# Retrieval Methods of Natural Language Based on Automatic Indexing

Dan Wang, Xiaorong Yang, Jian Ma, Liping Zhang

# Retrieval methods of natural language based on automatic indexing

Dan Wang[1,a,*]，Xiaorong Yang[1,b]，Jian Ma[1,c]，Liping Zhang[1,d]

[1] Institute of Agricultural Information，Chinese Academy of Agricultural Sciences，Beijing 100081，China Key Laboratory of Agricultural Information Service Technology（2006-2010）, Ministry of Agriculture, The People’s Republic of China

[a]wangdan01@caas.cn, [b]yangxiaorong@caas.cn, [c]majian@caas.cn, [d]zhangliping@caas.cn

**Abstract.** Since natural language enter the computer retrieval system，due to the natural language retrieval is not restricted by professional experience, knowledge background, retrieval experience by users, and above reasons favored by the users. As the title of the Chinese literature is the concentrated reflection of Chinese literature content, it reflects the central idea of the literature. Retrieval methods of natural language described in this article is limited to literature title in subject indexing. The basic idea of this method is, with automatic indexing methods respectively the literature title in the database of retrieval system used in natural language retrieval for automatic word indexing. To control the concept of a given keyword, namely meaning transformation, form the final indexing words. Then, using the vector space model for the index data in the database will be "or" operation to retrieve, forming a document set B. For each document title in set B for automatic indexing, the title of each article for automatic indexing, indexing terms for the formation and retrieval of natural language indexing terms similarity calculation, sorted according to similarity of each document in set B. The first best match the requirements presented to the user documentation. This method is a simple and practical method of natural language retrieval.

**Keywords:** Automatic indexing; Natural language Retrieval methods

## 1 Introduction

When the computer retrieval system has just entered the practical stage, people soon find its defects in retrieval time lag, retrieval feedback results, and to develop more convenient and efficient online retrieval system in the terminal. However, in the online era, the adverse effects of a full-time retrieval personnel and user questions needs have become a new reality questions. To this end, people have developed a variety of user-friendly man-machine interface. Today, access to the network retrieval, the user of retrieval system has undergone fundamental changes, the end-users who have different ages, different occupations, different knowledge backgrounds, different experiences have become increasingly demanding for retrieval system of convenience, immediacy and transparency. Thus, retrieval system (user interface) with the ability to understand natural language are increasingly welcomed by the majority of users, become an important part of the network retrieval system.

Natural language also known as "everyday language", it is a tool for expression and exchange of ideas in everyday life for the long-term social practice, it is very wide application in information retrieval. From the user's perspective, natural language search is users use the words, phrases or natural statement of daily life for questions. From the technical side, natural language retrieval is the natural language processing technology applied in information retrieval system of information organization and indexing, and output[1]. In information retrieval, the so-called natural language is relative to the case of controlled language, natural language is essentially a raw and standardized treatment of uncontrolled language. The whole process from the point of view of information retrieval, natural language search is including two aspects of natural language indexing and natural language query question. Natural language search is a direct order from the source document as the index identifies the content, users can directly use the natural language questions and complete a form to retrieve information retrieval.

## 2 Status of Natural Language Retrieval

Natural language retrieval was born in computer search, arising from the date it would have equal shares and information retrieval language. It is because natural retrieving language has its own advantages, it has long attracted people's attention, so that domestic and foreign scholars and experts to study it. Study abroad can be traced back in the 1960s, research focused on the automatic indexing achieve human indexing effect, the main representative of the study from the initial American scholar Salton and Bely, and later the American scholar Sparck, John, Tait, Fagan, Croft, Turtle, Lewis and so on; To the 1990s, TRC (Text Retrieval Conference Text Retrieval Conference) natural language search system began to participate in trials and competitions[2]. From TREC-1 meetings to TREC-6 meeting, the study of natural language search continues to move forward, its research focus on from the original statistical methods to the study of query expansion mode and flow index merge algorithm. The late 1990s, many foreign well-known databases such as Dialog, BIOSIS, ProQuest online, also started to provide natural language search interfaces in their own retrieval system and try natural language search. Many network-oriented information resource retrieval of test systems and the search engine uses a certain amount of natural language search technology, to a certain extent, to achieve a natural language search function, these test systems and search engines are: START、IRENA、FERRET、Ask Jeeves（http://www.ask.com）、Geoquery（http://www.cs.utexas.edu/users/ml/geo.html）、ixquick（http://www.ixquick.com）、Northern Light（http://www.northernlight.com）、Ask Northern Light a question、Electric library（http://www.elibrary.com）and so on.

