Chinese Word Segmentation for Agriculture

Kui Fang, Weiqiong Bu *(Corresponding Author)
Institute of Information Science & Technology, Hunan Agricultural University Changsha, P. R. China
Email:fk@hunau.net,buwq_molica@163.com

Wu Lou
Chinasoft International Co., Ltd, Changsha, P. R. China
Email: Louwu1985 @ 163.com

Lu-Ming Shen
Science College, Hunan Agricultural University, Changsha, P. R. China
Email: lum_s@126.com

***Abstract*—Based on the Hash mechanism, a new algorithm is presented, the algorithm can realize search, update, deletion and addition operations for dictionary. According to the characteristics of Chinese characters GB code, by preserving the GB code of first word in entry, this method effectively improves the utilization rate of the storage space. In the dictionary, the one-to-many corresponding relationships between dialects and agricultural keywords are built, the dialect words can be translated efficiently into the agricultural key words, so as the word segment accuracy is improved. In the time complexity, Chinese word segmentation algorithm for agriculture were compared with the algorithms for array, linked list and AVL tree.**

***Index Terms*—rural informatization, agricultural knowledge base, Chinese word segmentation, Hash function, word segmentation dictionary**

I. INTRODUCTION

During the 12th Five-Year Plan, agricultural informatization was intended to be the top priority of national economy and social development. Information technology used in agriculture has become a basic trend. Informational service is a major method to promote wider applications of agriculture-related science and technology for many countries. However, to farmers how to get valuable and interesting information from magnanimity information, that is one of hotspots in the field of agricultural informatization.

Knowledge Base is its central part of the agricultural information service platform. Chinese word segmentation[1] design is a key technology for the Knowledge Base Construction. For the agricultural knowledge base to build, in this paper, the design and implementation of agricultural professional segmentation is discussed.

There are three main ways for Chinese word segmentation design, Chinese word segmentation based dictionary Chinese word, the Chinese word segmentation based on the statistics and the Chinese word segmentation based on the understanding[2-3].

Chinese word segmentation based dictionary also known as mechanical word segmentation[4], including the forward maximum matching method, reverse maximum matching and minimum segmentation. The mechanical segmentation is not rely on the lexical, syntactic, and semantic knowledge, segmentation speed fast, simple, easy to implement, can be accurately cut separation of all word in the dictionary.

This way has a wide range of applications in a variety of Chinese information processing, The drawback is slow matching; intersection and combination ambiguity segmentation problem; word does not have a standard definition, there is no uniform standard word set; different dictionary ambiguity.

The advantages of this method is: not to be processed text field limit; do not need a machine readable dictionary; reduce the impact of unknown words, as long as there is adequate training text is easy to create and use. The drawback is that the need for a large number of pre-sub-word vocabulary, for the support, and training process in time and space overhead greatly.

The advantages of the based on statistical word segmentation method is[5]: not subject to the restrictions of the processing text ; do not need a machine readable dictionary; reduce the impact of unknown words, as long as there is adequate training text is easy to create and use. The drawback is the need for a large number of words pre-segmented support, and training process in time and space overhead greatly.

Chinese word segmentation based understanding is also called as knowledge segmentation, knowledge segmentation is an ideal method, it does not exist to the above problem, but the complexity of word segmentation algorithm is high, its effectiveness and feasibility need further validation.

In this paper, based on the advantages of the Hash table, a new algorithm of dictionary search based on the Hash mechanism is presented. The based on the characteristics of the GB code, it will save the storage space. By comparing a variety of segmentation algorithms in loaded, write, file size and operating time, the professional agricultural word segmentation accurate, and improve the utilization of space and time complexity.

doi:10.4304/jsw.8.5.1219-1226

II. Chinese Word Segmentation Dictionary for Agriculture

In the entry or retrieval of agricultural knowledge, how to accurately and efficiently extract keywords for agricultural knowledge, how to check the duplicate records in agricultural knowledge text base is the key technology.

