Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Time-conserving deduplicated data retrieval framework for the cloud computing environment

P. Swathika & J. Raja Sekar

To cite this article: P. Swathika & J. Raja Sekar (2023) Time-conserving deduplicated data retrieval framework for the cloud computing environment, Automatika, 64:4, 681-688, DOI: 10.1080/00051144.2023.2211439

To link to this article: https://doi.org/10.1080/00051144.2023.2211439

AUTOMATIKA
2023, VOL. 64, NO. 4, 681–688
https://doi.org/10.1080/00051144.2023.2211439

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Time-conserving deduplicated data retrieval framework for the cloud computing environment

P. Swathika[a] and J. Raja Sekar[b]

[a]Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, Sivakasi, India; [b]Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India

ABSTRACT

Cloud computing technology is quite inevitable in today's smart world. The excessive utilization of data mandates updated storage space, which is highly expensive and cloud storage is the best solution to it. As charges are levied for the utilized space, data redundancy must be avoided for the effective exploitation of cloud space. Data deduplication is a technique, which removes redundant data and conserves storage, bandwidth and charges. However, data retrieval upon deduplicated data is not well explored in the existing literature. This work attempts to present an effective retrieval framework for deduplicated data in a cloud environment by presenting two protocols namely Data Outsourcing Protocol (DOP) and Data Retrieval Protocol (DRP). The retrieval performance of the proposed approach is tested and compared with the existing approaches in terms of standard performance measures. The work performance of the proposed Deduplicated Data Retrieval (DDR) framework performs better in terms of retrieval precision, recall and time conservation rates. The average precision and recall rates attained by the proposed work are 97.9% and 95.75% respectively.

ARTICLE HISTORY
Received 20 February 2023
Accepted 3 May 2023

KEYWORDS
Cloud; data retrieval; data deduplication; data outsourcing; deduplicated data retrieval

1. Introduction

Cloud computing is often regarded as a potential infrastructure for Information Technology (IT) due to its robust features. It can gather and restructure vast storage, processing, communication and application resources. This enables cloud users to receive IT services in a flexible, pervasive, cost-effective and on-demand manner [1]. Attracted by these exceptional qualities, an increasing number of data owners are outsourcing their local document management systems to the public cloud. However, a concomitant difficulty is how to retain the utility of uploaded data while protecting the privacy of sensitive information [2]. All the data must be encrypted before they can be outsourced; hence, it is of the utmost importance to create adequate mechanisms to perform fundamental operations on the encrypted data collection.

Based on their functionalities, numerous encrypted document retrieval schemes have been proposed, including single-keyword Boolean search schemes [3,4] single keyword-ranked search schemes [5,6] and multi-keyword Boolean search schemes [7–9]. In terms of document retrieval, these approaches fall short of completely satisfying data users. In the actual world, it is fairly typical to utilize a collection of keywords to search for relevant files in a given field. In addition, the retrieved results are needed to be arranged according to their relevance to the specified keywords. Unfortunately, none of the aforementioned strategies can fully satisfy these needs.

Researchers have become more interested in privacy-preserving multi-keyword ranked document search strategies [10,11]. These techniques enable data users to recover encrypted documents based on a list of keywords, and the search operations are comparable to those for plaintext documents from the data users' perspective. In comparison to multi-keyword Boolean search, these schemes are more user-friendly and compatible with their retrieval practices. Nonetheless, these methods can be enhanced in the following ways: first, the majority of existing methods assume that all data users are reliable. This assumption is implausible in practice. In actuality, the cloud server can easily pose itself as a data consumer to obtain the secret keys from the data owner for a very minimal cost. Once the cloud server obtains the secret keys, all encrypted documents may be quickly decoded, which is a devastating blow to existing encryption techniques. This is the most essential reason for developing a realistic and new architecture for safe document retrieval in encrypted cloud file systems.

Second, most present systems are limited to a single document retrieval method, and the search experience for data users can be enhanced. Data users may need

CONTACT P. Swathika pswathika.ai@gmail.com, swathikap@mepcoeng.ac.in Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu 626 005, India

to search a collection of documents using filenames, authors, multiple keywords or any combination thereof. We can intuitively regard filenames and authors as common terms, similar to the majority of existing methods. However, this method may reduce the search's precision. Integration of the multi-keyword Boolean query schemes with the multi-keyword ranked search schemes is another feasible way.

