Journal of Computer and Communications, 2015, 3, 90-98
Published Online September 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.39010

Improved Research on Fuzzy Search over Encrypted Cloud Data Based on Keywords

Ping Zhang, Jianzhong Wang

College of Fundamental Education, Sichuan Normal University, Chengdu, China
Email: 835148187@qq.com

Received 20 August 2015; accepted 25 September 2015; published 28 September 2015

Abstract

A search strategy over encrypted cloud data based on keywords has been improved and has presented a method using different strategies on the client and the server to improve the search efficiency in this paper. The client uses the Chinese and English to achieve the synonym construction of the keywords, the establishment of the fuzzy-syllable words and synonyms set of keywords and the implementation of fuzzy search strategy over the encryption of cloud data based on keywords. The server side through the analysis of the user's query request provides keywords for users to choose and topic words and secondary words are picked out. System will match topic words with historical inquiry in time order, and then the new query result of the request is directly gained. The analysis of the simulation experiment shows that the fuzzy search strategy can make better use of historical results on the basis of privacy protection for the realization of efficient data search, saving the search time and improving the efficiency of search.

Keywords

Cloud Data, Fuzzy Search, Keywords, Synonyms, Searchable Encryption

1. Introduction

With the arrival of the era of big data and cloud data, more and more ordinary users and enterprise storage in local data files, cloud storage service with convenient using, cost saving and the other advantages are welcomed by more and more users. But the uses of cloud storage services are also facing some problems. For example, some of the business secrets of the enterprise must be able to protect the privacy to prevent illegal use, so the data is generally encrypted locally and then outsourced to the cloud storage server, which brings a lot of trouble to the data use. Because of the limitation of network bandwidth and local storage capacity, users may not download all data to the local and then decrypt it before using it and how to improve the search efficiency etc. In the

How to cite this paper: Zhang, P. and Wang, J.Z. (2015) Improved Research on Fuzzy Search over Encrypted Cloud Data Based on Keywords. *Journal of Computer and Communications*, **3**, 90-98. http://dx.doi.org/10.4236/jcc.2015.39010

face of the huge amount of information decrypted and the need of accurate, efficient and personalized search information and other needs, domestic and foreign researchers have proposed a lot of search methods.

A fuzzy keyword search method for encryption data in cloud computing is proposed in the literature [1]. In the paper [1], a new technique based on fuzzy keyword search using the similarity between the key words of the edit distance is proposed. In the paper [2], a range search method based on the packet-based range is proposed. The architecture, mechanism and model evaluation of cloud computing security are described in the literature [3]. A privacy preserving method based on encryption cloud data is proposed in the literature [4]. The literature [5] presents a hybridization of Searchable Encryption and Attribute Based Encryption techniques. The proposed model supports a personalized and secure multi-user access to outsourced data, presenting high search performance. A secure and effective range query method based on R-tree for the outsourcing database is proposed in the paper [6]. In intelligent search, in the paper [7] a new technology of intelligent text search is proposed, including information retrieval, information extraction and information filtering. In the literature [8] the improved word segmentation algorithm and the improved correlation algorithm are proposed by analyzing the existing mainstream Chinese word segmentation algorithm and Lucene correlation ranking algorithm. A new intelligent cloud search optimization algorithm is designed in the paper [9].

A new search model is proposed in the paper [10]. The problem of low efficiency of large data query is improved and an effective search method based on large data is improved in the literature [11], which can improve the search efficiency. In the paper [12], a fuzzy search scheme based on key words is proposed, which realizes the search for the Chinese fuzzy tone and synonymous keywords and uses the pseudo random function to protect the private key. In a word, semantic search method and encryption of data search research have made fruitful achievements, but a safe and efficient fuzzy search method based on the key words fuzzy sets has not been designed. Therefore, in this paper, a fuzzy search strategy based on key encrypted cloud data on the basis of fuzzy search strategy over encrypted cloud data based on the keywords is proposed, exploring a method to improve the search efficiency in the cloud storage environment on the client and the server by different search strategies.