Agricultural professional vocabulary is often too uncommon, common dictionary only collected a small amount of agricultural terminology[6], which resulted in the word segmentation is not accurate, it is difficult to accurately extract Agriculture Keywords. In this article, the fast and efficient translation of the dialect is also need to focus on issues.

2.1 Chinese Characters GB Code

GB2312 (1980) standard basic set includes a total of 682 graphic characters of non-Chinese characters and the 6763 Chinese characters.

The standard basic set is divided into 94 partitions, each district contains 94-bit. Using bit and position on the Chinese characters coding, known as area code. Each Chinese characters and symbols from two bytes to represent. The first byte is called “high byte” (Also known as partition byte), the second byte is called “low byte”(also known as bit byte). Suppose that high byte of a Chinese characters GB code is HighGB, and low byte is LowGB, then has following the relationship

HighGB=partition code + 20H
LowGB=bit code + 20H

2.2 Representation of Dictionary in Memory

According to the characteristics of GB code, using 6763 blocks corresponding to 6763 characters, which is the first word in a sentence. In computer memory, 6763 Chinese characters are represented by a two-dimensional array, each word as a unit in the array. The subscripts of the array are determined by partition bytes and bit bytes of each word, the computing method as follows:

Array subscript (high dimension) = partition byte-176
Array subscript (low dimension) = bit byte-161

In the dictionary, we must be aware of the fact that agriculture specialized vocabulary and agricultural dialect vocabulary is a one-to-many relationship. That is to say, many agricultural dialect vocabulary correspond an agricultural specialized vocabulary. In the memory, the structure of dictionary class as follows in figure 1.

CharacterArray Class: Chinese character indexes class, a two-dimensional array, index all the Chinese characters in the GB2312.

FirstWord Class: FirstWord: the first word,Num: the number of vocabulary with FirstWord as the first word; isWord: the vocabulary or not, wordProperty: the vocabulary attributes; hashTable: the list of all vocabulary with FirstWord as the first word.

Terms Class: Vocabulary or word line. wordName.
Mykeyword Class: Specialized terms.

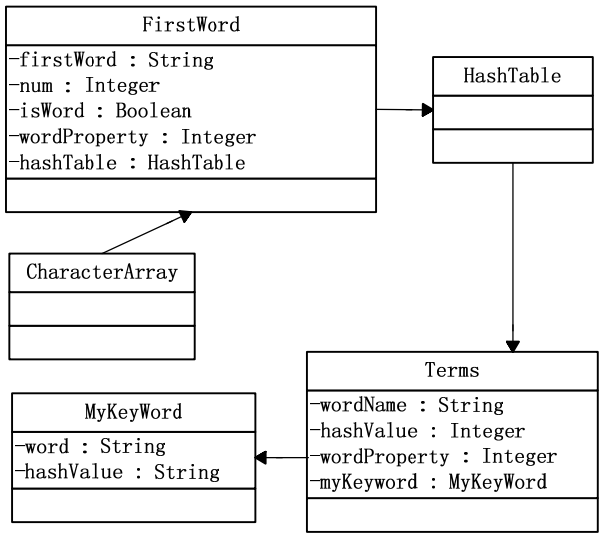


Figure 1. Structure of dictionary class

The intuitive performance graph of the dictionary in memory is shown as in figures 2.