Before the 1990s, in the domestic field of information retrieval research on natural language search in natural language indexing, other studies have concentrated on the theoretical discussion on the text indexing by natural language. After the middle of 1990, there have been some studies on the user interface. Professor Zhang Qiyu was the earlier focus on natural language search scholars, he made a more in-depth study

on a variety of factors natural language indexing information retrieval efficiency. He proposed text type, the search range, the search terms of the degree of specificity, the wording of the text is not standardized, different indexing methods and the degree of control of natural language search system would have an impact on [3]. National Taiwan University Department of Library Chen Guanghua used LOB Corpus as the training corpus, used SUSANNE Corpus as test Corpus, studied the natural language retrieval on the syntactic level [4]. In recent years, there are some practical web search engine to provide natural language search in China, there are TRS retrieval systems, Eureka search engine and Naxun Chinese news search engine. It is worth mentioning that the "Baidu knows" is the most influential of the Q & A platform - natural language search system. As of September 15, 2012, "baidu knows" " the use of natural language retrieval methods has solved the problem of 200 million [5].

## 3 The key issues of natural language search

Natural language search includes two aspects, namely natural language indexing and users retrieve by using a form of natural language questions. These two aspects can work independently, in technology implementation respectively, at the same time, they have close connection each other. The former is to standardize the indexing of natural language retrieval. The latter is to provide a natural language interface for users to ask questions, make information retrieval system to retrieve user needs to understand natural language in the form of expression, and processes the user's natural language questions. To solve these two areas, we depend on the following key technology research and development [6].

### 3.1 Subject Indexing.

One of the key issues of natural language search is how to extract most accurately fuller expression of documentation related to the topic words from the document. As well as the relationship between the words in a document expressing the theme concept, and this relationship is stored in the index, to support subsequent retrieval.

### 3.2 Question treatment.

Another key issue of natural language search is user's natural language questions by expressed understanding of computers. Ideal retrieval system should be able to "understand" the real search request which users use natural language expression. Not only retrieval system understands the significance of the user clear statement, but also understands the hidden meaning in natural language questions to be expressed. Thus, end users do not need to bother to go more express retrieval needs, learning tedious search command format.

### 3.3 Questions and index matching.

Ask and index of effective matching is another difficulty in natural language retrieval. Specific matching algorithm depends on the structure of the index and quiz process technology. Meanwhile, adopted retrieval model will largely affect the matching algorithm and effect.

### 3.4 Control Concept.

In essence, natural language search is a concept search, it requires a certain conceptual system or knowledge database to support. Knowledge of knowledge database can help solve the problem of natural language questions differentially expressed, that is a solution to the information source text and user questions related to the use of different words to express the concept of problem. The synonyms and near synonyms in the knowledge base can achieve control of the concept of words, to eliminate the difference of word brings the retrieval accuracy.

## 4 Natural Language Indexing

### 4.1 Factors affecting the quality of indexing.

In the information retrieval system, indexing methods and results have a greater impact on the retrieval results. Currently there is no indexing full-text indexing and word extraction indexing form keyword index of automatic indexing method. Quality of indexing directly affects the natural language retrieval results, influencing factors quality of indexing:

#### 4.1.1 Indexing depth

Indexing depth is used for indexing a document identification (keyword) number. It reflects the indexing on the degree of comprehensive and specifically to the theme of the document analysis. In general, the depth of indexing (indexing terms more) is the greater, the recall is the higher.

#### 4.1.2 The relationship between words of indexing words

The theme of the document is composed of multiple indexing words together common expression. There are certain grammar and restrictions relations between indexing words. For example, the position and the order of indexing words and relationship of index terms and synonyms and so on. In the process of automatic indexing, the more accurate analysis of the relationship between the word of the word, the more in the index expression, the complete retrieval result is better.

