IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 20, 2017, accepted January 8, 2018, date of publication January 30, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2797131

Cloud Information Retrieval: Model Description and Scheme Design

ZHEN YANG[1], (Member, IEEE), JILIANG TANG[2], AND HUAN LIU[3], (Fellow, IEEE)
[1]College of Computer Science, Beijing University of Technology, Beijing, 100124 China
[2]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA
[3]School of Computing, Informatics, and Decision Systems Engineering, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ 85281 USA

Corresponding author: Zhen Yang (yangzhen@bjut.edu.cn)

This work was supported by the National Natural Science Foundation of China under 61671030.

ABSTRACT The fast development of cloud technology has brought about a new trend in the field of information service: more and more information is being transferred to the cloud as requested. However, the data, such as texts, images, sounds, and videos, before being moved to the cloud, in most cases, has to be encrypted so that intelligible information will not be obtained from unauthorized accesses. While having done a nice work in protecting the data privacy of its owners, this encrypting process, has produced a great challenge for retrieval of the document stored via traditional IR model based on document, query and relevance. In order to retrieve encrypted information from cloud, an alternative retrieval system is needed. To satisfy such a need, we have: 1) build a cloud information retrieval framework characterized by its retrieval risk formula, which, enables, for the very first time to the best of our knowledge, an effective retrieval of keywords from encrypted cloud data without undermining key word privacy and retrieval performance; and 2) upgraded the existing searchable encryption scheme that can only support simple equality queries on encrypted data and has been proved to perform slightly better than random selection, so that it can now support the state-of-art information retrieval methods, such as vector space, probabilistic, and language model. To evaluate the effect of the system proposed above, we've conducted a wide range of experiments on benchmark data sets, of which the results shows that solution can fulfill its purposes quite well in various settings.

INDEX TERMS Information retrieval, cloud computing, searchable encryption, keyword extraction, query expansion.

I. INTRODUCTION

With the rapid growth of internet information service, cloud computing has become a prevalent method of delivering software, data storage, and computing services. In particular, more and more sensitive information are being transferred to cloud. However, in most public cloud dominant architectures, it is cloud service provider (CSP) that manages and holds users' data, raising a great concern for privacy of data due to the lack of mutual trust between the data owner and CSP. In order to protect privacy, confidential files are usually encrypted prior to out-sourcing. Such a move, while alleviating the worry about information safety, have created a need for retrieving encrypted files. Data owners often find it necessary to share their outsourced data with a large number of users, who then have to retrieve the documents from the cloud. Considering the following scenario in the cloud in Figure 1, the company X needs to share the data, such as texts, images, sounds, and videos, among their employees (such as Alice, Bob, and Charles) in the cloud. In this scenario, the company X is data owner, the employees are data users, and the CSP is the cloud server. Since the Company X and CSP are not in the same trusted domain, data has to be encrypted before out-sourcing for data privacy, which forces the users to perform document retrieval from the encrypted data. As a result of this encrypting process, how to efficiently retrieve from cloud files has become a pragmatically important task, which is also technically challenging as shown below.

The existing approach to retrieval of encrypted files is to selectively retrieve files through keyword-based search instead of retrieving all the encrypted files back and then have them decrypted which would be impractical in the cloud computing scenarios. Specifically the encrypted document retrieval goes through the following procedures [2], [5], [11]:

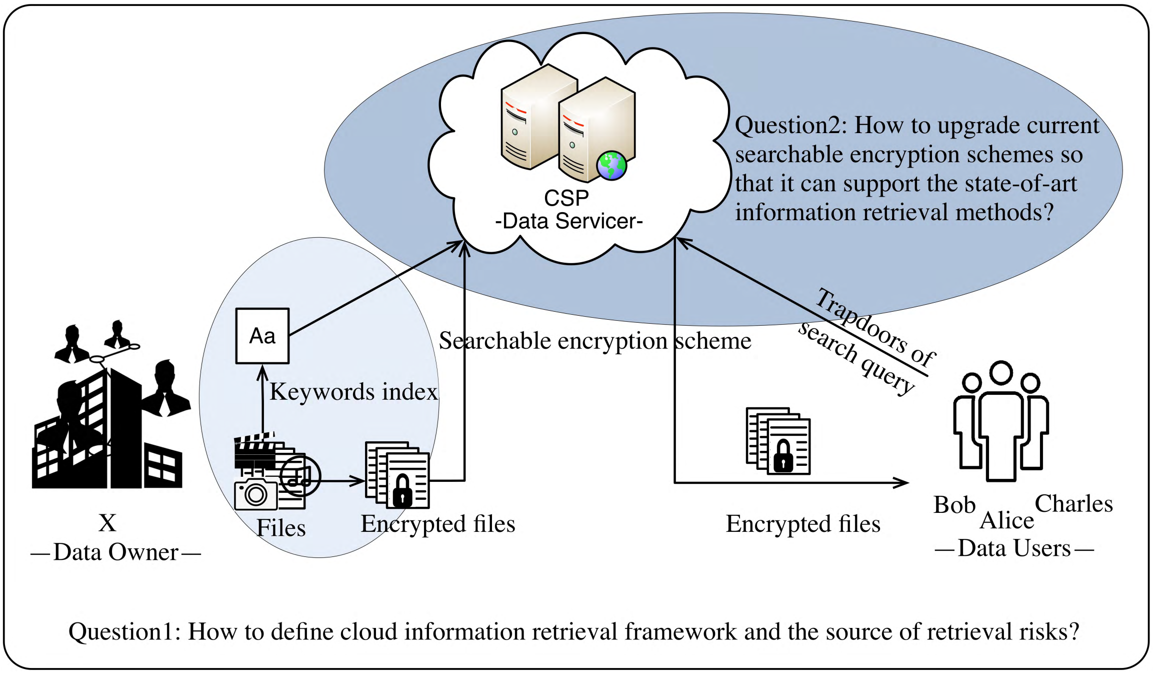

CSP
-Data Servicer-
Question2: How to upgrade current searchable encryption schemes so that it can support the state-of-art information retrieval methods?
Aa
Searchable encryption scheme
Trapdoors of search query
Keywords index
X
—Data Owner—
Files
Encrypted files
Encrypted files
Bob
Alice
Charles
—Data Users—
Question1: How to define cloud information retrieval framework and the source of retrieval risks?

FIGURE 1. Scenario of searchable encryption documents retrieval architecture in the cloud.

(1) Only a few or even one keyword(s) from documents are summarized for building up an index for keywords associated with these documents, which is different from plaintext document retrieval; (2) The encrypted index and documents are transferred to the cloud through searchable encryption scheme; (3) Users can retrieve the documents they need by integrating the trapdoors of queries with index information where both document content and keyword privacy are well-preserved.

