

**Privacy-Preserving Multi-keyword Top-k Similarity Search Over Encrypted Data**

**Abstract:**

Cloud computing provides individuals and enterprises massive computing power and scalable storage capacities to support a variety of big data applications in domains like health care and scientific research, therefore more and more data owners are involved to outsource their data on cloud servers for great convenience in data management and mining. However, data sets like health records in electronic documents usually contain sensitive information, which brings about privacy concerns if the documents are released or shared to partially untrusted third-parties in cloud. A practical and widely used technique for data privacy preservation is to encrypt data before outsourcing to the cloud servers, which however reduces the data utility and makes many traditional data analytic operators like keyword-based top-k document retrieval obsolete. In this paper, we investigate the multi-keyword top-k search problem for big data encryption against privacy breaches, and attempt to identify an efficient and secure solution to this problem. Specifically, for the privacy concern of query data, we construct a special tree-based index structure and design a random traversal algorithm, which makes even the same query to produce different visiting paths on the index, and can also maintain the accuracy of queries unchanged under stronger privacy. For improving the query efficiency, we propose a group multi-keyword top-k search scheme based on the idea of partition, where a group of tree-based indexes are constructed for all documents. Finally, we combine these methods together into an efficient and secure approach to address our proposed top-k similarity search. Extensive experimental results on real-life data sets demonstrate that our proposed approach can significantly improve the capability of defending the privacy breaches, the scalability and the time efficiency of query processing over the state-of-the-art methods.