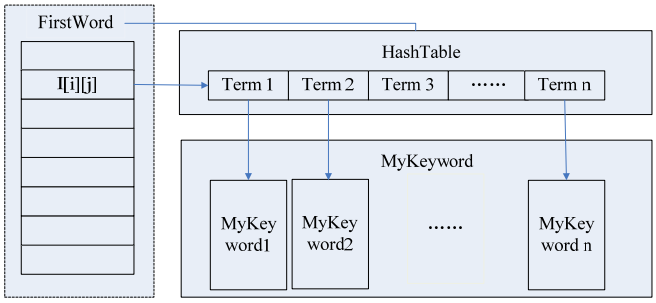


Figure 2. Representation of Dictionary in memory

2.3 Representation of Dictionary in Disk

The resident dictionary in memory can avoid loading the dictionary in the each segmentation. This way can improve the efficiency of the segmentation. The dictionary of Chinese word segmentation must be designed as a singleton, all the word segmentation method call with a dictionary object. The dictionary in Chinese word segmentation will be loaded into memory from disk, the results of increase, deletion and change for the dictionary will be saved into memory, In this process, memory and disk data synchronization. Meanwhile, the corresponding relations of the dialect words and key words will be established in the database, and will be saved in the agricultural knowledge base as the backup. The data in the database table must correspond with dictionary, so the need to create a function in the dictionary data and database data synchronization operation.

The representation of the dictionary in disk as shown in figure 3

```
Ah 2 F # ha 185254 # yowei 417365 1 #
A 49 F # laboren 570550 # site 407401 1 # ba 176214 # ma 19
Ai 7 F # ji 188176 # likesong 586643 1 # mier 377466 1 # fenbeige
```

Fig ure 3. Representation of Dictionary in Disk

In above figure, the first word in each line is Chinese characters commonly used in 6763. The first word of each line follows a numerical value, which indicates the number of entries led by the word. The symbol "#" is a separator. Other Chinese characters denote a string removed first word. "F" and "T" represents whether the vocabulary is agricultural vocabulary or not. Remove the first word.

III. CHINESE WORD SEGMENTATION DESIGN FOR AGRICULTURE

3.1 Hash Function for Dictionary

Hash is an important storage method[7], is also a kind of important research method. Hash search is a search algorithm more efficient way. In this paper, the system locate term through own design of the Hash function, thus realizing the dictionary lookup and update operation.

Suppose that the surplus word is n after the first word in term is removed, the I[i][j] denotes the high bit and low bit for ith word, then the Hash function can be expressed as

$$\text{Hash(Value)} = \sum_{i=1}^{n} (I[i][0]*1000 + I[i][1])$$

3.2 Dictionary Structure

In the design of dictionary structure, we mainly consider the following three aspects[8-9]:

(1) Space complexity;

(2) Time complexity for research;

(3) Time complexity of the translation of dialect words.

We consider above these aspects factors. The dictionary structure of system is designed as shown as in figure 4.

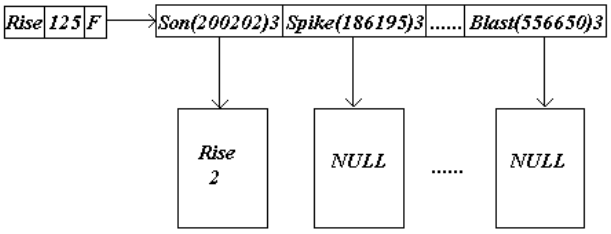


Figure 4. Dictionary Structure

In above figure, “rice” is the first word for example.:

(1) 125 denote that the dictionary exits 125 vocabulary headed by the word "rice"

(2) "F" indicates the "rice" word is a single vocabulary or not;

(3)”son”,” spike”,” blast” represents the vocabulary headed by the word “rice”,” ear of rice”, “Rice blast”;

(4) * is identification symbol. The numerical value 3, 2 , 1 indicates respectively the dialect word, the agriculture keyword , the common word;

(5) (200203) and (186195) denotes respectively the hash values of “son” and “spike”;

(6) The "son" uses the pointer to the "rice", said rice just dialect word, its keyword is "rice", thus, Thus, we have established a one-to-many relationship. If the word does not exist, pointed to the empty object.

3.3 Dictionary Construction

This dictionary sources from the Internet free dictionary,120000 entries, a small amount of agricultural vocabulary contained. The dictionary construction steps are as follows:

(1) Initialization, establish the dictionary object Dictionary, index object FirstWord, read the entries from the free dictionary;

(2) Read the dictionary a line, take the first word, and get the first word of the GB code;

(3) If FirstWord object in memory exits above the first word, then go to (5);

(4) The first word is added to the FirstWord object, and establishes the new Hash table object associated with it;

(5) After the first word is removed in entry, compute Hash value of the string, save in key of Hash table, and save the string in Value;

(6) Judge the attributes of vocabulary. 3, 2, 1 represents respectively the attributes value of dialect words , agriculture Keyword and common word;

(7) If the entry is dialect word, establish a new Mykeyword class, and save

the corresponding agricultural specialized vocabulary;

(8) Save the memory data to disk, form a dictionary.

