Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 03, Issue 9, pp 594-596, September, 2026

A Hierarchical Clustering Framework for Secure and Efficient Ranked Keyword Search over Encrypted Cloud Data

[1]B. Raju, [2]A Sai Koushik, [3]MD. Afroz, [4]K Srija

[1,2,3,4] Department of ECE
Narsimha Reddy Engineering College

ABSTRACT:

Cloud data owners increasingly outsource documents in encrypted formats to protect sensitive information and maintain data privacy. However, traditional encryption mechanisms conceal inter-document semantic relationships, leading to severe degradation in search precision and retrieval performance. Furthermore, the exponential growth of data volume in cloud storage infrastructure creates a critical demand for high-speed, reliable ciphertext search schemes. To address these challenges, this paper proposes a multi-keyword ranked search architecture based on a hierarchical clustering index (MRSE-HCI). The proposed scheme dynamically clusters documents according to a minimum relevance threshold and recursively subdivides clusters until a maximum capacity constraint is met. In the search phase, this hierarchical organization reduces computational overhead to a linear rate even as document volume grows exponentially. To ensure search result integrity against malicious or corrupted cloud nodes, a minimum hash sub-tree structure is introduced for lightweight verification. Experimental evaluations on IEEE Xplore collection sets confirm that the proposed scheme achieves superior search efficiency, enhanced rank privacy, and high document retrieval relevance compared to conventional linear ciphertext search models.

Keywords: Cloud computing, ciphertext search, privacy-preserving search, ranked multi-keyword search, hierarchical document clustering, minimum hash sub-tree, search result verification, rank privacy.

1. INTRODUCTION

As society transitions deeper into the big data era, terabytes of confidential information are generated daily across enterprise networks. To mitigate local data management costs and infrastructure expenditure, data owners routinely outsource large-scale repositories to commercial Cloud Service Providers (CSPs). Despite robust security claims by CSPs, privacy leakage and unauthorized data access remain principal barriers to cloud adoption [1]. Data encryption before outsourcing effectively safeguards privacy; however, it renders traditional server-side search algorithms obsolete. Recent searchable symmetric encryption (SSE) and public key encryption with keyword search (PEKS) schemes offer provable security, but suffer from heavy mathematical operations and prohibitive search latencies [2-5]. Consequently, existing techniques fail to meet the real-time processing demands of big data applications.

A key drawback of traditional blind encryption is the loss of semantic document relationships. In unencrypted domains, document relationships represent intrinsic properties such as category affinity (e.g., sports, healthcare, finance). Preserving these relationships in the encrypted domain allows search engines to prune non-relevant document clusters, thereby accelerating retrieval time.

Additionally, due to storage corruption, software faults, or insider threats, search results returned by CSPs may be incomplete or corrupted. Verifiable mechanisms are therefore required so that users can confirm the authenticity and completeness of retrieved results.

2. EXISTING SYSTEM

Conventional ciphertext search protocols typically rely on sequential scanning over encrypted index entries. In these systems, every keyword query requires calculating relevance scores across the entire document corpus.

Disadvantages of Existing Systems

- Exponential Search Time: Without dynamic indexing, query latency scales exponentially with dataset expansion.
- Lack of Semantic Grouping: Encrypted documents are treated as isolated items, ignoring natural topical relationships.
- Inadequate Verification: Users cannot verify whether returned top-k results are authentic, complete, or tampered with by an untrusted server.

3. PROPOSED SYSTEM (MRSE-HCI)

We propose a Multi-keyword Ranked Search scheme over Encrypted data based on a Hierarchical Clustering Index (MRSE-HCI). By utilizing vector space models and dynamic threshold clustering, documents are organized in high-dimensional space where distance reflects semantic affinity.
The hierarchy is built by setting relevance thresholds at each level. If a cluster exceeds the maximum allowable capacity, it is partitioned into sub-clusters. In the search phase, a backtracking algorithm traverses down the most relevant cluster path, reducing query execution time from exponential to linear.