2. System Framework and Parameter Definition of the New Search Strategy over Encryption Cloud Data

New strategy based on the encrypted data search scheme in the client analyses fuzzy tone and polysemy characteristic of Chinese characteristics, uses Chinese and English to realize the synonym construction of keywords and establishes fuzzy tone words and synonym set of keywords, in the implementation of Chinese fuzzy tone and synonymous keywords search and the pseudo random function to protect the private key. When the user's query request reaches the server side, query request graph analyzed, keywords provided for user to choose, the key words that are used in the end are determined, the topic words and auxiliary words picked up. The topic words and auxiliary words have been matched with the historical inquiry in order, the query result of which is returned to the client.

System framework of fuzzy search strategy based on the keywords of the encrypted cloud data is shown in **Figure 1**. The framework consists of three main entities, which are data owners, users and cloud service providers. Among them, data owners are individuals or enterprise users, who store the collection of data files $F = (f_1, f_2, \cdots, f_n)$ in the cloud server. The keywords set associated with the file set F will be pre-defined and expressed as $W = \{w_1, w_2, \cdots, w_p\}$. In order to ensure that sensitive data is not used by unauthorized people, the data set *F* will be encrypted before the outsourcing to the cloud server. Due to the presence of a large number of sound words and synonyms in Chinese, in order to improve the cloud data utilization efficiency and success rate of retrieval, the system provides encryption data fuzzy sound synonyms and fuzzy search function. In the implementation of the service, the data owner will use search request to generate a private key PK to and distribute to the user, such as team members or employees. When the private key distribution is completed, for any input keyword *W*, in order to search relevant documents collection safely, authorized users use the private key PK and one-way generating function to convert query keywords into an encrypted search request and then submit to the cloud server. The cloud server executes the search without decrypting data and sends data collection FIDW related to a keyword *W* or *W* of fuzzy sound or synonym object files searched to searchers.

In this paper, we use the free cloud platform of the Amazon provided as the experimental platform and use the journal magazine as the search object to carry out the experiment. Through the comparison of this paper scheme

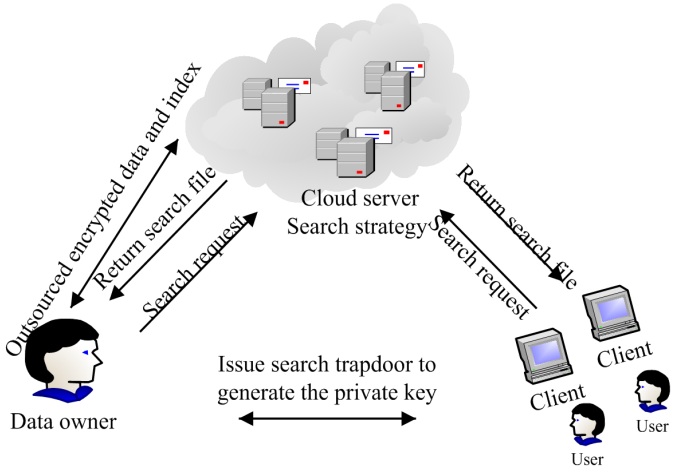

Figure 1. New strategy system framework for fuzzy search.

and the method of no using the fuzzy search technology, the validity and practicability of the scheme are verified. When the index this paper scheme is built, with the increase of the number of keywords, the occupancy of CPU and memory gradually increase and the time consumed increases. The time consumed for building an index is shown in **Figure 3**. The time consumption of the keywords query is shown in **Figure 4**. The success rate of the query is shown in **Figure 5**.

The search scheme still follows the security definitions involved in the traditional symmetric encryption. In addition to search results and search model, other related content connected with the file stored and its index should not be leaked.

The meaning and representation of the parameters in the system of the new search strategy over encryption cloud data are show in **Table 1**.

3. Implementation Scheme of the New Search Strategy over Encryption Cloud Data

The goal of fuzzy search is to return all the similar results to the keywords according to the keywords of different users. In this paper, we use the fractional step scheme to reduce the difficulty of fuzzy matching over the encrypted cloud data. First, data owners in the client construct fuzzy key words set, which includes Pinyin keywords, keywords and fuzzy tone, keywords corresponding English words, corresponding index information keywords and file ID table, Chinese and English keywords comparison table and pinyin and fuzzy tone comparison table. Second, based on the fuzzy keyword collection, a safe and efficient fuzzy search method is design, making full use of historical records to improve search efficiency.