Third, data duplication issue is not given attention. Duplication is a problem for cloud storage since it consumes unnecessary space and bandwidth. When the duplicates are detected and removed, considerable savings can be expected in terms of space, time and computation.

In light of the current scenario, the existing solutions pay more attention towards security and privacy preservation, while data retrieval is not given more importance. Additionally, the existing works support equality queries over encrypted data. This article considers the aforementioned points and presents a Deduplicated Data Retrieval (DDR) framework, which involves data deduplication, storage and retrieval phases. This work presents two protocols such as data outsourcing (DOP) and retrieval protocols (DRP). The DOP deals with operations such as pre-processing, extraction of keywords, formation of index and encrypted index. The DRP functions concerning retrieval such as query formation, detailed query formation, encrypted data retrieval and ranking. The highlights of this work are listed as follows:

- Data retrieval frameworks for deduplicated data are quite scarce in cloud computing environment.
- Inherits the advantages of deduplication and retrieval, such that duplicates are eliminated and retrieval of deduplicated data reduces storage, computational and time complexities.
- Data retrieval precision rates are quite convincing and the time consumption is tolerable.

The remainder of this paper is organized as follows: the related works concerning data retrieval for cloud systems are studied in Section 2. The proposed DDR framework is elaborated in Section 3, while its work efficiency is evaluated in Section 4. Section 5 concludes the article.

2. Review of literature

Data deduplication is a storage optimization technique used to enhance storage efficiency. This section examines common data deduplication algorithms for cloud storage, such as [12–15], identifies factors that affect storage capacity and discusses various data deduplication scenarios. In [16], an efficient ranked retrieval scheme for multiple keywords was presented with privacy protection for cloud computing. This work considered multiple data owners and performed multi-keyword search, while returning the leading "*n*" results to the users without the keywords-related information. Security analysis was made to justify the performance of the work. A privacy-preserving content-based image retrieval scheme for cloud storage was proposed in [17]. Their work focussed on encrypted storage and image retrieval, while ensuring privacy. The retrieval precision rates were analysed and this work claimed that it reduced the time and space complexities as well.

In [18], a privacy-preserving data retrieval model was presented for the cloud-based Internet of Things (IoT). Their work built an implicit index being maintained by the edge servers and the privacy-preserving hierarchical model to preserve privacy was presented. The computational cost of this work was proven to be minimal. In [19], a scheme design for cloud information retrieval was presented. A retrieval risk formula was presented to retrieve keywords from the encrypted cloud data, which could enhance the search experience. A homomorphic encryption-based cipher text retrieval scheme was presented for hybrid cloud environment with multiple data owners in [20]. Here, both the public and private cloud servers coordinated with each other to perform data retrieval. The multi-owner and multi-keyword-based search was presented based on an encrypted binary index tree structure and homomorphic encryption scheme. In [21], a large-scale image retrieval method for cloud computing was presented. The binary signatures of image descriptors were generated by hamming embedding algorithm and a frequency histogram with binary descriptors was formed for feature representation. The image retrieval accuracy rate was enhanced by random sampling and min-hash algorithms.

In [22], a cost-effective data retrieval scheme was presented based on Named Data Networking (NDN) for Vehicular Cloud (VC). In their work, the vehicles utilized unicast model to collect data from the nearest VC through a retrieval process. Additionally, the mobility of vehicles was also supported. They claimed that it reduced the retrieval cost, while increasing the accuracy rates. A cipher text retrieval system was presented in [23] by considering an encrypted heterogeneous database in cloud-based IoT scheme. Their work relied on integration middleware that could support cross-language and cross-platform queries for different databases. The data integration method considered cross-database queries over encrypted data. A secure product information retrieval scheme was presented for cloud computing in [24]. The security of commercial data was considered by this work, while allowing the users to retrieve product information. Their work supported both the identifier and feature-based product searches by building two encrypted index trees. In [25], a secure phrase search model for encrypted

data of cloud-based IoT was presented. The phrase search determined the location relationship between words by presenting homomorphic encryption and bilinear mapping. Additionally, the user's search pattern was protected by a probabilistic trapdoor generation algorithm. The search accuracy of this work was enhanced, while suffering from moderate overheads.

