.e. it is difficult to scan for encrypted files. Various schemes exist in the literature, but both are limited to unmarried key-word searches or are inefficient in terms of the time needed for searches. The authors propose an effective method of safe data retrieval using the principle of bucketization. This limits the amount of comparisons that are consistent with sub-linear problems.

ISSN: 2233-7857 IJFGCN

Pimpalkar, A. P., & Raj, R. J. R. (2020) proposed Record review and associated projects have recently emerged as important areas of observation. Nowadays, the problem of researchers is that there is a huge amount of data generated every minute and second, as people constantly share their minds, reviews some of the items that are related to them. Social media data, however, remains unstructured, disseminated, and difficult to work with, and want to create a solid base so that they can be used as useful data on a particular issue. Processing such unstructured data in this space in phrases such as noise, co-relevance, emoticons, folklore, and slang is pretty tough and thus requires proper pre-processing records before having the right feelings. Dataset is extracted from Kaggle and Twitter, pre-processed using NLTK and Scikit-analyze, and option and extraction functions are performed for Bag of Words (BOW), Term Frequency (TF) and Inverse Document Frequency (IDF) schemes. It is also emphasised in this research that the choice and illustration of features in the sense of various pre-processing techniques have an advantageous impact on the output of the class.

Thangavel, M., et al. (2015) describes about the security in a cloud garage as a vital feature of cloud storage. This work offers a new multi-degree encryption technique to ensure cloud protection. The Modified-RSA (M-RSA) key technology algorithm is applied in this scheme to produce segment keys. Initially file is encrypted via the public key portion of the MRSA. For convenient communication, a Modified Vernam Cipher (MVC) is added to the file and stored within the cloud. Ahead apply for; the document shall be received with the help of the use of converse MVC accompanied by MRSA decryption using phase personal keys. This scheme was investigated in one of the ways of document entry. Proposed algorithms secure the system from a gross pressure attack. In particular, the computational time of the M-RSA is much less than the standard RSA collection of laws. Frequency and spectral analysis show that MVC is efficient and powerful.

Fatmawati, T., Zaman, B., & Werdiningsih, I. (2017) introduce Working to develop generation, the technique of finding data with textual content is simple, since the textual content of the news is not only accessible in print media, along with newspapers, but also in electronic media that can be accessed using the quest engine. A term is frequently used as a query in the process for finding the related documents on the search engine. The number of phrases that make up the issue of the phrase and its role has an impact on the significance of the file generated. As a consequence, the quality of the data collected will be impaired. On the basis of the above issue, the purpose for this study is to investigate the implementation of the popular phrase index system for data retrieval. The framework is built with pre-processing, indexing, measurement of the term weighting and calculation of the cosine similarity. The device will then show the file search effects in a sequence, mainly based on the cosine similarity.

3. THE PROPOSED SCHEME

Increase output by extracting a specific collection of words for use in text indexing from the resulting document. If full-text material illustration is implemented, all words are used for indexing. Indexing is an important technique: the ability of the user to locate documents on a unique issue is restricted by the indexing mechanism that has produced index words for this situation [11].

ISSN: 2233-7857 IJFGCN

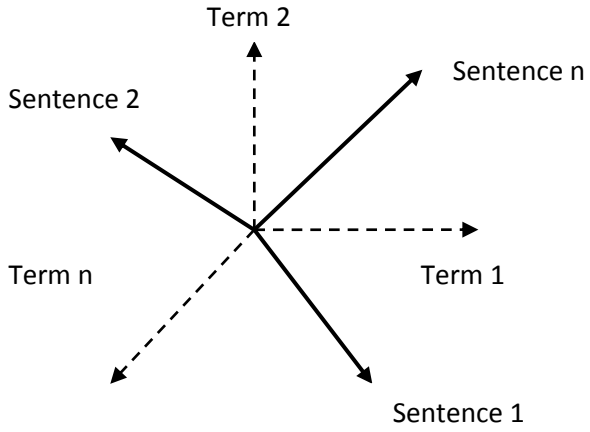

Figure 1: - Vector Space Model (VSM)

Figure 1 displays the vector space model; each phrase/period is the axis/measuring. The text/document is interpreted as a vector within a multi-dimensional space. The number of specific phrases is analogous to a variety of measurements.