3.1. Set up Fuzzy Pinyin Keywords Set

The establishment of keywords set is the precondition for the efficient fuzzy search. Keyword w is given and similar constraints d to generate $SP_{w,d}$ and SY_w. The following epsilon $w' \in SP_{w,d}$, meet $ED\left(PY_w, PY_{w'}\right) \le d$; $w' \in SP_w$, meet $syn\left(w\right) = syn\left(w'\right)$. The Chinese characters consist of pinyin initials and finals, as shown in **Table 2**.

Combination of consonants and vowels with specific rules, specifically, for example, consonant is m, the set of vowels only is {a, o, i, u, ai, ei, ao, ie, an, en, in, ang, eng, ing, ian, iao}, number of vowels fixed .The easiest error in Chinese pinyin spelling is fuzzy sound, such as initials in Alice retroflex retroflex peace, n and l, the anterior nasal and after nasal in the vowels, such as ian, iang , uan and uang. Therefore, the simplest way to set up the fuzzy sound keywords is to list the possible phonetic combinations and then to find out the keywords which are the same as those of the combination.

Assume d = 2 given by the user, the input pinyin of keywords w is nin, according to the alphabet constitute rules to generate the corresponding fuzzy sound keyword combination $SP_{w,d} = \left\{w_1', w_2', \cdots\right\}$, in which w_i' pinyin should be included in the set {nen, len, neng, in leng, *en}.

Table 1. The meaning and representation of the system parameters of the new search strategy over encryption cloud data.

Parameter	*Parameter meaning and representation*
F	A collection of files that are outsourced represents a collection of n data files $F=(f_1, f_2,\cdots, f_n)$.
W	A collection of different keywords that are extracted from the document collection F, representing a collection of p words $W=\{w_1, w_2,\cdots, w_p\}$.
I	Index for the establishment of a fuzzy keyword search for privacy preserving.
Tw	After user input search keywords w, the search request is generated by the one-way function, that is Trapdoor.
FID_{wi}	The file set F contains an ID collection of documents of keywords w_i or near w_i-syllable word or synonym.
f(key,·),g(key,·)	Pseudo random function (PRF), defined as {0, 1}* × key→ (0, 1).
Enc(key,·) Dec(key,·)	Symmetric key encryption/decryption function based on semantic security.
$ED(w_1, w_2)$	A description method of string similarity. For two words w_1 and w_2 in the two to achieve the increase of the minimum number of the required to convert, modify or delete the operation of the characters. For a given word w and integer d, with $S_{w,d}$ representing similar thereto word w', to meet the $ED(w,w')\le d$.
$SP_{w,d}$	The fuzzy Pinyin of keywords w corresponding to the keyword set. For the given Chinese keywords w and integer d, the Pinyin PY_w corresponds to the fuzzy tone set $SP_{w,d}=\{w'_1, w'_2,\cdots\}$, to meet with the edit distance of keywords w's Pinyin that is less than d for all similar Pinyin keywords set expressed as $ED(PY_w, PY_{w'})\le d$. For any $w'_i\in SP_{w,d}$, the pinyin of w'_i is represented as $PY_{w'_i}\in S_{PY_{w,d}}$.
SYw	A set of keywords synonymous with keywords w. different keywords set $SYw=(w'_1, w'_2,\cdots)$. Describing the same thing in a language are converted into another language, w'_i generally corresponding to the same keywords w_e, which is called set $SYw=(w'_1, w'_2,\cdots)$, as a synonym set for keywords w, among them the $syn(w)=syn(w'_i)$, syn() is synonymous conversion function. This paper uses Chinese and English to achieve the conversion.
Fuzzy keyword search	A collection $EF=(ef_1, ef_2,\cdots, ef_n)$ consisted of n encrypted data files and different predefined keywords set $W=\{w_1, w_2,\cdots, w_p\}$ are given, search keywords w and d input and after the implementation of the synonymous keyword search it will return to the ID file set {$FIDw_i$,}, in which $w_i=w$, $w_i\in SY_w$ or $w_i\in SP_{w,d}$.
History	The interaction between the user and the cloud server, which is composed of F and a group of keywords, is expressed as $Hq=(F, w_1, w_2,\cdots, w_p)$.
View	According to the secret key K, historical records Hq given, cloud servers can only see encrypted history that is the view of Vk (Hq). It includes index I of document collection f, query keywords trapdoor T'_{1w}. Among them $w'\in SP_{w,d}$ and T'_{2w}. $w'\in SP_w$, encryption file set C, is expressed as $\{e_1,\cdots, e_n\}$.

