An Efficiency Keyword Search Scheme to improve user experience for Encrypted Data in Cloud

Jiangang Shu, Xingming Sun, Lu Zhou, Jin Wang

School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract: With the increasing popularity of cloud computing, more and more sensitive or private information has been outsourced onto the cloud server. For protecting data privacy, sensitive data usually has to be encrypted before outsourcing, which makes traditional search techniques based on plaintext useless. In response to the search over encrypted data, searchable encryption is a good solution in Information Security. However, most of existing searchable encryption schemes only support exact keyword search. That means they don’t support searching for different variants of the query word, which is a significant drawback and greatly affects data usability and user experience. In this paper, we formalize the problem of semantic keyword-based search over encrypted cloud data while preserving privacy. Semantic keyword-based search will greatly improve the user experience by returning all the documents containing semantically close keywords related to the query word. In our solution, we use the stemming algorithm to construct stem set, which reduces the dimension of index. And the symbol-based trie is also adopted in index construction to improve the search efficiency.

Keywords: Semantic search, searchable encryption, stemming algorithm, cloud computing

1 Introduction

With the advent of cloud computing, more and more sensitive or private information has been outsourced onto the cloud in order to benefit from unlimited storage space and enormous computation power. The data stored in the cloud may suffer from malicious use or unauthorized access by the cloud service provider, classically considered as “honest-but-curious”. In recent years, searchable encryption has attracted a lot of attention from a growing number of researchers. However, most of existing searchable encryption schemes only support exact keyword search, except the fuzzy keyword search scheme proposed by Li et al. [11]. That is, it is common that the user’s searching input might not exactly match those pre-set keywords due to different variants of the word with semantically close meaning, such as “compute”, “computed”,...
“computing”, etc. The native way to support semantically close keyword search is to generate all the keywords within proximity in a semantic sense from the original keyword and send all the semantically close keywords as queries to the cloud server. This basic solution achieves the basic goal of semantically close keyword search however it is ineffective.

In this paper, we formalize the problem of semantic keyword-based search over encrypted cloud data while preserving privacy. Moreover, we combine the searchable encryption technique with a stemming algorithm. This method eliminates the need of enumerating all the semantically close keywords both in index construction and query generation, and also reduces the dimension of the index. In addition, we propose an efficient semantic keyword-based search scheme, which greatly enhances searching flexibility by returning all the documents containing the semantically close keywords related to the given query word. Rigorous privacy analysis shows that our scheme is secure. Through experiments on the real-world dataset and performance analysis, we show our scheme is quite efficient and infeasible.

2 Related Work

Searchable encryption is a new information security technique and it can enable users search over encrypted outsourced data through keywords without decrypting the data at first. The first practical searchable encryption scheme in symmetric setting is proposed by Song et al. [1], in which keywords are encrypted by a deterministic encryption algorithm under a two-layer construction and users have to go through the whole document collection to search a certain keyword. After that, Goh [2], Chang et al. [3] and Curtmola et al. [4] propose the improvement schemes based on similar index to improve search efficiency. Boneh et al. [5] propose the first public key-based searchable encryption scheme, where anyone owning the private key can search the data encrypted by the public key. In recent years, many schemes and security definitions have been put forward under different applications [6-20]. Among those schemes, some interesting issues such as conjunctive or range query [6-9], multi-keyword search [10, 15, 17], result secure ranking [13, 14, 17], fuzzy keyword search [11, 12, 15], similarity search [16] and preferred search [18] have been proposed and discussed. Specially, in order to address the problems of minor typos and format inconsistence, Li et al. [11] propose a fuzzy keyword search scheme which combines edit distance with wildcard-based technique to construct fuzzy set. This scheme might address the problem of different variants of a word to a certain extent, but the performance of the scheme is largely affected by the variable $d$ of edit distance and the dimension of index also increases a lot.
3 Design Goals and Threat Model

3.1. Design goals

In this paper, we address the problem of semantic keyword-based search service over the encrypted cloud data. Specifically, we focus on the following goals:

- **Flexible Search** Our scheme should provide the flexible search service, which can return all the documents containing the semantically close keywords related a given query keyword.