In summary, the basic thinking behind the procedures is to meet the the needs of could service users for both sound - information security and efficient document utilization. To meet the need, many searchable encryption schemes, such as PEKS [2], Fuzzy EKS [11], HVE-PEKS [3], SCF-PEKS [9], and PERKS [22], MRSE [4], etc., have been developed already. However, most of them, as we find, are suffering from the drawback of giving not as much attention to retrieval performance as to security and privacy protection of data in cloud document retrieval. We thus decide that a retrieval system capable of high storage security and retrieval performance is in great need and we propose that for development of such a system, the following two questions need to well answered.

- Question 1: How to define cloud information retrieval framework and the source of retrieval risks? Most of the existing solutions focus on security and privacy protection of data and therefore are not functioning well in terms of their retrieval performance. For improvement of the performance, the framework and its risk need to be well defined.
- Question 2: How to upgrade current searchable encryption schemes so that it can support the state-of-art information retrieval methods? The existing schemes can only support simple equality queries on encrypted data and therefore only performs slightly better performance than random selection. However, because of the strict security policy and extremely short keyword index, construction of sate-of-art retrieval models is quite a challenge.

In the efforts to answer these questions, we've come to realize that cloud information retrieval can be formulated as a task of extremely short text retrieval with strict security and privacy policies and we've developed a retrieval system of which the superior performance is well established via a wide range of experiments conducted by us on multiple benchmark data sets. The major contributions of our efforts can be summarized as below:

- We build a cloud information retrieval (CIR) framework and define its retrieval risk formally. To the best of our knowledge, we are the first to study the problem of effective keyword retrieval over encrypted cloud data with high users' privacy and retrieval performance
- We design a searchable encryption scheme that can support the state-of-art retrieval methods. We illustrate the way to integrate the state-of-the-art retrieval models to cloud information retrieval framework. We also discuss the communication complexity of these retrieval models in CIR.

The remainder of the paper is organized as follows: In Section 2, the relevant research literature is review. In Section 3, the cloud information retrieval framework (CIR) is formalized and introduced. In Section 4, the details of searchable encryption schemes, including document outsourcing protocol and document retrieval protocol, are

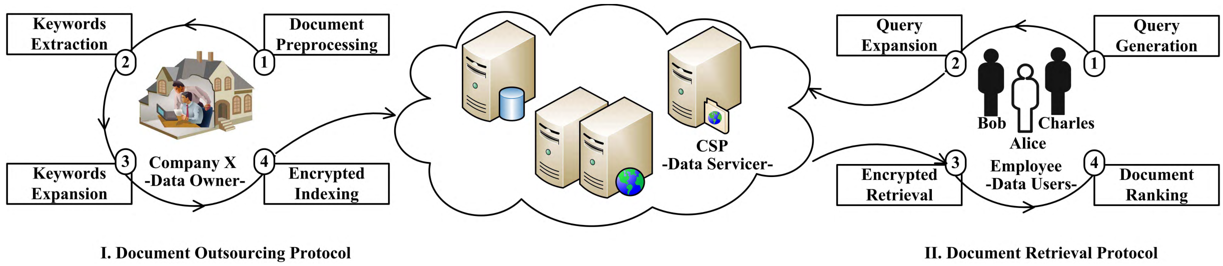


FIGURE 2. Typical framework of cloud information retrieval.

discussed. In Section 5, a wide range of experiments are conducted to evaluate the proposed system on multiple benchmark data sets. In Section 6, the concluding remarks are given and future research tasks are proposed.

II. RELATED WORK

Information retrieval is defined as a task of finding documents of an unstructured nature that satisfies an information need from large collections [14]. Much research has been done since Salton's early work on SMART [18] in the 1960s. Some important concepts in information retrieval were developed as parts of research on the SMART system, including the vector space model (*tf-idf* weighting), relevance feedback, and Rocchio classification. Since then document retrieval models as well as their applications have been thoroughly discussed. Over the last 40 years, many well-known primary retrieval models including probabilistic modeling (such as okapi BM25) [17], and language modeling [7], have also been developed. Moreover, researchers also ambitiously attempt to develop new information retrieval technologies to support more diverse retrieval applications including text, pictures, audio, speech and video, etc [13], [15], [24]. Especially, an advanced and novel feature extraction strategy has been proposed in [13]. And IR technology has been used to rank social media sites based on combined search engine query results [24]. Meanwhile, the TREC (Text Retrieval Conference), TDT (Topic Detection and Tracking), NLP&CC, ACE (Automatic Content Extraction), MUC (Message Understanding Conference), and NTCIR (NII-NACSIS Test Collection for Information Retrieval Systems) workshops also provide forums for document retrieval researchers and objective criteria to evaluate newest systems and methods.

Firstly, we need to redefine the problem of cloud information retrieval. Unfortunately, though many searchable encryption schemes have been proposed, such as PEKS [2], Fuzzy EKS [11], HVE-PEKS [3], SCF-PEKS [9], and PERKS [22], MRSE [4], most works focus on the facet of security and privacy protection of data in cloud information retrieval, while ignoring the fact that the retrieving performance is also crucial in addition to the need for security and privacy preservation.

Second, we need to design new searchable encryption scheme. Unfortunately, little attention was given to these issues although the current solutions only support simple equality queries on encrypted data that provide a slight better result than random selection. Boneth [3] assumed the retrieval was a matching between words in emails' subject and users' query. Many other researchers make similar assumptions [2]–[4], [9], [22]. Li *et al.* [11] proposed a fuzzy keyword search scheme to enhance system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. However, how to built the closest possible matching set is still an open problem.

III. CLOUD INFORMATION RETRIEVAL (CIR)

A. DEFINITION OF CLOUD INFORMATION RETRIEVAL

Mathematically, cloud information retrieval (CIR) can be denoted as a septuple $\{S, D, K, C, Q, F, R(Q, c_i)\}$, where:

- S is a security and privacy preserving policy.
- D is a set of representations for the documents in the collection.
- K is a searchable keyword field, namely the union set of all keywords of each document.
- C is a set of searchable encrypted representations for the documents in the collection D.
- Q is a set of logical representations for the user information needs, namely queries.
- F is a framework for modeling document representations, queries, and their relationships.
- $R(Q, c_i)$ is a ranking function that determines the ordering of c_i among C with regards to the query Q.

Thus, the **CIR is defined as a special information retrieval task (F), i.e., finding encrypted documents (C) stored in the cloud with a searchable, extremely sparse keyword index (K) that satisfies the brief information need (Q) under the restriction of security and privacy preserving policy (S).**

Supposing the strict security protocols are well satisfied and the keywords are properly extracted, which we will discuss in §3 and §4, we face a extremely short keyword index which even contains a few or even one keyword(s), because users hope to obtain their interested information with less communication for communication protocol security in the cloud, which poses a great challenge in measuring the relevance between searchable keyword index and query. To remedy this, as shown in Figure 2, we propose a CIR framework. If the company X (Data owner) wants to exchange or share some documents (Data) with employee (Data users) via cloud (Data server), he can follow the protocols below.