Table 2. The composition of Chinese characters.

Phonetic composition	*Set*
Initials	{b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w}
Single vowel	{a, o, e, i, u, ii}
Compound finals	{ ai, ei, ui, ao, ou, iu, ie, iie, er}
The anterior nasal vowel	{an, en, in, un, iin}
After nasal vowel	{ang, eng, ing, ong}

3.2. Set up a Set of Synonyms

There are a lot of synonyms or near synonyms in Chinese, and the words in the dictionary cannot be reflected, such as “计算机”, “电脑” and “微机”. These three Chinese words’ meanings are consistent with one another. File F associated with the three words should all returned to the user when users in cloud data search, but how to achieve similar synonym comparison has not a good way. Through the comparison of the expression similarity between Chinese and English in describing the same thing , the differential expression of language is proposed

to achieve synonyms conversion method, for example, the English words of three words above are all computer, so if the Chinese keywords w_i corresponding the same English language translation, so these words are synonymous.

If user input keywords w, executive function $syn(w)$, system translates w into English words w_e, then look up Chinese keywords whose English translation in bilingual table is w_e and return to the set of keywords SY_w. After generating the corresponding fuzzy tone and synonym set, encrypted function Enc(key) used, the encrypted $SP_{w,d}$ and SY_w ,together with the encrypted file are sent to the cloud to be saved.

3.3. Encryption Using Random Number Algorithm

The literature [13] design the random tree data structure providing cloud computing environment of privacy protection and the algorithm OPEART (order-preserving encryption based on random tree) is constructed. The algorithm by introducing a random realization of data encryption supports >, <, ≥ any relation operation of encryption data. Security analysis and performance evaluation in Literature [13] show that algorithm OPEART is indistinguishability under distinct and neighboring chosen plaintext attack and can be efficiently implemented on data encryption operations. In this paper, we introduce a secure OPEART algorithm to provide privacy protection in cloud computing environment.

3.4. Generate Search Requests

When the user enters the keyword w, the system performs a fuzzy search and returns the corresponding file ID collection $\{FIDw_i\}$. The formation of a search request generation process and keyword index are similar, namely according to the input of the w and d, fuzzy Pinyin and synonyms generating function to get fuzzy Pinyin keywords set $SP_{w,d}$ and synonym set SY_w. $SP_{w,d}$, w and SY_{wa} are encrypted to generate search trapdoor, submitting to the cloud server, namely search request generation are completed.

3.5. Fuzzy Search Scheme

In cloud service system, in order to avoid cloud access to sensitive information, establishment of a part of the work such as search index, the trapdoor generation in the client implementation. But executing search in large amounts of data is very consuming in resource, which can be handed to the cloud server to complete. The encryption cloud data fuzzy search implementation algorithm based on the keywords is as follows:

{

The data owner randomly selects two numbers a and b as the private key PK.

Establish index. The index $I_1 = \left\{ f\left(a, w_i'\right), Enc\left(PK_{wi'}, FID_{wi}\right)\right\}$, $w' \in SP_{w,d}$.

$I_2 = \left\{ f\left(a, w_i'\right), Enc\left(PK_{wi'}, FID_{wi}\right)\right\}$, $w' \in SP_w$, in which $1 \le i \le p$, Encryption key $PK_{wi'} = g\left(b, w_i'\right)$.

The index table I_1, I_2 and the encryption of data files to the cloud server storage.