A practical framework for secure data retrieval was presented for encrypted cloud storage systems in [26]. The document retrieval system was integrated to a proxy server for improving the system's security. Their approach created two AVL trees for managing file names and authors, while an HRF tree was used to manage document vectors. The HRF tree was manipulated by Depth-First search algorithm and the tree was encrypted by Enhanced Asymmetric Scalar Product Preserving Encryption (EASPE) algorithm. These three index trees were inter-linked to support search requests. A certificateless keyword searching scheme for fog-based IoT was proposed in [27]. They developed a public key encryption method that was demonstrated to be secure against inside keyword guessing attacks (IKGA). The symmetric encryption key generation and distribution methods were presented inclusive of search and retrieval methods.

In [28], a searchable encryption scheme based on double-layered blockchain was presented for multi-cloud storage. The encrypted documents along with indexes were stored in Interplanetary File System (IPFS) with hash values and IPFS addresses in block chain. Verification scheme was also provided to check the data integrity of the retrieved file. In [29], an open-source cloud tool was utilized to store and retrieve files based on digital bipartite and digit compact prefix indexing methods. OpenStack was employed for setting up multiple nodes. A secure keyword search and retrieval scheme upon hashed encrypted data was presented in [30]. The data owner outsourced the hashed encrypted data in addition to a searchable index tree to cloud server. The data users could search and retrieve the data. A centralized keyword search scheme for cloud applications was proposed in [31]. In their work, a centralized manager searched and accessed all the data from authorized users. This work was featured by short ciphertext and search result verification.

In [32], a three-tiered architecture that could support data storage and retrieval was presented. Initially, the optimal data chunk size was computed by the cuckoo search algorithm and the chunk existence was verified by the hash value produced by SHA3-512. The data deduplication speed was increased by Merkle Hash Tree with MapReduce (MHT), while data storage and retrieval were performed. A secure data sharing and retrieval system for cloud data was presented in [33], which was based on blockchain technique.

Motivated by these existing works, this paper presents a data storage and retrieval system for cloud environment, which focuses to enhance the user search experience with faster and accurate retrieval rates.

3. Proposed DDR framework for cloud data

This work presents a data retrieval framework for deduplicated data in the cloud computing environment. The retrieval frameworks upon deduplicated data are not well-explored, as the cloud data retrieval frameworks focus more on security and privacy preservation. This work functions upon deduplicated data such that the redundant data are removed and the retrieval process is effectively done in the minimal period. To achieve better data retrieval, this work presents two protocols for Data Outsourcing Protocol(DOP) and Data Retrieval Protocol (DRP). The overall flow of the proposed work is shown in Figure 1.

The DOP involves several phases, which include data pre-processing, keyterm extraction, index building and encrypted index formation. The DRP includes phases such as query formation, detailed query formation, encrypted retrieval and ranking. In this work, a data retrieval process relies on different entities such as security policy, deduplicated document collection, keyterm collection, queries, query representation and ranking function.

Hence, the DDR framework is dedicated to retrieve ranked relevant documents with respect to the query, out of a pool of deduplicated documents, while adhering to the security policy [34–38]. This work understands that the users exploit limited key terms for searching the document collection, owing to the security requirements, such that two protocols, Data Outsourcing Protocol (DOP) and Data Retrieval Protocol (DRP) are expounded (Table 1).

3.1. Data outsourcing protocol (DOP)

It is assumed that the data owner has outsourced the Deduplicated Documents (DD) to the Cloud Server (CS).

3.1.1. Document pre-processing

This phase obtains the deduplicated data and tokenizes it to perform operations such as stopping and stemming. A list of tokens is obtained after the process of tokenization, where the processes of stopping and stemming are concerned with the removal of meaningless (on its own such as conjunctions, prepositions, articles and so on) terms and clipping of words (removing -ing,-es, -al and so on).

3.1.2. Keyterm extraction

For every DD, the key terms are extracted and ranked with respect to relevance and the top-ranking terms are chosen as the searchable key-term entities(kt_i), which

