An Efficient Privacy-Preserving Ranked Keyword Search Method

Objective:

The objective of this system is cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving.

Abstract:

Cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving. Therefore it is essential to develop efficient and reliable ciphertext search techniques. One challenge is that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design ciphertext search schemes that can provide efficient and reliable online information retrieval on large volume of encrypted data. In this paper, a hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast ciphertext search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed in this paper. Experiments have been conducted using the collection set built from the IEEE Xplore. The results show that with a sharp increase of documents in the dataset the search time of the proposed method increases linearly whereas the search time of the traditional

method increases exponentially. Furthermore, the proposed method has an advantage over the traditional method in the rank privacy and relevance of retrieved documents.

Introduction:

As we step into the big data era, terabyte of data are produced world-wide per day. Enterprises and users who own a large amount of data usually choose to outsource their precious data to cloud facility in order to reduce data management cost and storage facility spending. As a result, data volume in cloud storage facilities is experiencing a dramatic increase. Although cloud server providers (CSPs) claim that their cloud service is armed with strong security measures, security and privacy are major obstacles preventing the wider acceptance of cloud computing service.

In this paper, a vector space model is used and every document is represented by a vector, which means every document can be seen as a point in a high dimensional space. Due to the relationship between different documents, all the documents can be divided into several categories. In other words, the points whose distances are short in the high dimensional space can be classified into a specific category. The search time can be largely reduced by selecting the desired category and abandoning the irrelevant categories. Comparing with all the documents in the dataset, the number of documents which user aims at is very small. Due to the small number of the desired documents, a specific category can be further divided into several sub-categories. Instead of using the traditional sequence search method, a backtracking algorithm is produced to search the target documents. Cloud server will first search the categories and get the minimum desired sub-category. Then the cloud server will select the desired k documents from the mini-mum desired sub-category. The value of k is previously decided by the user and sent to the cloud server. If current sub-

category cannot satisfy the k documents, cloud server will trace back to its parent and select the desired documents from its brother categories. This process will be executed recursively until the desired k documents are satisfied or the root is reached. To verify the integrity of the search result, a verifiable structure based on hash function is constructed. Every document will be hashed and the hash result will be used to represent the document. The hashed results of documents will be hashed again with the category information that these documents belong to and the result will be used to represent the current category. Similarly, every category will be represented by the hash result of the combination of current category information and sub-categories information. A virtual root is constructed to represent all the data and categories. The virtual root is denoted by the hash result of the concatenation of all the categories located in the first level. The virtual root will be signed so that it is verifiable. To verify the search result, user only needs to verify the virtual root, instead of verifying every document.