TF-IDF is a statistical degree that measures the importance of a word to a text in a group of documents. This is achieved by multiplying two metrics: how usually a word appears in a document, and the inverse document frequency of a word through a collection of files. TF-IDF (Term Frequency-Inverse Document Frequency) is developed for document search and data retrieval [10]. It performs by increasing the amount of times a word occurs in a record; however, it is balanced by the number of files that contain a word.

Each file within the collection to be searched is denoted in an equal approach. Sometimes stem is employed to apprehend variations of the same expression, in order that a huge variety of listed phrases can be in addition decreased. Words are normally weighted in terms of their context in terms. A now not unusual weighting approach is to assign disproportionate weights to terms that seem frequently in a document just from time to time somewhere else.

The similarity between and document stored inside the system and the non-public question is defined by the distinction between the document vector and the question vector. Documents are typically ranked by means of their proximity to the query. This is referred to as the ranking of relevance.

3.1. Pre-processing

Pre-processing is the most important activity of statistical retrieval to retrieve vital records from unstructured text statistics. But the traditional pre-processing strategy is no longer continuously producing top-satisfactory consequences due to the success of the named entity.

The purpose behind pre-processing is to symbolise each file as a function vector, i.e. to divide textual contents into man or woman phrases. Textual content documents are modelled as transactions in the proposed classifiers [12]. The necessary process for text indexing is nothing but to choose the keyword that is the character selection method. This step is vital in finding out the essential for the building of the next level, that is, the stage of classification. It is crucial to select the big key phrases that express what it method and to discard the phrases that do not help to differentiate between the documents.

ISSN: 2233-7857 IJFGCN

3.1.1. Document Indexing using E-TFIDF

The main objective of document indexing is to improve productivity by using extracts from the following document to use a selected collection of phrases for indexing the record. Document indexing consists of selecting a precise collection of keywords based on the entire corpus of records, and assigning weights to the one keywords for each individual record, which is why each file is reworked right into a keyword weight vector. Usually, the weight is correlated with the frequency of occurrence of the word in the file and the range of documents that use that term. The proposed work enhances the performance of the TF-IDF and its miles represented as more enhanced term frequency-inverse Document Frequency (E-TFIDF).

3.1.2. Term Weighting (TW) of E-TFIDF

Documents are interpreted as vectors in the vector space model. TW is a solution model that determines the fulfilment or malfunction of the classification machine. While special phrases have specific levels of meaning in textual content, the tw is often used as a key indicator.

The three important components that impact the importance of time in a file are the TF problem, the IDF factor, and the Document Duration Normalization (DDN) [7]. The term frequency of every phrase in the document (TF) is a weight that relies upon at the incidence of every word within the document. It communicates the value of the phrase in the text. The inverse record frequency of each word in the document database (IDF) is a weight that based on the circulation of every sentence in the document database. It communicates the context of every word in the record database [8]. E-TFIDF is a technique that uses both TF and IDF to calculate the weight of a time. The E-TFIDF scheme can be a very commonplace in-text magnificence area, and nearly all different weighting schemes are variants of the scheme.

The document collection is represented as 'DC', a word is defined as 'WO', and an entity document ‘doc’, the weight ‘wt’ is designed by means of Equation 1

$$\boldsymbol{wo_{do} = f_{WO,do} * log\left(\frac{|DO|}{f_{wt,do}}\right)} \qquad \ldots \boldsymbol{equ(1)}$$

Where TF represents the number of occurrences of 'wt' appears in a document ’do’. |DST| utilized here as the volume of the dataset. IDF mentioned as the quantity of documents in which 'wt' appears w, D in D.