- **Document out-sourcing protocol**
  Step I: Preprocessing documents. Collect the documents D to be out-sourced, tokenize the text which turns each document into a list of tokens, and do linguistic preprocessing.
  Step II: Extracting keywords. For each document $d_i \in D$, select its top ranked words as the searchable keyword field k_i. Thus K denotes the union set of all the k_i.
  Step III: Expanding keywords. For each keyword in K, we expand it with its nearest neighbors in K. The expanded keyword set is denoted as K^+.
  Step IV: Encrypted indexing. Index the encrypted document that each keyword term occurs in, namely C. With the predefined encryption scheme S, C and K^+, documents can be out-sourced to the cloud server for storage.
- **Document retrieval protocol**
  Step I: Query generating. Users summary their information need and submit a query Q.
  Step II: Query expanding. For query Q, expand each query with its nearest neighbors in K^+. And the expanded query is denoted as Q^+.
  Step III: Encrypted retrieving. As the predefined encryption scheme S, the CSP compares query with the index table and returns the entire possible encrypted file IDs.
  Step IV: Document ranking. Users decrypt the returned results, rank each document according to different IR models, and generate the final file IDs of most relevant files of interest.

Note that the details of these protocols will be discussed in §4.

B. RISK OF CLOUD INFORMATION RETRIEVAL

The document out-sourcing protocol and document retrieval protocol can be simplified as a retrieval framework F:

$$F:\ D\times K\times Q \mapsto [d_i \otimes k_i] \odot [Q \oplus K] \tag{1}$$

where the operator $\otimes$ denotes the mask operator that removes these words not in keyword index, $\oplus$ semantic expansion, and $\odot$ a ranking function defined according to different retrieval models.

With Bayesian retrieval risk theory [28], we can further investigate the retrieval risk in the retrieval framework F. As shown in black solid line in Figure 3, a classic retrieval system can be regarded as an interactive information service system that answers a user's query by ranking a list of documents. The user's query (Q) is viewed as the output of some probabilistic model (Φ) associated with the user U. Similarly, a document collection (D) is viewed as the output of some probabilistic model (Ψ) associated with a document source (S). Thus, the document retrieval can be viewed as two Markov chains:

- $U \to \Phi \to Q$ and
- $S \to \Psi \to D$.

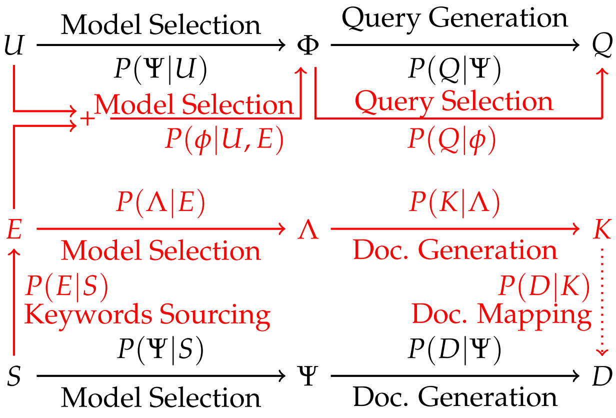


FIGURE 3. Diagram of cloud information retrieval.

However, in CIR, interactions between service systems and users become different. As shown in red solid line in Figure 3, the keyword index (K) can be viewed as the output of some probabilistic model(Λ) process associated with a keyword source (E) that is extracted from the source of documents (S) according to $P(E|S)$. An explicit query is only an abstract of the user's information needs and it's difficult to infer user's actual searching intents and interests. Therefore the user's query (Q) is viewed as the output of some probabilistic model (Φ) associated with the user (U) and expanded by the keyword source (E). Besides, as shown in red dotted line in Figure 3, after ranking the keyword index (K) by the similarity $P(K|Q)$, we also should map the keyword index to the documents and finally answer a user's query according to the mapping between keyword index and documents. Since these processes are not directly involved in the retrieval actions, they are denoted as dotted lines. The cloud information retrieval can be illustrated as below:

- $(U+E) \to \Phi \to Q$ and
- $S \to E \to \Psi \to K \dashrightarrow D$.

According to the risk minimization retrieval theory [27], the expected risk of a retrieval action r associated with a loss $L(r, \Phi, \Lambda)$ is given by

$$R(r|U,Q,E,K) = \int_{\Phi}\int_{\Lambda} L(r,\Phi,\Lambda)P(\Phi|U,E,Q)P(\Lambda|S,E,K)d\Phi d\Lambda \tag{2}$$

Thus we choose the optimal retrieval action r^* by the Bayes decision rule with the least expected risk:

$$r^* = \arg\min_r R(r|U,Q,E,K) \tag{3}$$

To achieve the least expected risk, we should make a good estimation of $P(\Phi|U,E,Q)$ and $P(\Lambda|S,E,K)$.

- For a good estimation of $P(\Lambda|S,E,K)$, we should provide a keyword detection method consistency with S. In fact, the performance of keyword extraction plays an important role in CIR, and many researches already showed that the randomly selected words in email subject or content of document are not valid as a keyword

TABLE 1. Notations used in the cloud information retrieval (CIR) framework.

$S = \{Setup(\lambda), Enc(sk,\cdot), Dec(sk,\cdot), T(\cdot)\}$	A symmetric encryption based security and privacy preserving policy:
$-\lambda$	- A security parameter;
$-Setup(\lambda)$	- A probabilistic algorithm that inputs a λ and outputs a secret key sk;
$-Enc(sk,\cdot)$	- The encryption algorithms;
$-Dec(sk,\cdot)$	- The decryption algorithms;
$-T(\cdot)$	- A one-way trapdoor function.
$D = \{d_1, d_2, \cdots, d_{\vert D\vert}\}$	A set of logical representations for the documents in the collection D.
$-d_i = \{w_1, w_2, \cdots, w_{\vert d_i\vert}\}$	- The i^th document in D.
$K = \{k_1, k_2, \cdots, k_{\vert K\vert}\}$	The searchable keywords field, namely the union set of all keywords of each document.
$-K^+$	- The expansion index of keywords: $\{T(t_i), Enc(sk, \{< t_j, Sim(t_i, t_j) >\}_{j=1}^{S})\}$.
$C = \{c_1, c_2, \cdots, c_{\vert C\vert}\}$	A set of searchable encrypted representations for the documents in the collection D.
$Q = \{q_1, q_2, \cdots, q_{\vert Q\vert}\}$	A set of logical representations for the user information needs, namely queries.
$-Q^+ = \{q_1, q_2, \cdots, q_{\vert Q\vert * S}\}$	- The $Q's$ S-tuple expansion;
F	A framework for modeling document representations, queries, and their relationships.
$R(Q, c_i)$	A ranking function defines the ordering of c_i among D with regards to the query Q.

candidate [2], [3]. In this paper, we use a entropy difference based metric (*ED*) [25] to evaluate and rank the relevance of words in a text. The *ED* measure considers the consistency between K and S, thus, the superior performance can be achieved.