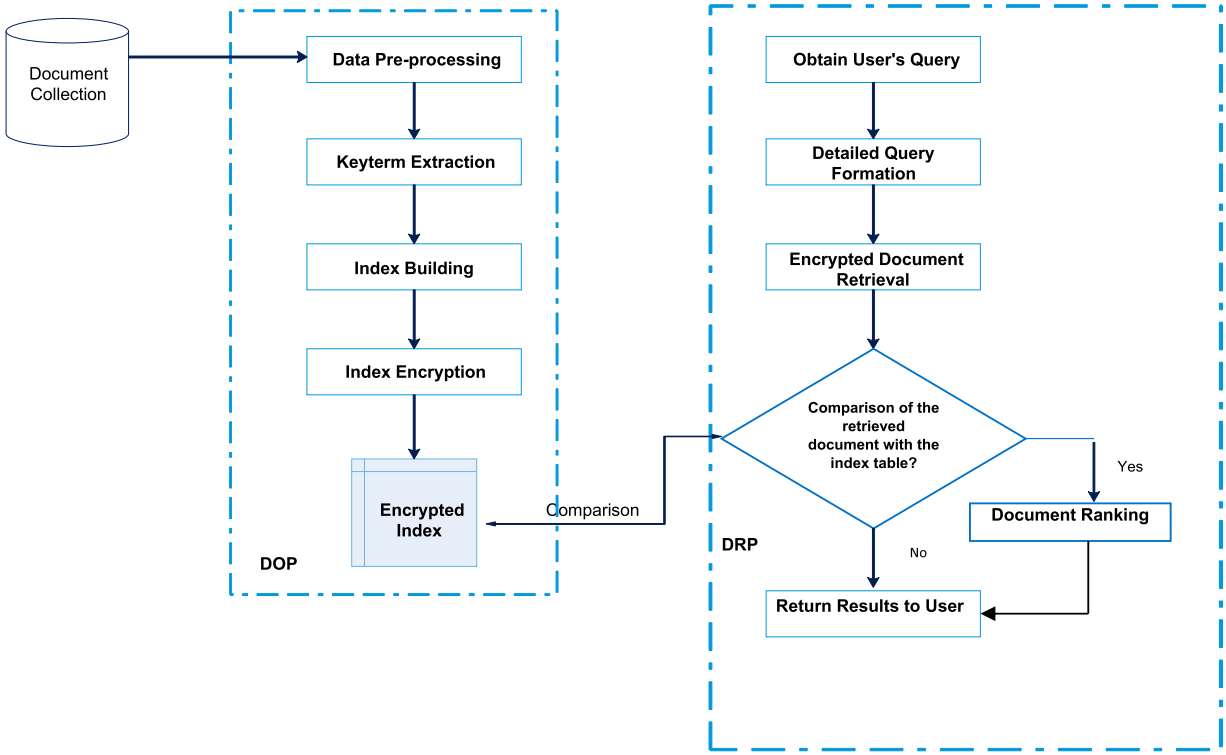

Figure 1. Overall flow of DDR framework.

Table 1. Notations and descriptions.

Notations	Description	Notations	Description
CBS	Corpus Based Similarity	***ENC***	Encryption process
CS	Cloud Server	kt_i	Key Term entities
CSP	Cloud Service Provider	*KT*	Key Term
DD	Deduplicated Documents	KT^+	Extended key term index
DDR	Deduplicated Data Retrieval	q_i	Set of user's query
DID	Document Identifier	*Q*	Query
DDR	Deduplicated Data Retrieval	sec_k	Secret key
DOP	Data Outsourcing Protocol	*SI*	Searchable Encrypted Index
DRP	Data Retrieval Protocol	*TD*	Trapdoor value
DU	Data Users		

can be represented by

$$KT = \{kt_1, kt_2, kt_3, \ldots, kt_n\} \tag{1}$$

3.1.3. Index building

Here, KT is processed to get rid of the problem of sparse index. For every key term $kt_i \in KT$, semantic extension is computed and the index is built, as presented in Equation (2).

$$\{TD(kt_i), ENC(\text{sec}_k, \{\langle kt_j, sim(kt_i, kt_j)\rangle\}_{j=1}^{L} \tag{2}$$

In the above equation, $TD(kt_i)$ denotes the trapdoor value of kt_i and kt_j is the nearest L semantic neighbours in KT semantic extension of key terms kt_i, where the corpus-based similarity is computed by

$$CBS(kt_i, kt_j) = \frac{\max\{logf(kt_i), logf(kt_j)\} - logf(kt_i, kt_j)}{\log N - \min\{logf(kt_i), logf(t_j)\}} \tag{3}$$

In the above equation, $f(kt_i)$ represents the total count of documents that contain the term kt_i in DD. The total count of documents that contain both kt_i and kt_j in DD are represented by (kt_i, kt_j). $N$is the total number of documents and the extended index of key terms are outsourced as KT^+.