#### 4.1.3 Indexing size

During indexing, text block which indexing object points can be called indexing unit. Indexing unit refers to how large blocks of text to generate a set of index terms, it reflects the index object size. On the premise of indexical meaning, the smaller indexing particle size, the more sophisticated, the better retrieval refers specifically. Several factors described above is directly related to this article.

### 4.2 Automatic indexing method.

Automatic indexing is to use the computer to give corresponding to deal with literature searching, the process of indexing is divided into classification indexing and subject indexing. Subject indexing is divided into title lexical, lexical unit, syria lexical and keyword method. The first three indexing methods belong assignment indexing method index, the latter belongs to the extraction indexing words. Assignment indexing requires a thesaurus for support. This article relates to the automatic indexing is assigned indexing words [7]. This article use the keywords table is by the agricultural information institute, Chinese academy of agricultural sciences "computer automatic indexing research" compiled by the "computer automatic indexing multifunctional agriculture word", there are " use, generation, genera, divide and reference" and other relations between each item, can cut out from the document keywords conversion, automatic indexing given keywords, synonym, synonyms, related words, and the quest for word to complete the indexing concept of control. There are three methods in the general automatic indexing. The first one is segmentation method based on string matching, the second one is the segmentation method based on understanding, the third one is segmentation method based on statistical word [8]. Segmentation method based on string matching is Chinese character string and entry words in dictionary match according to a certain strategy, given an indexing word after a successful match. According to different scanning direction, it can be divided into forward match and reverse match. According to the different length matching of different priority, it can be divided into the biggest (longest) and the minimum (minimum) match.

#### 4.2.1Forward longest matching method（MM methods）

The strings obtained by coarse segmentation have been verbatim scanned from left to right and match with Thesaurus, and the keywords of thesauri maximum matching as the primary keywords. For example, in thesaurus in the "cadres tenure" in Chinese, and also included "cadres"、"office"、"age". Longest matching method is that "A short length is not taken" the word extraction rules, only extracting "cadres tenure" is used.

#### 4.2.2 Reverse longest matching method（RMM methods）

Principle of RMM with MM method is the same. Difference is that the word of the scanning direction, it is taken from right-to-left matching substring. Statistics show that simply using the forward maximum matching error rate is 1/169, simple to use reverse maximum matching error rate is 1/245. Obviously, RMM method in the segmentation accuracy than MM method has been greatly improved.

#### 4.2.3 Minimum Segmentation

Both forward maximum matching and reverse maximum matching, guaranteed indexing words maximum benefit of indexing specificity, is advantageous to the indexing specificity. But the biggest index terms are possible split in order to extract the smaller index terms. This is likely to increase the indexing words, that is to improve the indexing depth. At the time of retrieval, which is beneficial to avoid leak phenomenon. For example, in thesauri "cadres working age" term is maximum matching and indexing, if we the further use of minimum-cut method, there are likely to increase "cadres", "working", "age" indexing terms, indexing terms has increased, can improve recall.

## 5 Natural language search implementation

Natural language search is including natural language indexing and natural language query question two aspects. Automatic indexing of natural language are discussed earlier, the remaining problem is the problem of natural language questions and queries. Automatic indexing of natural language are discussed earlier, next we discuss natural language questions and queries and so on.

### 5.1 Natural language query processing.

One of the characteristics of natural language retrieval is to allow users to use natural language retrieval requirements directly to the system. Due to the use of natural language, in the form of natural sentences express questions and system index in the index entry form is different, the expression of document both cannot match directly, So we requires a necessary processing of natural language questions. Discussion of the relevant issues:

#### 5.1.1 Indexing item level

When you retrieve using natural language questions, retrieval system requirements index entries in the language level is completely consistent with the natural language questions. In this paper the method to realize natural language retrieval is limited to the title.

#### 5.1.2 The control of indexing words

Natural language retrieval has the characteristics of concept retrieval, semantic retrieval. Therefore, in order to achieve better search results, we want to deal with natural language questions in concept. Namely, the use of a number of related linguistic dictionaries, such as thesaurus dictionary or conceptual relationships dictionary or related knowledge, we need control natural language words in document title, at the same time we also need control the language (words) on the users of natural language questions. The key words in everyday-language are converted into keywords, and give the corresponding synonyms, hypernym, synonyms, etc. In this

paper, the natural language search implementation used a "Computer automatic indexing vocabulary multifunctional agriculture".