3.4 Singleton Design Pattern

The singleton design is one of the many design patterns in the most widely used design pattern. The singleton is a creative model. It produced only one instance, and provided a global access point to access it. For some classes, Singleton is very important. For example, sometimes, the database connection or the Socket connection should be subject to certain restrictions, must maintain that only one connection exists in the same time. Singleton design pattern manage its unique instance through the class itself, this property provides a solution to the problem. Only one instance is a common object of the class, but the design of this class, it can only create an instance and provides full access to this instance.

The following code is a realization way for Singleton.

```
public class Singleton
   {
   private static final Singleton singleton = null;
private Singleton()
     {
     }
public static Singleton getInstance()
     {
if (singleton== null)
     {
singleton= new Singleton();
```

```
    }
return singleton;
    }
  }
```

In above examples, shows several features of Singleton:

Private constructor indicates that this class is impossible to form an instance. This means the class has only one instance. The all instances stored in its own the private members of the class. When take an instance, only need to use the Singleton.getInstance ( ).

3.5 Agricultural Entries Addition

The present Chinese word segmentation participle is mainly used in agriculture. In practice, agricultural specialized vocabulary is very rare, agricultural vocabulary in the dictionary is quite scarce, in the construction of agricultural knowledge text base, we design a professional field of key words and dialect, as long as these words are saved into the dictionary, then system can segmented accurately keyword and dialect of knowledge text base, with the growth of knowledge text base, the dictionary will be more and more perfect.

In view of the local dialect vocabulary, in entries Addition, the correspondence between dialects and agricultural key words is build, so as the dialect words can be translated efficiently into the agricultural key words (terminology). The entries adding generally occurs in the data preservation for knowledge text base, the procedure is as follows

Input: Agriculture dialect words and the corresponding agricultural Keywords;

Output: Addition succeed or failure

(1) To take the agriculture dialect words and the corresponding key words

(2) To Use entry lookup algorithm, If the entry already exists, change its attribute value, Otherwise, compute its Hash value, and save the entry , its Hash value and attribute value.

(3) To build the correspondence between dialects and agricultural key words, to set the dialect attribute value for 3, to establish a new Mykeyword class, to store the corresponding agricultural key words, to form a correspondence between dialect and agricultural key word.

3.6 Search Entry

The search algorithm is the most commonly used algorithm in the segmentation algorithm[10]. In ours algorithm, according the query string , get the GB code of the first word , this GB code directly as the index value can locate the entry line, then through the hash value of entry can do hash search, find after the hash search, synonyms search.

Input: a string to search for;

Output: the entry exists, returns an entry attribute value, otherwise it returns false;

(1) To take the first word of the input string word, calculate its GB code, get array subscript;

(2)According to the array subscript, we get the Hash table of all entries leaded by the first word.

(3)To computer the hash value of the string of entry removed the first word, If the hash value in Hash table, positioned directly to the string, then the string is a vocabulary, returns its attribute, otherwise, returns false;

IV. EVALUATION OF CHINESE WORD SEGMENTATION ALGORITHM FOR AGRICULTURAL

The algorithms for array, linked list and AVL tree word segmentation are realized. The time complexity for the search, insert, deletion, word segmentation algorithm for agriculture were compared with the algorithms for array, linked list and AVL tree, the results as shown as the following table 1.

TABLE1
TIME COMPLEXITY OF ALGORITHM WORD SEGMENTATION FOR AGRICULTURE

Algorithms	Search	Insert	Deletion
Array	**O(log n)**	**O(n)*O(log n)**	**O(n)*O(logn)**
Linked list	**O(n)**	**O(n)**	**O(n)**
AVL tree	**O(log n)**	**O(n)**	**O(n)**
agriculture segmentation agriculture	**O(1)**	**O(1)**	**O(1)**

The above algorithms operating rate as shown in the following table 2.