- For a good estimation of $P(\Phi|U, E, Q)$, we should find a method to expand the short query keeping consistency between Q and E with the least risk. We used a corpus-based query expansion method to measure the semantic similarity of two words or phrases in the domain of "corpus" philosophy. It is a straightforward application of Google similarity distance [6], which is a method for automatically extracting similarity of words and phrases using corpus based on frequency page counts.

IV. SEARCHABLE ENCRYPTION SCHEME IN CIR

In this section, we propose a novel searchable encryption scheme to support the state-of-art information retrieval methods, and discuss the communication complexity of these retrieval models in CIR.

If the company X (Data owner) wants to exchange or share some documents (Data) with employees (Data users) via cloud (Data server), they should obey the following document out-sourcing protocol and document retrieval protocol.

A. DOCUMENT OUT-SOURCING PROTOCOL

Given a collection of documents D, we assume that the authorization between the data owner and CSP is appropriately done. As shown in Figure 2, with the notations summarized in Table 1, data owner should out-source the encrypted version C to the cloud server for storage as follows:

- **Step I: Document preprocessing**
  Collect the documents D to be out-sourced, tokenize the text which turns each document into a list of tokens, and do linguistic preprocessing (such as case folding, stemming and lemmatization, and stopword eliminations, etc.). Set up a symmetric encryption based on security and privacy preserving policy $S = \{Setup(\lambda), Enc(sk, \cdot), Dec(sk, \cdot), T(\cdot)\}$ with a security parameter λ.
- **Step II: Keyword extraction**
  For each document $d_i \in D$, keyword extraction ranks the words according to its relevance and selects the top ranked words as the searchable keyword field k_i. Thus the K denotes the union set of all the keywords field k_i.
- **Step III: Keyword expansion**
  To alleviate the extremely sparse keyword index, we implement a query expansion based on the keyword field K. For each keyword $t_i \in K$, we compute its S-tuple semantic expansion and build the index $\{T(t_i), Enc(sk, \{< t_j, Sim(t_i, t_j) >\}_{j=1}^{S})\}$, where $T(t_i)$ is the trapdoor value of t_i, and t_j is its top S semantic nearest neighbors in K semantic expansion of keywords t_i with corpus based similarity $Sim(t_i, t_j)$:

$$Sim(t_i, t_j) = \frac{max\{\log f(t_i), \log f(t_j)\} - \log f(t_i, t_j)}{\log N - min\{\log f(t_i), \log f(t_j)\}} \tag{4}$$

  where $f(t_i)$ denotes the number of documents containing word t_i in D, $f(t_i, t_j)$ the number of documents containing both word t_i and t_j in D, N the total number of documents in D. Therefore, the expansion index of keywords is outsourced as K^+.
- **Step IV: Encrypted indexing**
  Index the encrypted documents that each keyword term occurs in, namely generating the searchable encrypted representations $C = \{c_1, c_2, \cdots, c_{|C|}\}$. For each $d_i \in D$, its searchable index $c_i \in C$ can be denoted as $\{\{T(t_j)\}_{j=1}^{|k_i|}, Enc(sk, FID(d_i) \parallel \{< t_j, P(t_j)) >\}_{j=1}^{|k_j|})\}$, where $T(t_j)$ is the trapdoor value of t_j, $FID(d_i)$ is the file ID of document d_i, and $P(t_j)$ is the statistical characteristic of t_j in d_j, such as term frequency, inverse document frequency, and probability, etc. Thereafter, the searchable encrypted documents C and keyword expansion K^+ can be out-sourced to the cloud server for storage.

B. DOCUMENT RETRIEVAL PROTOCOL

Assume that the authorization among the data owner, data user and CSP is appropriately done. Through a predefined set of distinct keyword K, the cloud server provides the search service for the authorized users over the encrypted data C. With query Q, as shown in Figure 2, the data user could retrieve the documents and return the search results according to the following steps:

- **Step I: Query generation**
  If a user wants to seek a specific document, he would summarize his information need as a query. Given query $Q = \{q_1, q_2, \cdots, q_{|Q|}\}$, it is a set of logical representations of user's information need.
- **Step II: Query expansion**
  The authorized user computes the trapdoor $T(q_i)$ of $q_i \in Q$ and sends it back to the cloud with a secret key sk shared between data owner and authorized users. Upon the CSP receiving the $T(q_i)$, the server compares it with the predefined index table $\{T(t_i), Enc(sk, \{< t_j, S(t_i, t_j) >\}_{j=1}^{S}\}$, then returns S-tuple encrypted keyword expansion $Enc(sk, \{< t_j, S(t_i, t_j) >\}_{j=1}^{S}$. The user decrypts the returned results and expands the Q with its S-tuple expansion: $Q^{+} = \{q_1, q_2, \cdots, q_{|Q|*S}\}$.
- **Step III: Encrypted retrieval**
  To search with Q^{+}, the authorized user computes the trapdoor $T(q_i)$ of $q_i \in Q+$ and sends it back to the cloud with a secret key sk shared between data owner and authorized users. When the CSP receives the $T(q_i)$, the server compares it with the index table and returns the entire possible encrypted file IDs: $Enc(sk, FID(d_i) \parallel \{< t_j, P(t_j)) >\}_{j=1}^{|k_j|})$.
- **Step IV: Document ranking**
  The user decrypts the returned results, ranks each document based on the $\{FID(d_i), < k_j, P(k_j) >\}$ according to different IR models, such as VSM, probabilistic modeling and language modeling with relevant scores, and generates the final file IDs of most relevant files. The user sends these IDs back to the CSP and the encrypted documents are returned. Finally, the user decrypts the returned results and receives the relevant files.

It is true that we focus on the design of encryption searchable scheme and don't conduct the rigorous security analysis of communication protocol. Since the proposed protocol falls into the same protocol family with Fuzzy EKS and PEKS, the rigorous security analysis is deduced with same consideration so that the scheme can be proved to be secure and privacy-preserving. Similar security proof process can be found in Ref. [2], [11], to save space, we omit the detailed proof.

C. RETRIEVAL MODELS IMPLEMENTATION

In above, we proposed the general searchable encryption scheme in CIR. Here we explore the details of how to support the state-of-the-art retrieval models into searchable encryption protocols.