3.1.4. Index encryption

The searchable encrypted representations are produced for indexing the encrypted documents that possess the key term, as represented by the following equation:

$$SI = \{si_1, si_2, si_3, \ldots, si_{|SI|}\} \tag{4}$$

For every document $dd_i \in DD$, the searchable index $si_i \in SI$ is computed by

$$\{\{TD(kt_j)\}_{j=1}^{KT_i}, ENC(Sec_k, DID(dd_i) \times ||\{ < kt_j, P(kt_j)) >\}_{j=1}^{KT_j})\} \tag{5}$$

In the above equation, the trapdoor value of kt_j is represented by $TD(kt_j)$, DID stands for the document identifier of dd_i and $P(kt_j)$ represents the statistical features of kt_j in dd_j, which includes term and inverse document frequency. Now, the searchable encrypted documents SI along with extended key term index KT^+ are outsourced to cloud storage. Hence, the process of

DOP is explained and the succeeding section describes the working principle of DRP.

3.2. Data retrieval protocol (DRP)

This work is based on the assumption that the authorization process is effectively done between Data Owner (DO), Data User (DU) and Cloud Service Provider (CSP). The cloud server enables the search process to the authorized users with the help of distinct set of keywords KT, upon the encrypted data SI. When the user passes a query Q, the cloud server returns relevant documents to the users. The below-given section discusses the phases involved in this protocol.

3.2.1. Query formation

When the user needs some information, the user forms a query and submits it to the cloud server. The query is a collection of words that strongly represent the required information and is represented as

$$Q = \{q_1, q_2, q_3, \ldots, q_{|Q|}\} \tag{6}$$

3.2.2. Detailed query formation

In this phase, the user calculates the trapdoor value of q_i represented by $TD(q_i)$, where $q_i \in Q$ and submits to the cloud along with the Sec_k shared between the DO and DU. When the CSP receives $TD(q_i)$, comparison is performed with the existing index table, as indicated as

$$TD(kt_i), ENC(\text{sec}_k, \{< kt_j, L(kt_i, kt_j) > \}_{j=1}^{L}) \tag{7}$$

As a result, the server returns an extended version with L tuples, as follows.

$$\text{ENC}(\text{sec}_k, \{< kt_j, L(kt_i, kt_j) >\}_{j=1}^{L}) \tag{8}$$

The user then decrypts the results returned by the server and expands the query with L tuples, as represented by

$$Q^{+} = \{q_1, q_2, q_3, \ldots, q_{|Q|*L}\} \tag{9}$$

3.2.3. Encrypted document retrieval

The user is then prompted to compute the trapdoor value of the extended version of the query and submit it to the server, which is indicated by

$$Sec_k(TD(q_i)); q_i \in Q^{+} \tag{10}$$

The server performs a comparison operation between it and index table and responds the user with matching encrypted document identifiers.

3.2.4. Document ranking

DU performs a decryption operation upon the returned results and ranks all the documents based on $\{DID(dd_i), kt_j, P(kt_j)\}$ with respect to retrieval models such as VSM and language modelling with relevance degree and the final DIDs are generated to the most relevant documents. These DIDs are forwarded to the server, after which the encrypted documents are returned. Now, the DU decrypts the results and obtains the necessary documents being requested.

4. Results and discussion

The performance of the proposed DDR is tested on two different datasets such as TREC and Cranfield corpuses [34,39]. The TREC corpus utilized in this work consists of approximately 1,35,000 documents and the query is formed with fields such as title, description and narration. This work extracts the title alone for consideration. The Cranfield corpus contains about 1400 documents concerning aerodynamics with 225 queries. The efficiency of the proposed work is tested in terms of retrieval precision, recall, F-measure and time consumption rates against existing works such as CIR [19], PIR [24] and document retrieval [26].