TABLE2
OPERATING RATE FOR THE ABOVE ALGORITHMS

Algorithms	File size/KB	Load/ms	Write/ms
Array	**834**	**710ms**	**340ms**
Linked list	**834**	**733ms**	**349ms**
AVL tree	**834**	**1350ms**	**790ms**
agriculture segmentation agriculture	**1210**	**750ms**	**304ms**

The operating time ratio of the word segmentation algorithm for agriculture were compared with the algorithms for array, linked list and AVL tree, the result as shown as the following figure 5.

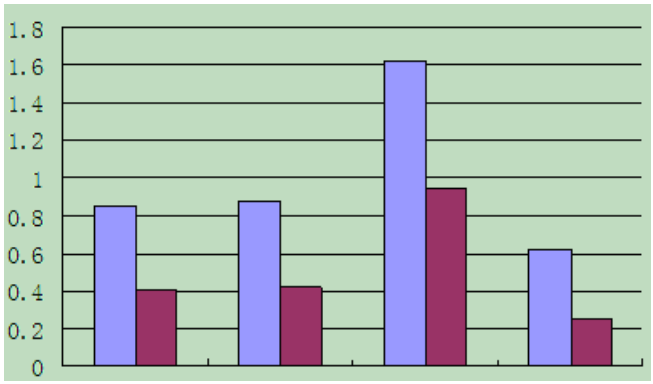


Figure 5. Operating Time Ratio

( ) indicates the load time ratio.

( ) indicates the write time ratio.

V. Information Retrieval

Information retrieval is through the traditional system retrieval technology to get the answer document set, the document set content to sort, and information extraction.

We will incorporate Chinese Word Segmentation for agriculture and the traditional retrieval system (Lucene system). The inverted index is established by using the Lucene system framework. For agricultural industry background and agricultural knowledge base, we design a sort algorithm to sort and display.

The retrieval sorting algorithm of agricultural knowledge base is as follows:

Definition 1 All sentences (problem) set is called as "the question word set", denoted by" C ". Word set " C " can be expressed as $C(t_1, t_2, \cdots\cdots t_n)$, where t_k is feature item, and $1 \le k \le n$.

Definition 2 The structural word set in agricultural knowledge base is called as" the record word set", denoted by" D ". The record word set " D " can be expressed as $D(t_1, t_2, \cdots\cdots t_n)$, where t_k is the feature item, and $1 \le k \le n$.

5.1 Keywords Analysis

The traditional retrieval system implemented by TF-IDF method, but the only limitation is to consider statistical characteristics of the words in the context. The higher the frequency of a feature item (key words), the feature item is more important, the feature item contains more representative of the topic. This is the meaning of TF (Term Frequency).

We usually think dialects, the keyword is more representative than the ordinary vocabulary; speaking from the part of speech, proper nouns is more representative than adverbs or other words.

In this paper, Vocabulary (commonly used words, dialect, professional keywords) are weighted by using the part of speech of words and word attributes, the algorithm feasibility and accuracy are increased.

Definition 3 Weight of feature t is expressed as Q_t . Suppose that the feature t is the professional vocabulary and nouns, given its weight Q_Z ;Suppose that the feature t is the dialect vocabulary and nouns, given its weight Q_f ;Suppose that the feature is commonly used words and nouns, professional words and dialect but not the noun, given its weight Q_c ;Suppose that the feature is the common vocabulary but not the noun, given its weight Q_o .

In this paper, TF-IDF value of the vocabulary $term_k$ is calculated as

$$TF - IDF(term_k) = Q \times \frac{f_k}{A_k} \times \log(\frac{S_k}{I_k})^2$$

where $TF - IDF(term_k)$ represents the TF-IDF value of feature item $term_k$ k, Q represents the weigh of the feature item $term_k$, the weigh is determined by definition3.The f_k represents frequency of the feature item $term_k$ in agricultural knowledge base, A_k represents the total number of words in Query data or knowledge base record, S_k represents the total number of records, I_k represents the number of the records with the feature item $term_k$.