Typical term weighting technique is a variation of the IDF and the TF. They're described as:

$$\boldsymbol{IDF = log\frac{number\ of\ documents\ in\ collection}{-number\ of\ documents\ with\ term} + 1} \qquad \ldots\ \boldsymbol{equ(2)}$$

$$\boldsymbol{TF(term, document) = frequency\ of\ a\ term\ in\ document} \quad \ldots \boldsymbol{equ(3)}$$

$$\boldsymbol{WEIGHT(term, document) = TF(term, document) * IDF(term)} \qquad \ldots \boldsymbol{equ\ (4)}$$

The IDF concept is that the less files that have the time, the more useful the term is to differentiate between the files that have it and those that do not have it. On the other hand, if the time usually exists in a study, then it is possible that the word is commonly used to describe the stuffing of the file. The highest weight is given to words that sometimes occur in a record, but now and then elsewhere. This is referred as the weighting method.

The most popular come up to relevance ratings in VSM is to assign each file according to score primarily on the computation of the weights of terms not remarkable to the document and the query. The

ISSN: 2233-7857 IJFGCN

words of the documents typically obtain their weight from the TF*IDF. The similarity among each file and the query is then calculated with the wording:

$$\boldsymbol{similarity}\ (\boldsymbol{Document}, \boldsymbol{query}) = \sum_{\boldsymbol{all\ query\ items}} \boldsymbol{Weight\ of\ a\ term\ in\ query} * \boldsymbol{Weight\ of\ term\ in\ document} \quad \ldots \boldsymbol{equ}\ (\mathbf{5})$$

The end outcome of TF/IDF is a vector with the numerous terms by the side of through their time weight.

Algorithm – E-TFIDF

Agree on TF, compute its subsequent Weight, and store it in
Weight Matrix (WM)
find out IDF
If IDF = zero then
get rid of the word from the Wordlist
get rid of the consequent TF from
the WM
Else
Compute TF/IDF and store normalized
TF/IDF in the consequent factor of the WM.

The new query is formulated by the usage of the chosen words for 2d-round retrieval. As a end result of the query expansion, a few applicable documents overlooked in the preliminary round can then be retrieved to progress the overall performance.

4. PERFORMANCE EVALUATION

The IR machine sends the ranked list of documents back to the consumer's query. The highest commonly used metric is the average degree of recall and accuracy dependent on significance. As far as the problem is concerned, the complete file part can be partitioned into four units: relevant to the individual and retrieved through the system; applicable but not retrieved; irrelevant and retrieved; nearby the point and not retrieved [13]. Recollection and accuracy are distinct on the basis of the above mentioned four units.

$$\boldsymbol{recall} = \frac{\boldsymbol{number\ of\ retrieved\ relevant\ documents}}{\boldsymbol{total\ number\ of\ relevant\ documents}}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{number\ of\ retrieved\ relevant\ documents}}{\boldsymbol{total\ number\ of\ retrieved\ documents}}$$

Recall shows the proportion of all applicable documents recovered since the collected works. Accuracy is termed as precision indicates how much of the recovered files are important. One of the problems with this compute is that the maximum number of documents in produce in the sequence is usually undisclosed.

The F-Measure is another way to combine and not to overlook the accuracy. The simple F-Measure Model is as follows:

ISSN: 2233-7857 IJFGCN

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precison}$$

IR systems usually display a change of mind and precision between bears, such that the more documents that are obtained, the more meaningless extra documents can be covered.