The user input PK, w and i, the client system to generate $SP_{w,d}$, SY_w and generate the trapdoor $T_{1w} = f\left(a, w'\right)$, $w' \in SP_{w,d}$. $T_{2w} = f\left(a, w'\right)$, $w' \in SP_w$. T_{1w}' and T_{2w}' will be sent to the cloud.

The cloud server will receive the trapdoor T_{1w}' and T_{2w}', respectively, with the index I_1 and I_2 are compared to obtain a matching document ID set $\{FID_{wi}\}$, and send the results to the client.

The user uses the corresponding $g\left(b, w'\right)$ to decrypt the file ID and retrieve the required file, called function Dec(key) decryption and then used.

}

3.6. Increasing the Server Side Search Strategy

When the query request arrives at the server, system will provide similar query keywords for the user to choose and the final query keywords are obtained. The keywords are constructed, subject words and auxiliary words extracted. Subject words and auxiliary words are matched with history query in order. System firstly match the keywords of query request with history query results, after which the shared history query results and a new date for the query results can be combined, and the combination of history query results as a query results. If you do not match, the result is directly the result of the query, and the results are recorded in the history. The basic framework for cloud data search is shown in **Figure 2**. The other see references [11].

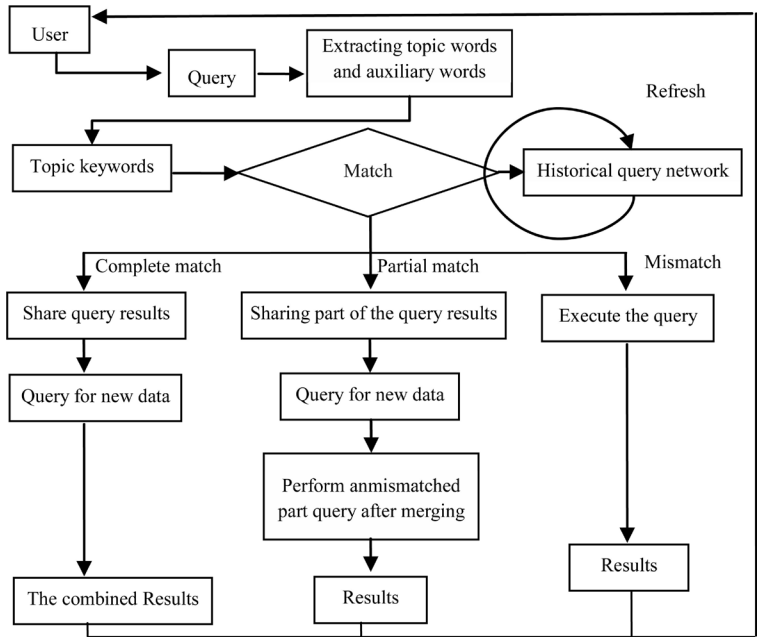

Figure 2. Basic frameworks for cloud search strategy.

The pseudo code of the cloud search algorithm is as follows:

{

The user proposes Query request. Provide similar keyword query for users to choose, users get a final query keywords. Keywords extracted topic keywords and auxiliary words.

System analysis query keywords, the query topic keywords and auxiliary keywords and historical query keywords match.

Execute different programs according to the matching results.

If the result is completely matched that it showed that the new query keywords have previously occurred, such previous queries obtained the query results can be directly to use for this query. That is to say, you can share the same history query results. At the same time, due to historical queries only for a certain period of time before the data query, there may be new data records generated after the history query, so a query still needs to be executed for the new data to get new results. The historical results and the new query results are merged to get the final result of the user.

If the result is partially matched that it showed that the key words of the user's new query request are extracted from the topic words and auxiliary words. The part of the theme words appeared before, so before the same query results through the query can be directly into the present query, that is, part of the history query results can be shared. Also due to historical queries only for a certain period of time before the data query, probably after the history query new data records may be generated, so the new data still need execute to obtain the new query results. The historical results and the new query results are merged to obtain the results of the query matching section. Then, in the result of the query matching part, it continues to perform the auxiliary word query and gets the final result of the user.