5.2 Weight of Keyword Position

In this system, the knowledge in storage, in accordance with a question-and-answer format storage. This means that a word may appear different location, it can appear in different locations in the title, key words, the causes, the answers.

The amount of information carried by a word depends on the different positions which it occurs, the word appears in a different location, the contribution value to the search result is different, and therefore its importance is also different.

We believe that the importance of the different positions of a word (the amount of information carried by the word) sorted according to their position, in descending order: keywords, titles, causes, answer. Weighted vocabulary, we give weigh value by the positions which it occurs.

Definition 4 The word position is called as the word domain. In our study, the words can appear domain title 21 fields (fields of the record in the database). Each domain have different contribution for the retrieval, the weighted value of word position is defined as w.

5.3 Calculation of Match Degree

In agricultural knowledge base, the knowledge is input by agricultural experts.

Although the establishment of the inverted index, but the questions and answer sentence of agricultural knowledge base and will not normally very long. The matching degree plays a reference role for matching the keywords of questions.

In this paper, the positive matching degrees and the reverse matching degree are respectively defined.

We defined the positive matching degree from the angle of the questions.The positive match degrees measure that how much matching keywords of questions in knowledge agricultural knowledge base. The more the same keyword there are between the question word set and the record word set, the more closer the user demands the record is.

We defined the positive matching degree from the angle of the records. The positive matching degree: The matching degree between the record word set and the problem word set.

We defined the reverse matching degree from the angle of the records. The reverse matching degree is a measure of the main drift. The more the same keyword there are between the record word set and the question record word set, the higher relevance there is between the record word set and the question word set.

Definition 5 Suppose that the problem word set contains N_C feature terms, the total number of the same feature item between the problem word set and the record word set is N , the positive matching degree is expressed as P_C .The P_C is defined as

$$P_C = \frac{N}{N_C}$$

From above the equation, we know: If the record question matches the keywords of the question is more, then P_C is the greater.

The reverse matching degree: The matching degree between the problem word set and the problem word set.

Definition 6 Suppose that the record word set contains N_D feature terms, the total number of the same feature item between the problem word set and the record word set is N , the positive match degree is expressed as P_D , and the P_D is defined as

$$P_D = \frac{N}{N_D}$$

From above the equation, we know: If the question matches the keywords of the record in the knowledge base is more, then P_D is the greater.

The calculating formula of matching degree is follows:

$$P = a\frac{N}{N_C} + b\frac{N}{N_D}$$

where: $a + b = 1$, and

$$a > 0, b > 0$$

5.4 Weight of Record

In our evaluation system, the importance and authoritative of the records are evaluated. This relates to the three Impact factor: the status of expert, hits and comments.

The more well-known experts in the field the expert is, the higher authority the record has. The expert is a chief expert in the field, the authoritative is highest.

The hits and comments reflect the current hotspots, and can also reflect the importance of the record.

In the retrieval, we will weight to record. The calculation formula of record weight is as follows:

$$Boost(r) = G(r) \times \log_a(\frac{C(r) \times D(r)}{C(r) + D(r)})$$

where r represents record, $Boost(r)$ represents weighted value of the record; $G(r)$ mean the status of expert, 1—the chief expert, 0.75—the expert, 0.5—the general expert; a represents the weighting factor of the record, the value will be determined by the situation of the actual project implementation , and the default value is 100; the $C(r)$ indicates hits; and $D(r)$ expressed comments.

5.5 Calculation of Relevance

(1)Calculation method of text relevant

In traditional retrieval systems, the relevance is calculated by the vector space model. VSM(Vector Space Model) is proposed by G.Salton[14-15] , which is used widely.

In the vector space model, the query and the candidate documents are seen as a vector in the vector space model, the most similar document to the query vector is determined as the query result. The relevance between two space vectors is the Cosine value of the angle between two vectors.