With the notations summarized in Table 1, given documents collection D, expanded query Q^{+}, the ranking methods with different retrieval models can be denoted as:

- Boolean Retrieval Modeling [23]
  AND operator: q_1 AND q_2 AND $\cdots q_{|Q|\times S}$,
  OR operator: q_1 OR q_2 OR $\cdots q_{|Q|\times S}$.
- VSM Retrieval Modeling [19]

$$R(Q^{+}, d_i) = \frac{\sum_{j=1}^{|Q|\times S} I(q_j, d_i) \cdot f(q_j, Q^{+}) \cdot f(q_j, d_i)}{\sqrt{\sum_{j=1}^{|Q|\times S} f(q_j, Q^{+})^2} \cdot \sqrt{\sum_{j=1}^{|Q|\times S} f(q_j, d_i)^2}} \tag{5}$$

- Probability Retrieval Modeling [8]

$$R(Q^{+}, d_i) = \sum_{j=1}^{|Q|\times S} IDF(q_j) \frac{I(q_j, d_i) \cdot f(q_j, d_i) \cdot (k_1 + 1)}{f(q_j, d_i) + k_1(1 - b + \frac{|d_i|}{avgdl})}$$

$$IDF(q_j) = \log \frac{N - df(q_j) + 0.5}{df(q_j) + 0.5} \tag{6}$$

- Language Modeling [1], [8]

$$R(Q^{+}, d_i) = \prod_{j=1}^{|Q|\times S} I(q_j, d_i) \cdot p(q_j|d_i) \tag{7}$$

$$p(q_j|d_i) = I(q_j, d_i)(\lambda \frac{f(q_j, d_i)}{N_d} + (1 - \lambda) \frac{f(q_j, D)}{N_D}) \tag{8}$$

where $I(q_j, d_i) = 1$ if q_j is a keyword of d_i and 0 otherwise, $f(q_j, d)$ $q_j's$ term frequency in the document d_i, $f(q_j, d)$ $q_j's$ term frequency in the entire collection D, $df(q_j)$ the number of documents containing q_j, $|d_i|$ the length of the document d_i in words, N the total number of documents in the collection, $avgdl$ the average document length in the text collection D. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$ [1], and $0 < \lambda < 1$.

The word frequency of each $q_j \in Q+$ should be weighed by

$$tf(q_j) = \begin{cases} Sim(q_j, q_o) \cdot tf(q_j) & \text{if } q_j \text{ is expanded by } q_o \\ tf(q_j) & \text{otherwise} \end{cases} \tag{9}$$

where $Sim(q_j, q_o)$ denotes the semantic similarity between q_j and q_o defined in Equation (4).

To support the retrieval models above, once the authorized user computes the trapdoor $T(q_j)$ of q_j and sends it back to the cloud with a secret key sk shared between data owner and authorized users, the CSP compares $T(q_j)$ with the index table and returns all the possible encrypted documents $Enc(sk, FID(d_i) \parallel \{< t_j, P(t_j) >\}_{j=1}^{|k_i|})$. More specifically,

- For the Boolean model, since extra information is not necessary, the CSP can just return the $Enc(sk, FID(d_i))$ and $P(t_j) = 0$ obviously has the least affection on the system usability.
- For the VSM retrieval model, the CSP should return all the possible encrypted documents and $P(t_j) = f(t_j, d_i)$.
- For the probability model, the CSP should returns all the possible encrypted documents and $P(t_j) =< f(t_j, d_i), df(t_j) >$.

- For the language model, the CSP should returns all the possible encrypted documents and $P(t_j) \quad =< f(t_j, d_i), f(t_j, D) >$.

In short, to support different models, the CSP and users have different communication load, and we can safely draw the conclusion that the communication cost of Boolean retrieval model < VSM retrieval model < probability retrieval model ≈ language model. Considering the cloud computing scenarios and security of communication protocol, users hope to obtain what he wants with less communication between the user and the CSP database. With the decrease of keywords, the communication cost can be further reduced. In our experiments in §5, we can see that the CIP can achieve acceptable performance even with one keyword.

V. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATA ANALYSIS

As shown in Table 2, three data sets are used for experimental evaluation purposes in this work.

- TREC Corpus - The first set is TREC topics 401-450 on TREC disks 5 (http://www.nist.gov/srd/nistsd23.cfm), which is distributed for the development and evaluation of Information retrieval (IR) systems and related natural language processing research. The document collections consist of the full texts of various newspapers, newswire articles and government proceedings. The data set is the basis of the TREC 8 information retrieval competition and contains FBIS set (Data provided from the Foreign Broadcast Information Service, approx. 130,000 documents and 470 MB) and LAT set (Los Angeles Times which randomly selected articles from 1989&1990, approx. 130,000 documents and 475 MB). The corpus consists of 50 information needs, evaluated with four levels relevance evaluation on different but overlapping sets of documents. The TREC query sets include three sections for each query: title, description and narrative. Since the narrative and description are about what document is relevant, in our experiments, only the title field are used as query.
- Cranfield Corpus - Cranfield corpus is a well known IR test collection (http://www.clairlib.org/index.php/Corpora) containing 1,400 aerodynamics' documents. The Cranfield data set is much smaller, and much more specialized, containing abstracts from technical papers on aeronautical engineering. 225 queries are provided, with gold standard or original ground truth relevance judgments. However, the relevance evaluation is assigned on four levels. Other statistical characteristics are summarized in Tab. 2.

2) PERFORMANCE EVALUATION

In information retrieval, precision and recall [14] are commonly used to describe the effectiveness of information retrieval algorithms. However, these quantities clearly vary depending on the number of relevant documents returned.

TABLE 2. The statistical summary of three sets of documents: From left to right, total number of documents (TND), average length of document (ALD), number of queries (NQ), and average length of query (ALQ).

Corpus	TND	ALD	NQ	ALQ
FBIS	130,471	501.5733	50	Title: 2.38
				Desc: 6.58
LAT	131,896	499.1752	50	Title: 2.38
				Desc: 6.58
Cranfield	1400	147.17	225	9.1733

Returning just one document would result in a high precision but a low recall; vice versa. Meanwhile, measuring recall is much more difficult because we have to know the number of relevant documents in the entire collection, which means that all documents in the entire collection must be assessed [20]. Therefore, in our experiments, we choose two common metrics to evaluate the effectiveness of information retrieval algorithms: top-N accuracy rate $p(n)$) and average precision (AP).

$$p(n) = \frac{\#\ of\ \ relevant\ \ doc\ \ in\ \ top\ n\ \ results}{n} \tag{10}$$

$$AP = \frac{1}{R}\sum_{n=1}^{L} P(n) \times r(n) \tag{11}$$

Where n is the number of returned documents, L the number of documents in the ranking, R the total number of relevant documents. $r(n)$ is equal to 1 if the document in the position n of the ranking is relevant and 0 otherwise.