If the result is mismatched that it showed that the new query request keywords without historical records can be shared, the need to re-execute all the queries, the results obtained by the user.

Feedback the user query to the user.

To realize the update of historical inquiry network.

}

4. The Experimental Analysis of the New Method over Encrypting the Cloud Data

4.1. Security Analysis

When the user enters the same search request, the encrypted search scheme in the cloud will always return the

same search results. Although the cloud server cannot see what it is in the end, in the interaction with the user, it can still create access patterns and search mode. Therefore, the system's security should ensure that the other contents in addition to the pattern and search requests are leaked. The literature [13] has proved that the fuzzy keyword search scheme is in line with the non adaptive security requirements. Non-adaptive attack model only considers cloud server adversary, the existing literature proves that the adversary cannot choose search requests and previous search results based on trapdoor. This is acceptable, because only the user with an authorized private key can generate a search trapdoor.

In general, the security of this scheme is reflected in that for two historical records with the same trajectory, the cloud server cannot distinguish the views between them. In other words, the cloud server cannot extract more information from the information in the query, so the program is safe. The literature [13] has proved that the fuzzy keyword search scheme satisfies the non adaptive semantic security.

4.2. Time and Space Consumption

In this paper, the keywords fuzzy search scheme adds the cloud search method on the basis of literature [13], but time complexity and space consumption are in the same level with the original program. The time complexity of the index construction of the pre-processing stage is related to the number of keywords n. That is O (n). The size of the index is also related to the number of words. That is O (n). In searching, due to the cloud server system support for multiple threads, you can also search the synonyms; fuzzy syllable word retrieval without increasing too much time consumption, so search time complex degree is O (n).

4.3. Experimental Comparison

In this paper, we use the free cloud platform of the Amazon provided as the experimental platform and use the journal magazine as the search object to carry out the experiment. Through the comparison of this paper scheme and the method of no using the fuzzy search technology, the validity and practicability of the scheme are verified. When the index this paper scheme is built, with the increase of the number of keywords, the occupancy of CPU and memory gradually increase and the time consumed increases. The time consumed for building an index is shown in **Figure 3**. The time consumption of the keywords query is shown in **Figure 4**. The success rate of the query is shown in **Figure 5**.

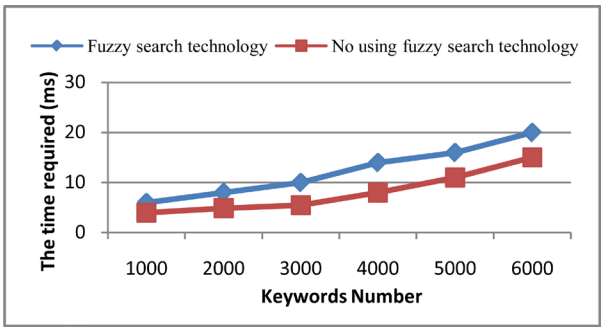

Figure 3. Time consumption for building an index.

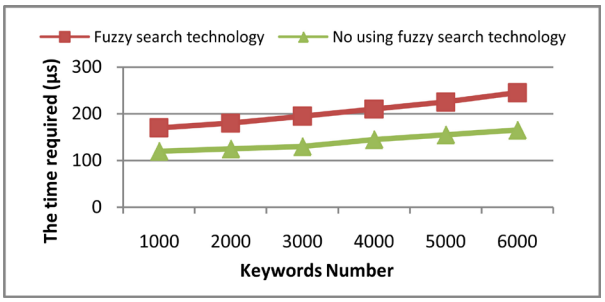

Figure 4. Time consumption of query.

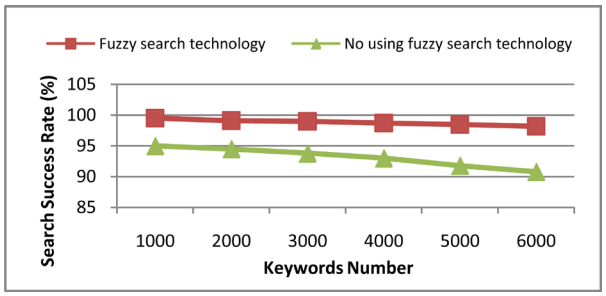

Figure 5. Keyword search success rate

From the above experimental results we can see, the storage to achieve storage of the synonyms of keywords and fuzzy speech sounds is appropriately increased through the search scheme, although the preprocessing and search time has increased, but it greatly improves the searching success rate.