Salton presents TF-IDF, which measures the weights of the feature term through their frequency, and is used widely. TF-IDF inherits the advantages and disadvantages of TF and IDF.

The text relevant is a common calculation method for the correlation. In this paper, text relevant is calculated by the vector space model

The calculation formula of the feature item weight of Questions is as follows:

$$W_C = P_C \times TF - IDF(term_k)$$

$$= P_C \times (\log\frac{S_k}{I_k})^2 \times \frac{F_k}{A_k} \times w \times Q$$

For feature in the record word set D , the weight calculation formula is

$$W_D = P_D \times TF - IDF(term_k)$$

$$= P_D \times (\log\frac{S_k}{I_k})^2 \times \frac{F_k}{A_k} \times w \times Q$$

We will substitute W_D the cosine formula of the vector space model, obtain:

$$\text{Cosine：} Sim(d,q) = \frac{d \bullet q}{\| d \| \times \| q \|}$$

$$= \frac{\sum_i (W_C(i) \times W_D(i))}{\sqrt{\sum_i W_C^2(i) \times \sum_i W_D^2(i)}}$$

(2)Normalization Calculation

The normalization calculation formula relevance of the problem and records is follows:

$$S_C = \alpha \sum_{k=1}^{n} TF - IDF(term_k) \cdot w_k + \beta P \times Boost(r)$$

$$= \alpha \sum_{k=0}^{n} Q \times \frac{F_k}{A_k} \times (\log \frac{S_k}{I_k})^2 \cdot w_k + \beta (a \frac{N}{N_C} + b \frac{N}{N_D}) \times Boost(r)$$

Where w_k is the weighted value of term_k position, $\alpha + \beta = 1$.

If we define

$$\alpha = n \frac{1}{n+N} \quad \beta = N \frac{1}{n+N}$$

Then

$$S_C = \frac{n}{n+N} \sum_{k=0}^{n} Q \times \frac{F_k}{A_k} \times (\log \frac{S_k}{I_k})^2 \cdot w_k + \frac{N}{n+N} (a \frac{N}{N_C} + b \frac{N}{N_D}) \times Boost(r)$$

Where n indicates the valid vocabulary in the question word set, N presents the total number of the same feature item between the problem word set and the record word set.

(3)Improvement of lucene Scoring Formula

In the Lucene retrieval system, the system measure query results by scoring candidate record. The higher the score, the record is closer to user demands, so the record will show also the better.

The Lucene scoring formula is only suitable for ordinary search. In this paper, we have improved the lucene scoring formula, and achieved good results.

The improved lucene scoring formula is as follows:

$$score(r) = P \times Boost(r) \times \sum_{k=0}^{n} (Q \times \frac{F_k}{A_k} \times (\log \frac{S_k}{I_k})^2)$$

VI. CONCLUSION

During the 12th Five-Year Plan, in china, the agricultural informatization was intended to be the top priority of national economy and social development. Information technology used in agriculture has become a basic trend. Informational service is a major method to promote wider applications of agriculture-related science and technology for many countries. However, the farmers how to get valuable and interesting information from magnanimous information is hotspots in the course of agricultural informatization.

Information technology services has been the main way to promote agricultural science and technology of agriculture in many countries, as it clearly became the important task in our social development and the national economy improvement during the 'twelfth-five', we keeps attempting the application of information technology in the agricultural informatization.

This paper is aimed at rural science and technology information integrated service platform design and implementation, according to the characteristics in agricultural field, design and realize the Chinese word segmentation. The experiment shows the Word segmentation is accurate, but also improves the utilization rate of the space and time complexity of word segmentation. The word segmentation can provide effective help for agricultural extension and agricultural science and technology information services, but still needs the unceasing improvement and the optimization, in order to improve the segmentation accuracy.

ACKNOWLEDGMENT

This paper is supported by the "Twelfth Five-Year" national science and technology support program (No 2011BAD21B03)and the scientific key planning project of Hunan province (No 2009NK4038).