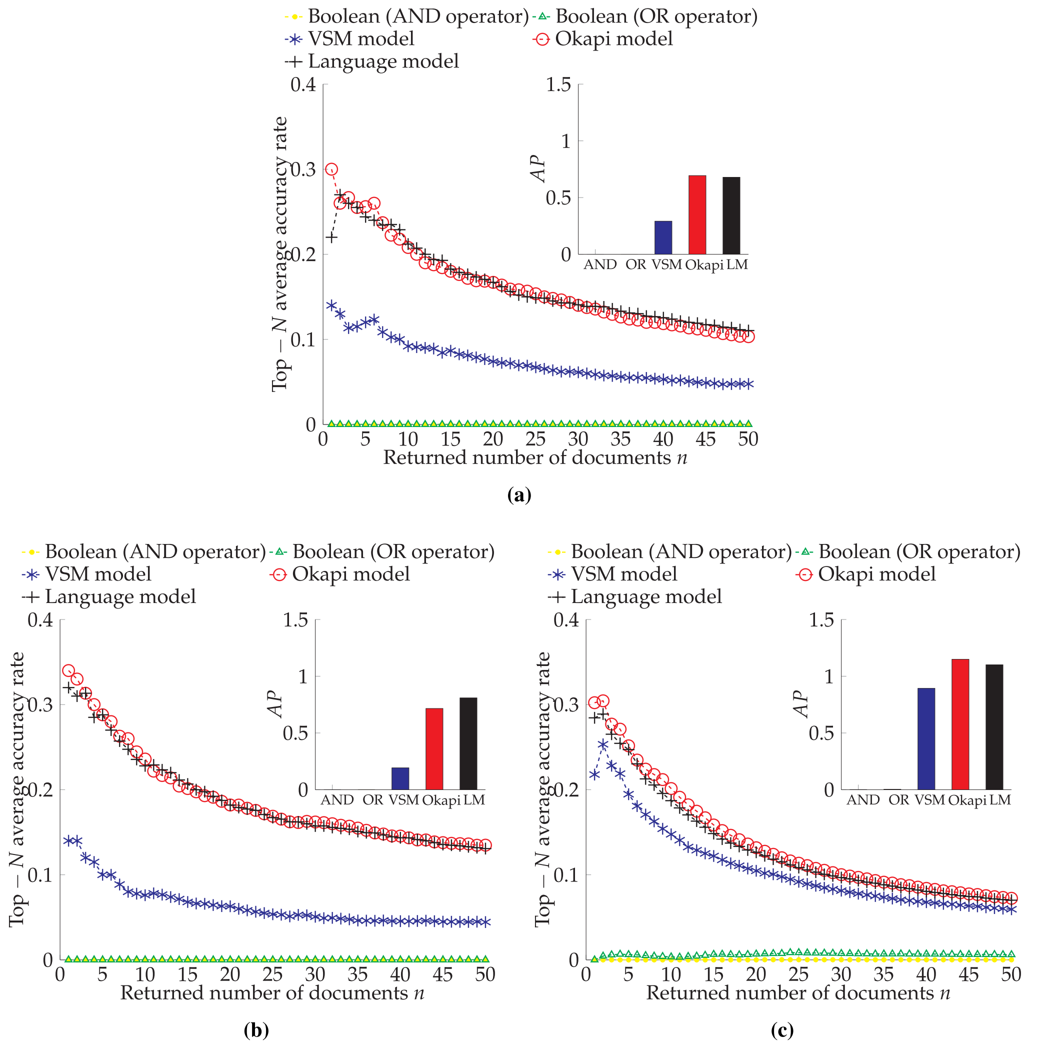
Thus, $p(n)$ denotes the fraction of correct documents in the top n results, and AP denotes the average of the precisions at the ranking positions where each relevant document is retrieved. Obviously, AP clearly varies depending on the number of relevant documents returned. However, it's impossible to assess the entire collection. So we calculate the Average precision (AP) for the top 50 (L=50) returned documents without loss of generality.

A Boolean retrieval model does not have a built-in way of ranking matched. i.e., it can't decide which document that satisfies $m + 1$ clauses in the query is more relevant than a document that satisfies the m clauses. But there may be some intrinsic property of a document that can serve as the basis of an useful ranking. In our experiments, like PubMed (http://www.ncbi.nlm.nih.gov/pubmed), we rank the returned results in order of publication recency, i.e., most recent first. For these documents without time stamps, we just assign random time labels to them.

3) EXPERIMENTAL SETTING

We conduct extensive experiments to evaluate the effectiveness of the proposed framework CIR. Through the experiments, we aim to answer the following two questions:

1) How effective is the proposed framework?
2) How do these parameters affect the performance of CIR?



(a)

(b)

(c)

FIGURE 4. Performance evaluation of basic retrieval models in CIR. Figure (a)-(c) show the top-*N* accuracy rate of the documents retrieval by Boolean model, VSM model, probability model (Okapi BM25), and language model in three data sets (TREC-LAT Corpus, TREC-FBIS and Cranfield Corpus). Inset: the same, but for average precision (*AP*) instead.

To answer these questions, we explore the performance of how well CIR supports the state-of-art retrieval models. We empirically evaluate the performance of the four basic retrieval models, including Boolean model, VSM model, probability model, and language model, in CIR. These comparisons would provide us with evidence about the effectiveness of CIR.

B. PERFORMANCE EVALUATION OF THE STATE-OF-ART RETRIEVAL MODELS

In this section, we examine the performance of CIR with different retrieval models, including the Boolean model (AND operator), Boolean model (OR operator), VSM model, probability model (Okapi BM25, $k1 = 1.2, b = 0.75$), and language model ($\lambda = 0.5$). We performed top-N accuracy rate (p) and average precision (AP) experiments on three data sets: TREC-LAT Corpus, TREC-FBIS corpus and Cranfield Corpus. We perform some preprocessing: all punctuation symbols were removed from the text, all words were changed to those lowercase and then a simple tokenization method based on whitespaces was applied.