- **Privacy-Preserving** The general goal is to protect the cloud server from learning additional sensitive information from document collection, index and search request. Specifically, we are concerned with document privacy, index privacy and query confidentiality.

- **Efficiency** The proposed schemes should be achieved with low communication and computation both on the data owner and user.

3.2. Threat model

In our scheme, the cloud server is considered as “honest-but-curious” like many previous works [10, 17] of searchable encryption. That means, the cloud server will execute designated protocol honestly and correctly, but it is eager to get some sensitive information by inferring and analyzing data or index in its storage and the search request received during the protocol.

4 Construction of Semantic Keyword-based Search

The key idea behind the semantic keyword-based search contains that design an efficient semantic keyword-based search scheme based on the construction of the stem set, which can return all the documents containing the semantically close keywords related to a given query word. The scheme contains the following algorithms.

- **Setup** In this initialization phase, the data owner initiates the scheme to generate a random key $sk \in \mathbb{F}^{k}$.

- **BuildIndex**
  - The data owner firstly computes $T_{s_i} = f(sk, s_i)$ for each $s_i \in S$, $1 \leq i \leq r$. Then he divides them into symbols as $T_{s_i} = [a_{i,1}, a_{i,2}, \ldots, a_{i,L}]$ and builds up the symbol-based trie $G$ covering all the stems of $s_i \in S$.
– He attaches each $FID_{s_i} = Enc(sk, \{FID_{w_i} || w_i\}_{w_i \in W(s_i)})$ to $G$ and outsources it, together with the encrypted document collection $C$, onto the cloud server.

- **GenQuery** For a given query word $w_i$, the data user performs the stemming process to attain the stem $s_i$. Then he computes $T_{s_i} = f(sk, s_i)$ and divides it into symbols as $T_{s_i} = \{a_{i_1}, a_{i_2}, ..., a_{i_L}\}$. At last, he submits the search request $T_{s_i}$ to the cloud server.

- **Search** Upon receiving the search request, the cloud server performs the search operation over the index $G$ and returns $FID_{s_i}$ to the data user. The user decrypts the returned results and retrieves relevant files.

5 **Performance analysis**

In order to evaluate performance of our proposed scheme, we set up the experiment on a publicly available real dataset: the Enron Dataset [21]. We implement all the algorithms in our paper on a 2.83GHz Intel Core(TM) processor, Windows 7 operating system with a RAM of 4G.

5.1. Index construction

Index construction conducted in the data owner mainly involves: computing the hash value of each stem, dividing the hash value into symbols and building the symbol-based trie covering those symbols. Figure 1 shows that the time cost to build index is nearly linear with the number of stems. Given the 18,711 stems, it just takes about 4.6s to build up the whole index.

5.2. Search

Figure 2 shows the time cost of the search operation. Note that searching irrelevant stems is much faster than searching relevant ones. The reason is that if there exists a mismatch between the symbols of the trapdoor and the index during the path exploration, the search operation will terminate at once.

6 **Conclusion**

In the paper, we discuss and address the problem of querying different variants of a keyword. Combining with the stemming algorithm, we propose a semantic keyword-based search scheme over encrypted cloud data. Given a query word, data
users can find all the documents containing the semantically close keywords or different variants through our scheme, which tackles the limitation of exact keyword search. Through experimental study on real dataset, our scheme is quite practical.

**Acknowledgements.** This work is supported by the NSFC (61232016, 61173141, 61173142, 61173136, 61103215, 61373132, 61373133), GYHY201206033, 201301030, 2013DFG12860, BC2013012 and PAPD fund.

![Fig. 1. Time of index construction](image1)

![Fig. 2. Time of search](image2)
References

19. Islam, M., Kuzu, M., Kantarcioglu, M.: Access pattern disclosure on searchable encryption: