International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

Multi keyword Searching Techniques for Top-K Retrieval in Encrypted Cloud

Student: Nakeertha Anusha (13H61D5820)
Guide: A. Jyothi
(Assistant Professor) Cvsr Engineering College

Abstract—
Cloud computing has emerging as a promising pattern for data outsourcing and high-quality data services. However, concerns of sensitive information on cloud potentially causes privacy problems. Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable symmetric encryption (SSE) allows retrieval of encrypted data over cloud. In this paper, we focus on addressing data privacy issues using SSE. For the first time, we formulate the privacy issue from the aspect of similarity relevance and scheme robustness. We observe that server-side ranking based on order-preserving encryption (OPE) inevitably leaks data privacy. To eliminate the leakage, we propose a two-round searchable encryption (TRSE) scheme that supports top-k multikeyword retrieval. In TRSE, we employ a vector space model and homomorphic encryption. The vector space model helps to provide sufficient search accuracy, and the homomorphic encryption enables users to involve in the ranking while the majority of computing work is done on the server side by operations only on ciphertext. As a result, information leakage can be eliminated and data security is ensured. Thorough security and performance analysis show that the proposed scheme guarantees high security and practical efficiency.

Index Terms—Cloud; data privacy; ranking; similarity relevance; homomorphic encryption; vector space model

INTRODUCTION

CLOUD computing, a critical pattern for advanced data service, has became a necessary feasibility for data users to outsource data. Controversies on privacy, however, have been incessantly presented as outsourcing of sensitive information including e-mails, health history and personal photos is explosively expanding. Reports of data loss and privacy breaches in cloud computing systems appear from time to time . The main threat on data privacy roots in the cloud itself . When users outsource their private data onto the cloud, the cloud service providers are able to control and monitor the data and the communication between users and the cloud at will, lawfully or unlawfully. Instances such as the secret NSA program, working with AT&T and Verizon, which recorded over 10 million phone calls between American citizens, cause uncertainty among privacy advocates, and the greater powers it gives to telecommunication companies to monitor user activity. To ensure privacy, users usually encrypt the data before outsourcing it onto cloud, which brings great challenges to effective data utilization. However, even if the encrypted data utilization is possible, users still need to communicate with the cloud and allow the cloud operate on the encrypted data, which potentially causes leakage of sensitive information.

Furthermore, in cloud computing, data owners may share their outsourced data with a number of users, who might want to only retrieve the data files they are interested in. One of the most popular ways to do so is through keyword-based retrieval. Keyword-based retrieval is a typical data service and widely applied in plaintext scenarios, in which users retrieve relevant files in

a file set based on keywords. However, it turns out to be a difficult task in cipher text scenario due to limited operations on encrypted data. Besides, to improve feasibility and save on the expense in the cloud paradigm, it is preferred to get the retrieval result with the most relevant files that match users' interest instead of all the files, which indicates that the files should be ranked in the order of relevance by users' interest and only the files with the highest relevances are sent back to users.

A series of searchable symmetric encryption (SSE) schemes have been proposed to enable search on ciphertext. Traditional SSE schemes enable users to securely retrieve the ciphertext, but these schemes support only Boolean keyword search, i.e., whether a keyword exists in a file or not, without considering the difference of relevance with the queried keyword of these files in the result. To improve security without sacrificing efficiency, schemes presented in show that they support top-k single keyword retrieval under various scenarios. The authors of made attempts to solve the problem of top-k multikeyword over encrypted cloud data. These schemes, however, suffer from two problems—Boolean representation and how to strike a balance between security and efficiency. In the former, files are ranked only by the number of retrieved keywords, which impairs search accuracy. In the latter, security is implicitly compromised to tradeoff for efficiency, which is particularly undesirable in security-oriented applications.

Preventing the cloud from involving in ranking and entrusting all the work to the user is a natural way to avoid information leakage. However, the limited computational power on the user side and the high computational overhead precludes information security. The issue of secure multikeyword top-k retrieval over encrypted cloud data, thus, is: How to make the cloud do more work during the process of retrieval without information leakage.

In this paper, we introduce the concepts of similarity relevance and scheme robustness to formulate the privacy issue in searchable encryption schemes, and then solve the insecurity problem by proposing a two-round searchable encryption (TRSE) scheme. Novel technologies in the cryptography community and information retrieval (IR) community are employed, including homomorphic encryption and vector space model. In the proposed scheme, the majority of computing work is done on the cloud while the user takes part in ranking, which guarantees top-k multikeyword retrieval over encrypted cloud data with high security and practical efficiency. Our contributions can be summarized as follows:

1. We propose the concepts of similarity relevance and scheme robustness. We, thus, perform the first attempt to formulate the privacy issue in searchable encryption, and we show server-side ranking based on order-preserving encryption (OPE) inevitably violates data privacy.
2. We propose a TRSE scheme, which fulfills the secure multikeyword top-k retrieval over encrypted cloud data. Specifically, for the first time, we employ relevance score to support multikeyword top-k retrieval.
3. Thorough analysis on security demonstrates the proposed scheme guarantees high data privacy. Furthermore, performance analysis and experimental results show that our scheme is efficient for practical utilization.

Relevance Scoring

Some of the multikeyword SSE schemes support only Boolean queries, i.e., a file either matches or does not match a query. Considering the large number of data users and documents in the cloud, it is necessary to allow multikeyword in the search query and return documents in the order of their relevancy with the queried keywords. Scoring is a natural way to weight the relevance. Based on the relevance score, files can then be ranked in either ascendingly or descendingly. Several models have been proposed to score and

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

rank files in IR community. Among these schemes, we adopt the most widely used one tf-idf weighting, which involves two attributes-term frequency and inverse document frequency.

Statistic Leakage

Although all data files, indices, and requests are in encrypted form before being outsourced onto cloud, the cloud server can still obtain additional information through statistical analysis. We denote the possible information leakage with statistic leakage. There are two possible statistic leakages, including term distribution and interdistribution. The term distribution of term t is t's frequency distribution of scores on each file iði 2 CÞ. The interdistribution of file f is file f's frequency distribution of scores of each term jðj 2 fÞ. Term distribution and interdistribution are specific . They can be deduced either directly from ciphertext or indirectly via statistical analysis over access and search pattern . Here, access pattern refers to which keywords and the corresponding files have been retrieved during each search request, and search pattern refers to whether the keywords retrieved between two request are the same. Based on our observation, distribution information implies a similarity relationship among terms or files. On the one hand, terms with similar term distribution always have simultaneous occurrence. For instance, obviously, the term "states" are very likely to co-occur with "united" in an official paperwork from the White House, and their term distribution, not surprisingly, are very same in a series of such a kind of paperwork. Given that this paperwork is encrypted but term distribution is not concealed, once an adversary somehow cracks out the plaintext of "united," he can reasonably guess the term that shares a similar term distribution with "united" may be "states." On the other hand, files with similar interdistribution are always the same category, e.g., two medical records from a dental office surely are the same category, and they are very likely to share a similar interdistribution (such as the titles of each entries are the same). Therefore, this specificity should be hidden from an untrusted cloud server.

TRSE DESIGN

Existing SSE schemes employ server-side ranking based on OPE to improve the efficiency of retrieval over encrypted cloud data. However, server-side ranking based on OPE violates the privacy of sensitive information, which is considered uncompromisable in the security-oriented thirdparty cloud computing scenario, i.e., security cannot be tradeoff for efficiency. To achieve data privacy, ranking has to be left to the user side. Traditional user-side schemes, however, load heavy computational burden and high communication overhead on the user side, due to the interaction between the server and the user including searchable index return and ranking score calculation. Thus, the user-side ranking schemes are challenged by practical use. A more server-siding scheme might be a better solution to privacy issues. We propose a new searchable encryption scheme, in which novel technologies in cryptography community and IR community are employed, including homomorphic encryption and the vector space model. In the proposed scheme, the data owner encrypts the searchable index with homomorphic encryption. When the cloud server receives a query consisting of multikeywords, it computes the scores from the encrypted index stored on cloud and then returns the encrypted scores of files to the data user. Next, the data user decrypts the scores and picks out the top-k highestscoring files' identifers to request to the cloud server. The retrieval takes a two-round communication between the cloud server and the data user. We, thus, name the scheme the TRSE scheme, in which ranking is done at the user side while scoring calculation is done at the server side.

Efficiency Improvement

The main appeal of the modified FHEI that we employ in the TRSE scheme is its conceptual

simplicity compared to Gentry's . This simplicity is achieved at the cost of a large key size. Although optimizations like modular reduction and compression can be employed to reduce the size of ciphertext, the key size is still too large for the practical system.

The user encrypts his trapdoor and sends the ciphertext to the cloud server. Therefore, the communication overhead will be very high if the encrypted trapdoor size is too large. To solve this problem and, thus, improve efficiency, a tradeoff of the security of search pattern may be needed unless a new encryption scheme that provides more reasonable ciphertext size becomes available. Researchers from cryptography community have made several attempts to move toward practical fully homomorphic encryption over integers. These progresses indicate that the efficiency of the TRSE scheme can be further improved.

CONCLUSION

In this paper, we motivate and solve the problem of secure multikeyword top-k retrieval over encrypted cloud data. We define similarity relevance and scheme robustness. Based on OPE invisibly leaking sensitive information, we devise a server-side ranking SSE scheme. We then propose a TRSE scheme employing the fully homomorphic encryption, which fulfills the security requirements of multikeyword top-k retrieval over the encrypted cloud data. By security analysis, we show that the proposed scheme guarantees data privacy. According to the efficiency evaluation of the proposed scheme over a real data set, extensive experimental results demonstrate that our scheme ensures practical efficiency.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53, no. 4, pp. 50-58, 2010.

[2] M. Arrington, "Gmail Disaster: Reports of Mass Email Deletions," http://www.techcrunch.com/2006/12/28/gmail-disasterreportsof- mass-email-deletions/, Dec. 2006.

[3] Amazon.com, "Amazon s3 Availability Event: July 20, 2008," http://status.aws.amazon.com/s3-20080720.html, 2008.

[4] RAWA News, "Massive Information Leak Shakes Washington over Afghan War," http://www.rawa.org/temp/runews/2010/ 08/20/massive-information-leak-shakes-washington-overafghan- war.html, 2010.

[5] AHN, "Romney Hits Obama for Security Information Leakage," http://gantdaily.com/2012/07/25/romney-hits-obama-forsecurity- information-leakage/, 2012.