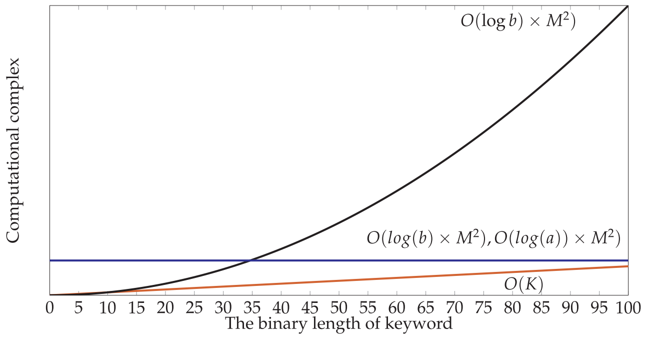
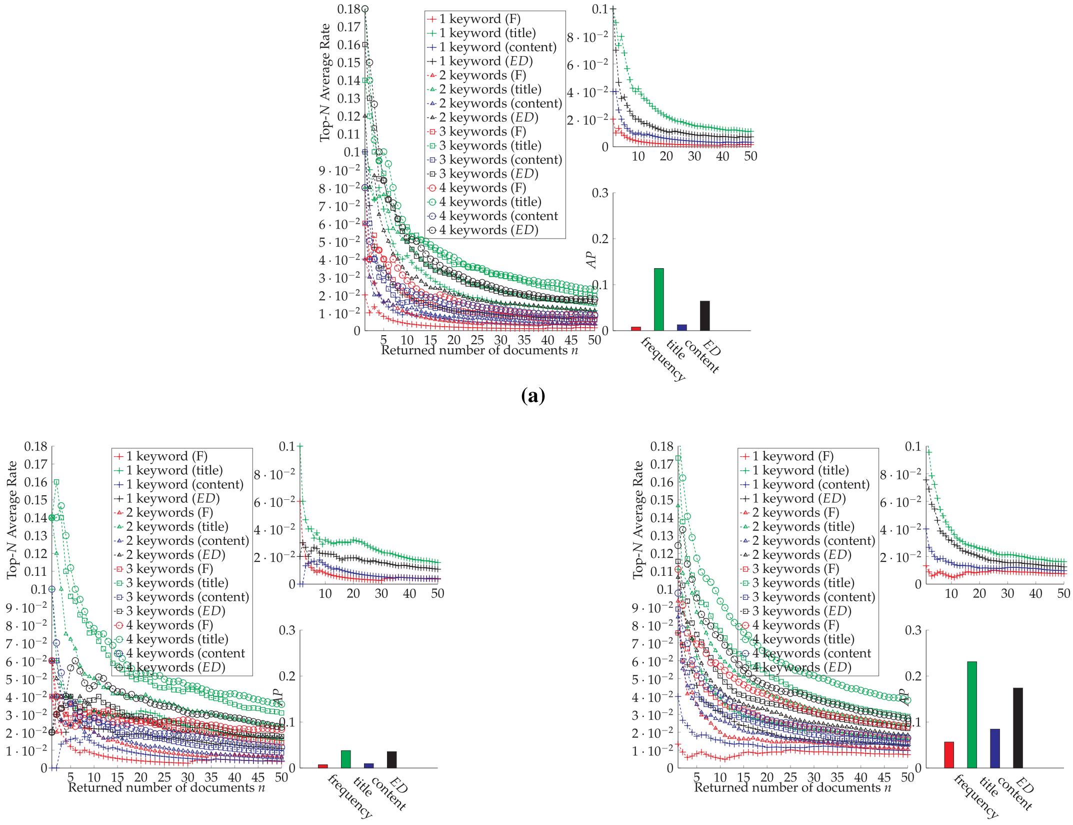


FIGURE 5. Diagram of computational complex varying with the number of keywords.

The experimental results are shown in Figure 4, we make some observations:



(a)

(b)

(c)

FIGURE 6. **Performance evaluation of keywords based retrieval in CIR. The Figure (a)-(c) show the top-*N* accuracy rate of the documents retrieval using Okapi BM25 with 1-4 keywords evidence(s) for each document using four kinds of keywords extraction methods (keywords were extracted from high frequency words (F), random words in title (title), and random words in content (content) and *ED* (*ED*)) in three data sets (TREC-LAT Corpus, TREC-FBIS and Cranfield Corpus). Inset (top): the same, but with only one keywords evidence. Inset (bottom): the same, but for average precision (*AP*) for one keywords evidence instead.**

- The general retrieval performance of Boolean retrieval model $<$ VSM retrieval model $<$ language model $\approx$ probability retrieval model.
- The Boolean model has proved to be unsuited for the document retrieval in the cloud although it is widely used in current encryption document retrieval scheme.
- Since the probability model (Okapi BM25) achieves the best performance in most cases, we use it as baseline in following experiments.

C. PERFORMANCE EVALUATION OF KEYWORD BASED INFORMATION RETRIEVAL

In this section, we use a entropy difference based metric (*ED*) [25] and its improvement to evaluate and rank the relevance of words in a text. Then we examine how the number of keywords affect the cloud computational complexity. Thereafter, we investigated how the different choices of keywords affect the final retrieval evaluation.

1) COMPUTATIONAL COMPLEXITY VS. THE NUMBER OF KEYWORD

The information security of CIR is guaranteed by the predefined encryption policy S, which is extremely time consuming and becomes the performance bottleneck of cloud service systems. We examine how the number of keywords affect the cloud computational complexity.

In practice, a symmetric or asymmetric encryption scheme, such as RSA, DSA, 3DES, AES, PKCS [10], [21], can be used to build the encryption policy S as it met the requirement of searchable encryption scheme [2], [11]. Taking RSA [16] as example, which is one of the first practicable public-key cryptosystem and is widely used for secure data transmission, X is plain keyword and Y is its corresponding ciphertext, $(p*q, b)$ is the public key, and (a, p, q) is the secret key. The binary length of X, Y, and $p*q$ is K, L and M, respectively. Thus the RSA encryption and decryption include the following operators, including $X > Y, X = Y, X < Y, X+Y, X-Y$, $X \cdot Y$, X/Y, $gcd(X, Y)$, $X^b \ mod \ p*q$, $Y^a \ mod \ p*q$, etc, and their computational complexity is $O(K)$, $O(K)$, $O(K)$, $O(K)$, $O(K)$, $O(KL)$, $O(L(K-L))$, $O(K^2)$, $O((\log b) \times M^2)$, and $O((\log a) \times M^2)$, etc. Obviously, as shown in Figure 5, with the increase of the number of keywords, the computational complexity of each operator rapidly grows with the highest at $O(K^2)$. Considering the communication efficiency and information security in the cloud retrieval scenario, data owners often need to share their data with a large number of users, hence less keywords mean less computationally complex, CPU computation, and security risk.

2) RETRIEVAL PERFORMANCE COMPLEXITY VS. THE NUMBER OF KEYWORDS

We investigate how different choices of keywords affect the retrieval performance. With probability model (Okapi BM25) as the basic retrieval model, we empirically compare four kinds of keyword extraction methods including keywords extracted from high frequency words, random words in title, and random words in content and *ED* in three datasets (TREC-LAT Corpus, TREC-FBIS and Cranfield Corpus). Figures 6(a)-(c) show the top-*N* accuracy rate of the documents retrieval using Okapi BM25 with the number of keywords (1-4 keyword).

- The keyword based on *ED* metric shows better performance than these methods based on keywords extracted from high frequency words and random words in content. It can achieve about an average of 10-15% accuracy rate in top-5 returned documents and 5-10% accuracy rate in top 10 returned documents, with less than four keywords for each document which is acceptable and suits application in documents retrieval in the cloud.
- Moreover, since all documents in three experimental corpora above have titles, we also compare keyword based retrieval using *ED* metric and titles, as shown in Figure 6, in most cases the *ED* metric based methods achieve comparable results with the title based methods. The titles have high relevance to the content of the documents, which suggests that the headline covers the main elements of the document. It also reflects the good performance of *ED* metric based methods. However, the titles are not always available in most practical applications, thus the *ED* metric based method is a better choice in practice.
- The inserted figures in Figure 6 show the top-*N* accuracy rate and average precision (*AP*) for all kinds of keyword detection metrics with only one keyword extracted from each document. The *ED* metric based method also shows better performance than these methods based on high frequency words and random words in content, and show comparable performance to the title based method.