5. Conclusion

In this paper, a new method based on keywords is presented according to the actual demand of the user's Chinese search in the cloud storage environment. Through the implementation of the scheme of fuzzy keyword search function, the problems of fuzzy sound and synonymous issues existing in the input text and the words that the user wants to find in the Chinese environment are easy to solve. By using the pseudo random function to effectively avoid the information leakage problem in the process of the query, it can ensure the safety. Therefore, this scheme has high security, good practicability and high search success rate.

Acknowledgements

In this paper, the research was sponsored by the Department of Education Project of Sichuan Province.

References

[1] Li, J., Wang, Q., Wang, C., Ren, K. and Lou, W.J. (2010) Fuzzy Keyword Search over Encrypted Data in Cloud Computing. *Proceedings of IEEE INFOCOM Mini-Conference*, San Diego, 15-19 March 2010, 441-445.

[2] Hore, B., Mehrotra, S., Canim, M., and Kantarcioglu, M. (2012) Secure Multidimensional Range Queries over Outsourced Data. *The VLDB Journal*, **21**, 333-358. http://dx.doi.org/10.1007/s00778-011-0245-7

[3] Lin, C., Su, W.B., Meng, K., Liu, Q. and Liu, W.D. (2013) Cloud Computing Security: Architecture, Mechanism and Modeling Evaluation. *Chinese Journal of Computer*, **36**, 1765-1784. http://dx.doi.org/10.3724/SP.J.1016.2013.01765

[4] Cao, N., Wang, C., Li, M., Ren, K. and Lou, W.J. (2011) Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data. *Proceedings of IEEE INFOCOM* 2011, Shanghai, 10-15 April 2011, 829-837.

[5] Tebibel, T. and Kaci, A. (2015) Parallel Search over Encrypted Data under Attribute Based Encryption on the Cloud Computing. *Computers & Security*, in press. http://dx.doi.org/10.1016/j.cose.2015.04.007

[6] Wang, P. and Ravishankar, C.V. (2013) Secure and Efficient Range Queries on Outsourced Databases Using R-Trees. 29*th International Conference on Data Engineering* (*ICDE*), Brisbane, 8-12 April 2013, 314-325.

[7] Wang, Z.Y., Xu,Y.R. and Guo, J. (2012) New Technologies of Intelligent Text Search. *Transactions on Intelligent Systems*, **7**, 40-49.

[8] Wu, J.M., Han, Y.H. and Ji, D.D. (2013) Research and Design of Search Engine for Digital Works Based on Lucene. *Computer Engineering & Science*, **35**, 166-172.

[9] Yin, Z. and Cao, J. (2012) Clouds Search Optimization Algorithm with Difference Quotient Information and Its Convergence Analysis. *Computer Science*, **39**, 252-255.

[10] Zhou, B., Liu, Y.Q., Zhang, M., Jin, Y.J. and Ma, S.P. (2011) Incorporating Web Browsing Activities into Anchor Texts for Web. *Information Retrieval*, **14**, 290-314. http://dx.doi.org/10.1007/s10791-010-9151-7

[11] Zhang, P. and Wang, J.Z. (2014) An Improved Efficient Search Method Based on Big Data. *Application Research of Computer*, **31**, 2331-2333.

[12] Fang, Z.J., Zhou, S. and Xia, Z.H. (2015) Research on Fuzzy Search over Encrypted Cloud Data Based on Keywords.

Computer Science, **42**, 136-139.

[13] Huang, R.W., Gui, X.L., Chen, N.J. and Yao, J. (2015) Encryption Algorithm Supporting Relational Calculations in Cloud Computing. *Journal of Software*, **26**, 1181-1195. (In Chinese) http://www.jos.org.cn/1000-9825/4656.htm