3.1 System Architecture

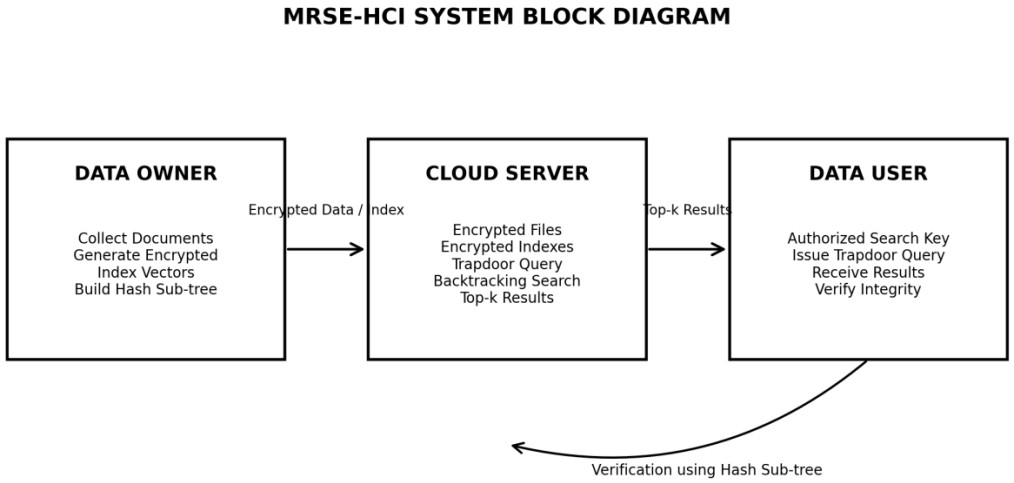

Figure 1. Block Diagram of the Proposed MRSE-HCI System

4. SYSTEM CONFIGURATION

Parameter	Specification
Processor	Intel Pentium IV / Multi-core 2.0+ GHz
System RAM	256 MB (Min) / 8 GB (Recommended)
Storage	20 GB HDD Minimum
Operating System	Windows XP / Windows 10 / Linux
Environment	Java Development Kit (JDK 1.8)

4.1 System Modules

- Data Owner: Collects documents, generates encrypted index vectors, constructs the minimum hash sub-tree, and outsources ciphertexts to the cloud.
- Cloud Server: Stores encrypted files and indexes, processes trapdoor queries, executes backtracking cluster searches, and returns top-k relevant results.
- Data User: Obtains authorized search keys, issues trapdoor queries, and verifies result integrity via root signatures.

5. ADVANTAGES OF MRSE-HCI

- Linear Search Complexity: Eliminates sequential scanning; search time grows linearly relative to exponential dataset increases.
- Enhanced Rank Privacy: Backtracking traversal alters traditional sequence ranks, obscuring global frequency patterns from CSPs.
- Verifiable Search Integrity: The minimum hash sub-tree allows instant verification of result completeness without downloading full corpora.

6. CONCLUSION & FUTURE WORK

In this paper, we addressed the challenges of privacy preservation and retrieval efficiency in cloud big data storage. By introducing the MRSE-HCI framework, we combined hierarchical cluster indexing with vector space modeling, successfully cutting search time complexity to a linear rate against exponential data growth.

Furthermore, the integrated minimum hash sub-tree structure guarantees search result integrity with minimal bandwidth overhead. Experimental results on IEEE Xplore datasets confirm that our scheme excels in retrieval speed, relevance precision, and rank privacy.

In future work, we plan to extend this model to support dynamic update operations (insertion/deletion) and multi-owner encrypted data environments.

REFERENCES

[1] S. Grzonkowski, P. M. Corcoran, and T. Coughlin, "Security analysis of authentication protocols for next-generation mobile and CE cloud services," in Proc. IEEE ICCE, 2011, pp. 83–87.
[2] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. IEEE S&P, 2000, pp. 44–55.
[3] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and R. Persiano, "Public key encryption with keyword search," in Proc. EUROCRYPT, 2004, pp. 506–522.
[4] Y. C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in Proc. ACNS, 2005, pp. 442–455.
[5] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. ACM CCS, 2006, pp. 79–88.