The effectiveness of any information retrieval systems is evaluated by precision and recall rates. The precision and recall rates can be improved by reducing the False Positive (FP) and False Negative (FN) rates, while improving the True Positive (TP) and True Negative (TN) rates. The efficiency of the retrieval system depends on the number of relevant documents returned in response to the submitted query by the user. F-measure rate depends on the precision and recall rates of the system. Precision is computed by computing the ratio of the count of retrieved relevant documents to the actual number of documents that are retrieved. Recall is the standard measure of a retrieval system, which is the total count of retrieved relevant documents to the total count of relevant documents in the dataset. The formulae to compute these performance measures are denoted as

$$F_m = \frac{2 \times P \times R}{P + R} \tag{11}$$

Here, P and R denote the precision and recall rates are expressed as

$$P = \frac{Total\ relevant\ documents\ retrieved}{Total\ documents\ retrieved} \tag{12}$$

$$R = \frac{Total\ relevant\ documents\ retrieved}{total\ relevant\ documents} \tag{13}$$

The precision and recall rates attained by the proposed work are tabulated in Table 2 and shown in Figure 2.

In Cloud Information Retrieval (CIR), the efficiency of information retrieval algorithms is typically measured in terms of precision and recall. This CIR [19] technique used TREC and Cranfield corpuses datasets and evaluate the precision and recall. However, precision and recall vary depending on the number of relevant documents returned. A single document return

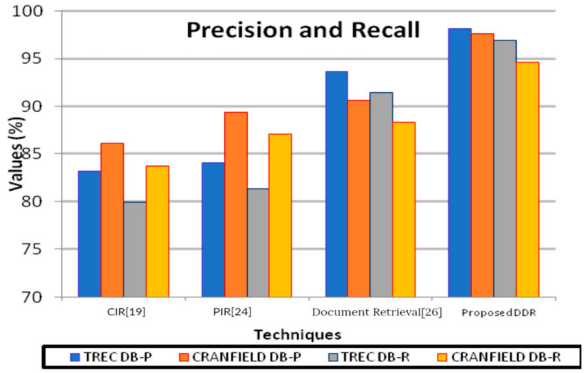

Figure 2. Precision and recall rate comparison.

Table 2. Performance analysis w.r.t precision and recall rates.

	Precision rates (%)		Recall (%)	
Techniques/Perf. Measures	Trec	Cranfield	Trec	Cranfield
CIR [19]	83.2	86.1	79.9	83.7
PIR [24]	84.6	89.4	81.3	87.1
Document retrieval [26]	93.6	90.3	91.4	88.3
Proposed DDR	**98.2**	**97.6**	**96.9**	**94.6**

would have a high recall but a low precision, and vice versa. In Product Information Retrieval (PIR) [24], the cloud server does not own the independent secret keys that are used to symmetrically encrypt the product information files. However, the returned documents would have a high recall but low precision. In document retrieval [26], the files are encrypted and stored. While performing retrieval, the query is evaluated and returns the relevant document to the user. In document retrieval [26], the experimental results were obtained from TREC and Cranfield corpuses datasets with precision and recall as performance metrics.

Based on the above results, it is observed that the proposed DDR framework shows better precision and recall rates. This indicates that the proposed DDR framework returns relevant documents in response to the user's query. The reasons for attaining this result are the employment of two protocols concerning data outsourcing and retrieval, which considers extended version of key terms.

In the above figure, P and R stand for precision and recall rates of the data retrieval systems. It is difficult to detect P and R rates, as the relevant documents must be known in prior. This is done as groundwork and the relevant documents are figured out before the execution of the system. The proposed DDR framework shows satisfactory results, as it returns relevant documents with respect to the user's query.

The P and R rates are the base inputs for the computation of F-measure rates. When a retrieval system proves better precision and recall rates, obviously the F-measure rates are increased. The time consumption analysis of the proposed DDR system is shown in Table 3.

Table 3. Time consumption analysis.

Techniques	Time consumption (ms)
CIR [19]	2586
PIR [24]	2491
Document retrieval [26]	2468
Proposed DDR	2436

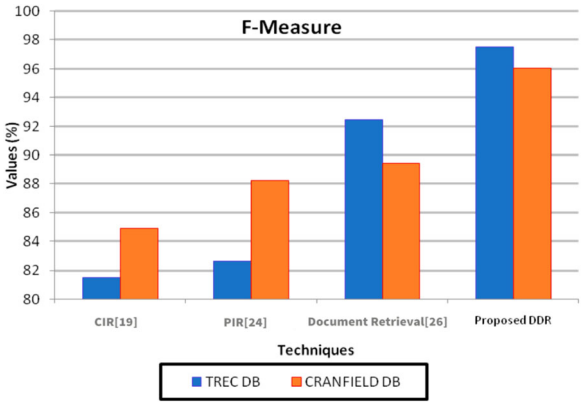

Figure 3. F-measure comparison analysis.