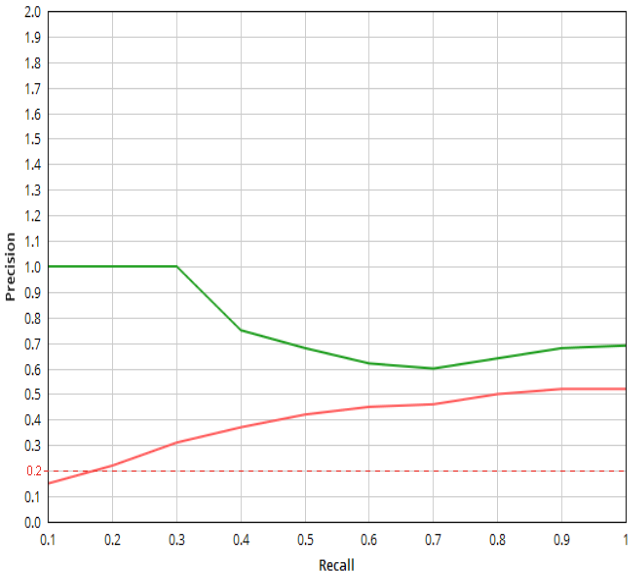

Figure 2: - Precision Vs. Recall

Figure 2 shows that the IR1 unit performs much better than the IR2 system because it has higher precision values at each stage. The redline is the proposed E-TFIDF approach and the novice line is the actual TF-IDF technique.

5. CONCLUSION

Among several data retrieval modes, the VSM is the majority accepted method of relevance computation and adopts the E-TFIDF form for element mining. This form is commonly used to retrieve multiple keywords in plain text. Indexing is the method for ordering files containing comparable keywords. The indexing listing consists of vast quantities of data until the statistical owner imports the data to the cloud. Indexes are easily replaced by a cloud server, enormously. By indexing, the time needed to look at files that maintain keywords is minimized.

6. REFERENCES

[1] Strzalkowski, T. (1995). Natural language data retrieval. Data Processing & Management, 31(3), 397-417.

[2] Handa, R., Rama Krishna, C., & Aggarwal, N. (2018). An efficient approach for secure data retrieval on cloud. Journal of Intelligent & Fuzzy Systems, 34(3), 1345-1353.

[3] Pimpalkar, A. P., & Raj, R. J. R. (2020). Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 9(2), 49-68.

[4] Thangavel, M., Varalakshmi, P., Murrali, M., & Nithya, K. (2015). Secure file storage and retrieval in cloud. International Journal of Data and Computer Security, 7(2-4), 177-195.

ISSN: 2233-7857 IJFGCN

[5] Fatmawati, T., Zaman, B., & Werdiningsih, I. (2017, August). Implementation of the common phrase index method on the phrase query for data retrieval. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020027). AIP Publishing LLC.

[6] Zhang, W., Lin, Y., Xiao, S., Wu, J., & Zhou, S. (2015). Privacy-preserving ranked multi-keyword search for multiple data owners in cloud computing. IEEE Transactions on Computers, 65(5), 1566-1577.

[7] Zhang, W., Xiao, S., Lin, Y., Zhou, T., & Zhou, S. (2014, June). Secure ranked multi-keyword search for multiple data owners in cloud computing. In 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (pp. 276-286). IEEE.

[8] Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2013). Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Transactions on parallel and distributed systems, 25(1), 222-233.

[9] Yang, Y., Lu, H., & Weng, J. (2011, November). Multi-user private keyword search for cloud computing. In 2011 IEEE Third International Conference on Cloud Computing Technology and Science (pp. 264-271). IEEE.

[10] Chuah, M., & Hu, W. (2011, June). Privacy-aware bedtree based solution for fuzzy multi-keyword search over encrypted data. In 2011 31st International Conference on Distributed Computing Systems Workshops (pp. 273-281). IEEE.

[11] Fu, Z., Sun, X., Liu, Q., Zhou, L., & Shu, J. (2015). Achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing. IEICE Transactions on Communications, 98(1), 190-200.

[12] Yang, C., Zhang, W., Xu, J., Xu, J., & Yu, N. (2012, November). A fast privacy-preserving multi-keyword search scheme on cloud data. In 2012 International Conference on Cloud and Service Computing (pp. 104-110). IEEE.

[13] Sun, W., Wang, B., Cao, N., Li, M., Lou, W., Hou, Y. T., & Li, H. (2013, May). Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. In Proceedings of the 8th ACM SIGSAC symposium on Data, computer and communications security (pp. 71-82).

[14] Xia, Z., Wang, X., Sun, X., & Wang, Q. (2015). A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. IEEE transactions on parallel and distributed systems, 27(2), 340-352.

ISSN: 2233-7857 IJFGCN