In conclusion, in the CIR, the proposed keyword based retrieval methods show a better chance to respond to actual changes and needs in demand of retrieval in the cloud. Only with less than four keywords extracted from each document, we can achieve acceptable performance. However, as mentioned above, considering the cloud computing scenarios and security of communication protocol, users hope to obtain what he wants with less communication between the user and the CSP database. This means that users want to retrieve documents of interest based on a few or even one keyword(s).

VI. CONCLUSION

There is a fairly long history of trying to find methods to selectively retrieve files of interests from within large collections. Over the last 40 years, a lot of well known primary concepts and models in information retrieval have been developed [12], [26]. Modern information retrieval techniques have achieved great success particularly by popular online search engines. However, due to the rapid growth of internet usage, and decentralized computing, storage and management characteristics of modern information service, more and more sensitive information are being transferred to the cloud. Due to the lack of the mutual trust between the data owner and the cloud service provider (CSP), data usually have to be encrypted prior to outsourcing for data privacy and combating unsolicited accesses, which brings about tremendous challenges to data usage especially for document retrieval.

Efficiently retrieving encrypted files from cloud is a challenging task because of the need to achieve both high security and retrieval performance. We have proposed, in this study,a new retrieval system to address the challenge. To the best of our knowledge, we formalize for the first time the problem of effective keyword retrieve over encrypted cloud data while maintaining keyword privacy and retrieval performance. A wide range of experiments have shown that the proposed system can achieve superior performance in various settings.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY, USA: ACM Press, 1999.

[2] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *Advances in Cryptology—EUROCRYPT*, vol. 3027. Berlin, Germany: Springer-Verlag, 2004, pp. 506–522.

[3] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," *Theory of Cryptography*, vol. 4392. Berlin, Germany: Springer-Verlag, 2007, pp. 535–554.

[4] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 222–233, Jan. 2014.

[5] Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in *Applied Cryptography and Network Security*, vol. 3531. Berlin, Germany: Springer-Verlag, 2005, pp. 442–455.

[6] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[7] T. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[8] W. B. Croft and D. J. Harper, "Using probabilistic models of document retrieval without relevance information," *J. Document.*, vol. 35, no. 4, pp. 285–295, 1979.

[9] L. Fang, W. Susilo, C. Ge, and J. Wang, "Public key encryption with keyword search secure against keyword guessing attacks without random oracle," *Inf. Sci.*, vol. 238, pp. 221–241, Jul. 2013.

[10] S. Goldwasser and M. Bellare, "Lecture notes on cryptography," in *Cryptography and Computer Security*. Cambridge, MA, USA: MIT, Aug. 1999, pp. 1996–1999.

[11] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–5.

[12] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

[13] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

[14] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[15] M. Mitra and B. B. Chaudhuri, "Information retrieval from documents: A survey," *Inf. Retr.*, vol. 2, nos. 2–3, pp. 141–163, 2000.

[16] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, pp. 120–126, Feb. 1978.

[17] S. Robertson and H. Zaragoza, *The Probabilistic Relevance Framework: BM25 and Beyond*. Norwell, MA, USA: Now Publishers, 2009.

[18] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw Hill, 1983.

[19] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[20] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 1995, pp. 138–146.

[21] D. R. Stinson, *Cryptography: Theory and Practice*. Boca Raton, FL, USA: CRC press, 2005.

[22] Q. Tang and L. Chen, "Public-key encryption with registered keyword search," in *Public Key Infrastructures, Services and Applications*, vol. 6391. Berlin, Germany: Springer-Verlag, 2010, pp. 163–178.

[23] W. Waller and D. H. Kraft, "A mathematical model of a weighted Boolean retrieval system," *Inf. Process. Manage.*, vol. 15, no. 5, pp. 235–245, 1979.

[24] Z. Yang, I. Jones, X. Hu, and H. Liu, "Finding the right social media site for questions," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 639–644.

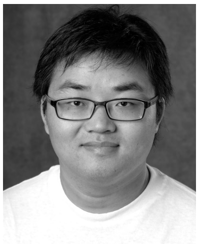
[25] Z. Yang, J. Lei, K. Fan, and Y. Lai, "Keyword extraction by entropy difference between the intrinsic and extrinsic mode," *Phys. A, Statist. Mech. Appl.*, vol. 392, no. 19, pp. 4523–4531, 2013.

[26] Z. Yang, C. Li, K. Fan, and J. Huang, "Exploiting multi-sources query expansion in microblogging filtering," *Neural Netw. World*, vol. 27, no. 1, pp. 59–76, 2017.

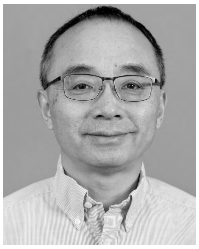
[27] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2001, pp. 334–342.

[28] C. Zhai and J. Lafferty, "A risk minimization framework for information retrieval," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 31–55, 2006.

ZHEN YANG (M'14) received the Ph.D. degree in signal processing from the Beijing University of Posts and Telecommunications. He is currently a Full Professor of computer science and engineering with the Beijing University of Technology. His research interests include data mining, machine learning, trusted computing, and content security. He has published over 30 papers in highly ranked journals and top conference proceedings. He is a Senior Member of the Chinese Institute of Electronics.



JILIANG TANG received the B.S. and M.S. degrees from the Beijing Institute of Technology in 2008 and 2010, respectively, and the Ph.D. degree in computer science at Arizona State University in 2015. He is currently an Assistant Professor of computer science and engineering with Michigan State University. His research interests include trust/distrust computing, signed network analysis, social computing, and data mining for social goods. He received the Best Paper Award of SIGKDD2016 and the Runner Up of SIGKDD Dissertation Award 2015. He was the Poster Chair of SIGKDD2016 and serves as regular journal reviewers and numerous conference program committees. He co-presented three tutorials in KDD2014, WWW2014, and Recsys2014, and has published innovative works in highly ranked journals and top conference proceedings that have received extensive coverage in the media.



HUAN LIU (F'12) received the B.Eng. degree in computer science and electrical engineering from Shanghai Jiaotong University and the Ph.D. degree in computer science from the University of Southern California. He is currently a Professor of computer science and engineering at Arizona State University. He was recognized for excellence in teaching and research in computer science and engineering at Arizona State University. His research interests include data mining, machine learning, social computing, and artificial intelligence, investigating problems that arise in many real-world applications with high-dimensional data of disparate forms, such as social media, group interaction and modeling, data preprocessing (feature selection), and text/web mining. His well-cited publications include books, book chapters, and encyclopedia entries and conference, and journal papers. He serves on journal editorial boards and numerous conference program committees, and is a Founding Organizer of the International Conference Series on Social Computing, Behavioral-Cultural Modeling, and Prediction.