As observed in Table 3, the time consumption of the proposed DDR is found to be minimal, when compared to the existing works. Though the time difference is minimal, the retrieval relevance rate of the proposed work is better than the existing approaches. Hence, the proposed DOP and DRP perform better, which results in improved retrieval results in the reasonable period. Figure 3 shows the F-measure rate of the proposed work.

The F-measure rate comparative analysis results of the proposed work against existing works are shown in Figure 3.

5. Conclusion

This work presents a data retrieval framework for deduplicated cloud data. Most of the existing cloud data retrieval systems focus on privacy preservation and security, while this work considers retrieval performance. Additionally, data retrieval frameworks for deduplicated cloud data are scarce in the existing literature. This work presents two protocols meant for data outsourcing and retrieval. The data outsourcing protocol is concerned with documents and index terms, while the retrieval protocol handles queries and index terms. The performance of the work is tested in terms of precision, recall, f-measure and time consumption rates. The attained results are compared with the existing works and the proposed DDR framework outperforms the existing works, due to the formation of an effective extended version of index and relevance computation. In the future, this work is planned to incorporate security feature by enforcing access control policy.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Tabrizchi H, Kuchaki Rafsanjani M. A survey on security challenges in cloud computing: issues, threats, and solutions. J Supercomput. 2020;76(12):9493–9532.

[2] Basu S, Bardhan A, Gupta K, et al. Cloud computing security challenges &solutions-A survey. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2018, January. p 347–356.

[3] Li F, Ma J, Miao Y, et al. Towards efficient verifiable Boolean search over encrypted cloud data. IEEE Transact Cloud Comput. 2021;11(1):839–853.

[4] Chen Z, Zhang F, Zhang P, et al. Multi-user Boolean searchable encryption supporting fast ranking in mobile clouds. Comput Commun. 2020;164:100–113.

[5] Wang C, Cao N, Ren K, et al. Enabling secure and efficient ranked keyword search over outsourced cloud data. IEEE Trans Parallel Distrib Syst. 2012 Aug;23(8):1467–1479.

[6] Wang C, Cao N, Li J, et al. "Secure ranked keyword search over encrypted cloud data," in Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst; 2010. p. 253–262.

[7] Zarezadeh M, Mala H, Ashouri-Talouki M. Multi-keyword ranked searchable encryption scheme with access control for cloud storage. Peer-to-Peer NetworkingAppl. 2020;13(1):207–218.

[8] Yin H, Qin Z, Zhang J, et al. Secure conjunctive multi-keyword ranked search over encrypted cloud data for multiple data owners. Future Gener Comput Syst. 2019;100:689–700.

[9] Xiao T, Han D, He J, et al. Multi-Keyword ranked search based on mapping set matching in cloud ciphertext storage system. Conn Sci. 2021;33(1):95–112.

[10] Cao N, Wang C, Li M, et al. Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Trans Parallel Distrib Syst. 2014;25(1):222–233.

[11] Xia Z, Wang X, Sun X, et al. A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. IEEE Trans Parallel Distrib Syst. 2016;27(2): 340–352.

[12] Li J, Chen X, Li M, et al. Secure deduplictaion with efficient and reliable convergent key management. IEEE Trans Parall Distr. 2014 Jun;25(6):1615–1625.

[13] Xia W, Zhou Y, Jiang H, et al. "FastCDC: a fast and efficient Content-defined chunking approach for data deduplication", USENIX Annual Technical Conference (USENIC ATCí16), Denver, Co, USA; 2016 June.

[14] Chen R, Mu Y, Yang G, et al. BL-MLE: block-level message locked encryption for secure large file deduplication. IEEE Trans Inf Foren Sec. 2015 Aug;10(2):2643–2652.

[15] Shin Y, Kim K. Efficient and secure file deduplication in cloud storage. IEICE Trans Inf Syst. 2014 Feb;E97-D(2):184–197.

[16] Sun J, Hu S, Nie X, et al. Efficient ranked multi-keyword retrieval with privacy protection for multiple data owners in cloud computing. IEEE Syst J. 2019;14(2):1728–1739.

[17] Ferreira B, Rodrigues J, Leitao J, et al. Practical privacy-preserving content-based retrieval in cloud image repositories. IEEE Transact Cloud Comput. 2017;7(3):784–798.