REFERENCES

[1] Chu-Ren Huang and Nianwen Xue, "Words without Boundaries: Computational Approaches to Chinese Word Segmentation," *Language and Linguistics Compass*,Vol.6,pp. 494–505, August 2012.

[2] S. Maosong, and Z. Jiayan, "A critical appraisal of the research on Chinese word segmentation," *Contemporary Linguistics*, Vol.3, pp. 22-32, February 2001.

[3] P.Chuan Chang and M.Galley, "Optimizing Chinese Word Segmentation for Machine Translation Performance," *Proceedings of third Workshop on Statistical Machine Translation*, pp. 224－232, Ohio State University, USA, June 2008

[4] Yang Jiana and Zang Jin-song. "Two times backtracking Chinese word segmentation method," *Application Research of Computers*, Vol.26, pp.3320-3323, September 2009.

[5] H. holling and H, blank, "Rule-based item design of statistical word problems: A review and first implementation," *Psychology Science Quarterly*, Vol. 50, pp. 363-378, September 2008.

[6] Zheng Zezhi and Zheng Yongku. "Design and Fulfilling for Management System of Current Chinese Semantic Dictionary," Computer Engineering, Vol.12, No.6, pp.25-29, June 2001.

[7] Guo Xianghao and Zhong Yixin. "A Fast Algorithm for Chinese Words Automatic Segment Based on Two letters word family Structure," *J. of the China Society for Scientific and Technology Information*, Vol.17, pp. 352-357, May 1998.

[8] Li Qinghu and Chen Yujian, "A New Dictionary Mechanism for Chinese Word Segmentation Journal of Chinese Information Processing," *Magnetism*, Vol.17, pp.23-29, December 2003.

[9] Chen Guilin and Wang Yongcheng, "A Kind of Highly Efficient Data Structure for Chinese Electronic Thesaurus," *J. Computer Research and Development*, Vol.37, pp.82-89,January 2000.

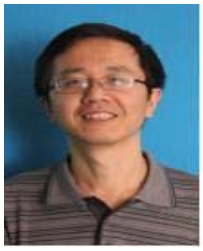
[10] Wu Shengyuan, "A New Chinese Phrase Segmentation Method," *J. Computer Research and Development*, Vol.33, pp.62-71, December 1996.

[11] R. Sproat and C. Shi, "A stochastic finite-state word segmentation algorithm for Chinese," *Magnetism, Computational Linguistic*s, Vol.22, pp.377-404, September 1996.
[12] Yang Erhong and Fang Ying, "The Evaluation of Chinese Word Segmentation and POS Tagging," *Magnetism, Journal of Chinese Information Processing*, Vol.20, pp. 44-49, January 2006.
[13] Y. Yang, "An evaluation of the statistical approaches to text categorization," *Information Retrieval*, vol.1, pp.76-88, February 1999.
[14] G. Salton, B. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol.24, pp.513-523, October 1998.
[15] K. Nigam and A. Mccallum, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, Vol.39, pp.104-134, April 1999.

Kui Fang was born in 1963. He received the Ph.D. in Computer science and technology from National University of Defense Technology of China, in 2000, and the M.S Degree. degree in Computational mathematics from Xi'an Jiaotong University of China, in 1985. He is now a professor of Computer science and Technology at Hunan Agricultural University. His major research interests include intelligent information processing, computer graphics,.

Weiqiong Bu was born in 1987. She is currently working towards the M.S. at Hunan Agricultural University.

Wu Lou was born in 1985. He received the M.S. Degree in Hunan agricultural University, China, in 2008. He is Lecturer at Hunan Agricultural University. His major research interests include intelligent information processing, computer graphics.



Lu-Ming Shen was born in 1973. He received the M.S. Degree in applied mathematics from Wuhan University, China, in 2001. He is currently working towards the Ph.D. degree at Huazhong University of Science and Technology. He is now an Associate Professor of applied mathematics at Hunan Agricultural University. His major research interests include Fractal geometry and its applications, computer graphics