[18] Wang T, Yang Q, Shen X, et al. A privacy-enhanced retrieval technology for the cloud-assisted internet of things. IEEE Trans Ind Inf. 2021;18(7):4981–4989.

[19] Yang Z, Tang J, Liu H. Cloud information retrieval: model description and scheme design. IEEE Access. 2018;6:15420–15430.

[20] He H, Chen R, Liu C, et al. An efficient ciphertext retrieval scheme based on homomorphic encryption for multiple data owners in hybrid cloud. IEEE Access. 2021;9:168547–168557.

[21] Xu Y, Zhao X, Gong J. A large-scale secure image retrieval method in cloud environment. IEEE Access. 2019;7:160082–160090.

[22] Wang X, Wang X, Wang D. Cost-efficient data retrieval based on integration of vc and ndn. IEEE Trans Veh Technol. 2021;70(1):967–976.

[23] Feng X, Ma J, Liu S, et al. Transparent ciphertext retrieval system supporting integration of encrypted heterogeneous database in cloud-assisted IoT. IEEE Internet Things J. 2021;9(5):3784–3798.

[24] Zhao YS, Zeng QA. Secure and efficient product information retrieval in cloud computing. IEEE Access. 2018;6:14747–14754.

[25] Shen M, Ma B, Zhu L, et al. Secure phrase search for intelligent processing of encrypted data in cloud-based IoT. IEEE Internet Things J. 2018;6(2):1998–2008.

[26] Fu J, Wang N, Cui B, et al. A practical framework for secure document retrieval in encrypted cloud file systems. IEEE Trans Parallel Distrib Syst. 2021;33(5):1246–1261.

[27] Olakanmi OO, Odeyemi KO. A certificateless keyword searchable encryption scheme in multi-user setting for fog-enhanced industrial internet of things. Transact Emerging Telecommun Technol. 2022;33(4):e4257.

[28] Fu S, Zhang C, Ao W. Searchable encryption scheme for multiple cloud storage using double-layer blockchain. Concurrency Comput: Practice Experience. 2022; 34(16):e5860.

[29] PriyaPonnuswamy P, VidhyaPriya R, Shabari Ram CP. File retrieval and storage in the open source cloud tool using digital bipartite and digit compact prefix indexing method. Concurrency Comput: Practice Exp. 2019;31(14):e5307.

[30] Sathyabalaji N, Komarasamy G, Raja DMS. Secure and privacy-preserving keyword search retrieval over hashed encrypted cloud data. Int J Commun Syst. 2020;33(5):1–13.

[31] Jiang P, Mu Y, Guo F, et al. Centralized keyword search on encrypted data for cloud applications. Security Commun Networks. 2016;9(18):5064–5084.

[32] Rasina Begum B, Chitra P. SEEDDUP: a three-tier SEcurE data DedUPlication architecture-based storage and retrieval for cross-domains over cloud. IETE J Res. 2023;69(4):2224–2241.

[33] Gajmal YM, Udayakumar R. Privacy and utility-assisted data protection strategy for secure data sharing and retrieval in cloud system. Inf Secur J: Global Perspect. 2022;31(4):451–465.

[34] https://github.com/samujjwaal/Cranfield-Vector-Space-Model

[35] Vinoth R, Jegatha Deborah L. A survey on efficient storage and retrieval system for the implementation of data deduplication in cloud. In: Pandian A, Palanisamy R, Ntalianis K, editors. Proceeding of the international conference on computer networks, big data and IoT (ICCBI - 2019). ICCBI 2019. Lecture notes on data

engineering and communications technologies. Cham: Springer International Publishing; 2020. p. 876–884. https://doi.org/10.1007/978-3-030-43192-1_95
[36] Lin Y, Mao Y, Zhang Y, et al. Secure deduplication schemes for content delivery in mobile edge computing. Comput Secur. 2022;114:102602.
[37] Almrezeq N. An enhanced approach to improve the security and performance for deduplication. Turkish J Comput Math Edu (TURCOMAT). 2021;12(6): 2866–2882.
[38] Tian G, Hu Y, Wei J, et al. Blockchain-based secure deduplication and shared auditing in decentralized storage. IEEE Trans Dependable Secure Comput. 2021;19(6):3941–3954.
[39] https://github.com/diazf/trec-data