### 5.2 Natural language retrieval match.

In information retrieval systems, indexing process is the back-end processing, it participles and matches for the source document, etc. It has established the index data for retrieval. Natural language question processing is the front-end work of system. It also participles and matches for user's natural language question, etc. It provides the interaction between user and the system interface. In the process of natural language retrieval, indexing words between index data and retrieval requirements matched under certain matching control mechanism. It completed to match both generated indexing word. There are three common retrieval models that are Boolean model retrieval, vector space model retrieval and probabilistic retrieval model retrieval. Vector space model has a simple structure, formal, and easy to implement features. The vector space model retrieval was used in this article. It would document retrieval item (such as titles) and user natural language questions were compared. That is, both the indexing terms would be compared, and use weight to calculate the similarity, and to judge search results by similarity.

### 5.3 The concept control of natural language.

In dealing with natural language question retrieval, Pattern matching calculation is commonly used, namely by keywords (index terms) in comparison to complete. In the process, if it is mechanically simple pattern matching processing, instead of using the concept of the corresponding control, pattern matching will have the following questions.
Choice of words there is no strict limit of natural language questions, words more vocabulary more cluttered, it will affect the search results. In order to solve these problems, we must further process natural language question in the matching process, namely, we must control at the conceptual level. Natural language query express retrieval concepts through natural language. The concept of a keyword in natural language is all kinds of relations with other key words. For example synonyms relations, hypernym relations, etc. For example, "microbial fertilizer" concept, based on the relationship between the upper and lower classes can be subdivided into "antibiotic fertilizer", "Rhizobium fertilizer" "nitrogen-fixing bacteria fertilizer," and so on. A simple literal match will cause missed. For example the "computer" and "electron brain" is different expression of the same thing, "potato" and "toodou" is also synonymous. If the user's natural language is "potatoes" (or computer), the word "toodou" (or electron brain) is a term describing a document will be missed. This approach can apply these same concepts are words matching retrieval, thus expanding the retrieval surface, thereby increasing the recall.

### 5.4 Natural language retrieval method.

Following implementation of natural language search will be described, in addition, natural language search to achieve this article is limited to the title of the document (title) levels.

#### 5.4.1 Natural language search process.

Using natural language questions directly match and retrieval with the body or abstracts of the document, and in any case no one article is hit. Using natural language questions directly match and retrieval with the title of the document. Perhaps occasionally one article is hit.
Such as natural language search method is certainly not desirable. Therefore, we need automatic indexing process for the natural language questions and databases' document. Natural language retrieval process diagram is as follows:

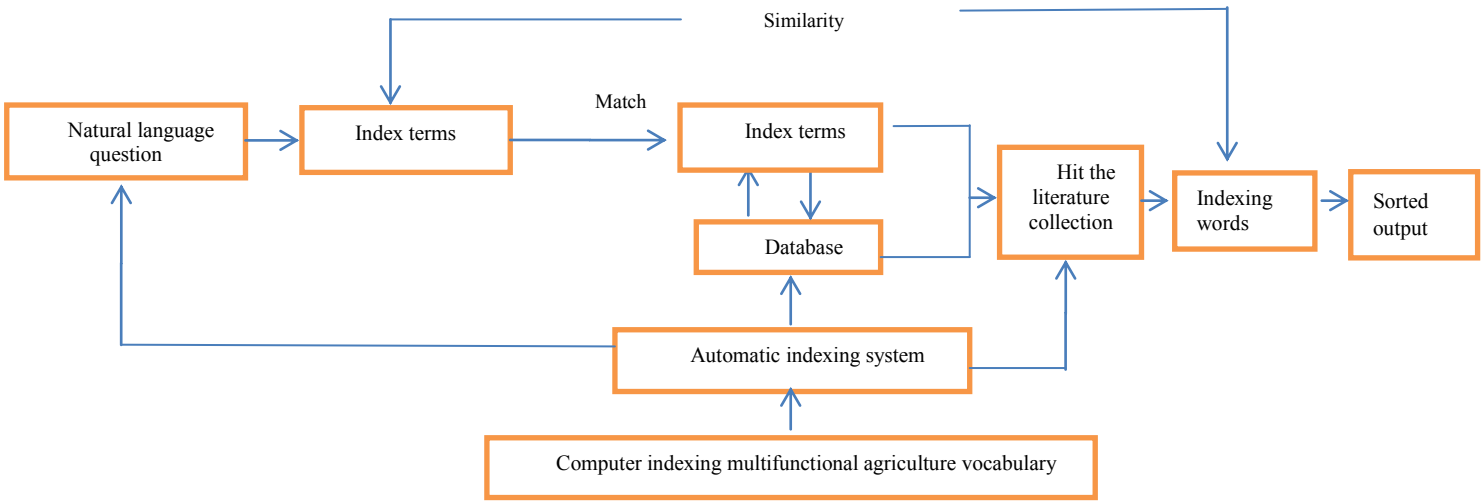


Figure 1 Natural language retrieval process diagram

First of all, the automatic indexing does automatic word segmentation for the title of the document and matches the words of thesauri. We will do meaning conversion process to index terms matched. We give index terms. While, we give the word hypernym, hyponym, synonym, synonyms or snare words as index terms. Then we do indexing process for indexing terms. This process is the systematic background does real-time processing to the database and the new information documents. To ensure that the title of the document should be indexing and index update for every new document in a database. In natural language retrieval, the system at the front desk also use the same automatic indexing system and indexing method to natural language question statement (or phrase) for indexing, and indexing words and the information source of indexing words "or" operation matching, matching literature that is hit literature, hit literature form a collection of hit documents, known as B collection.

#### 5.4.2 Hit literature output sorting.

Hit documents by the above methods are a number (n) articles in most cases. For example, indexing terms of natural language question in the statement are A, B, C three words, through A, B, C three words "or" operation to retrieve, any index terms of literature title containing A or B or C word, containing A and B or A and C or B and C words, or containing A and B and C words are retrieved. Faced with the hits literature (B collection), which was first presented to the user, but it has the sort of problem. Sort is based on the similarity of the title of hit document and the user's natural language question statement. Similarity calculation is as follows: Set the A string is indexing terms which natural language query statement provided, for a particular request statement, index terms of A string are fixed. In natural language retrieval, Set hit document is B collection, the title of each article in B collection is automatic indexed, and indexing words of every document form a C string, and indexing words of every document form a C string, so there are a number of C string in B collection, we compared A string with the C string, both indexing words overlap degree, we called similarity. The maximum similarity is 1, indexing terms of A string and indexing terms of C strings are one hundred percent the same. For each C string of the B collection, indexing words of C string and indexing words of A string one by one compares, calculate the ratio coincides both the index word (always less than 1). Thus, a characteristic natural language search results form a plurality of different sizes overlap ratio, Greater overlap percent, the higher the similarity. The coincidence percentage of high and low will be as the basis of the output sequence of the hits literature.

## 6 Test

Natural language retrieval was born in computer search, arising from the date it would have equal shares and information retrieval language. It is favored by the majority of users. Natural language search technology is widely used in utility database, news databases, etc. For example, in the field of agriculture practical technology database of crop management and cultivation techniques, plant protection technology, vegetable gardening management and technology and other aspects are more suitable for users to use natural language questions to ask questions for retrieval. "Baidu knows" is a very typical natural language question sentence retrieval practical database.

The author did a test by natural language retrieval methods in this paper. Test data is extracted from the CNKI database. Data mainly includes crop（wheat, corn, soybeans, cotton, potato, sesame seeds）cultivation techniques, pest control technology, gardening vegetable（cabbage, celery, tomato, carrot, rapeseed）cultivation techniques, livestock and poultry breeding, marine and freshwater aquaculture, and agro-processing, total capacity of data are more than 5,000. Automatic indexing thesaurus is Chinese Academy of Agricultural Sciences Institute of Agricultural Information compiled by the "Computer automatic indexing vocabulary multifunctional agriculture", a total of more than 40,000 keywords. There is a relation of the subject words, such as "use, generation, genera, division and reference", in addition to the category code and the corresponding net word. Using this method has

been tested on the above data, the test results are satisfactory, which is a higher precision. For example, the search statement as "potato cultivation technology", from the test sample data （there are 420 potato literature）and find relevant literature. According to the provided similarity method, followed by the top ranked "potato cultivation technology", "black potato cultivation technology", "precocious potato planting technology", "Heilongjiang province potato mechanized technology ", "potato cultivation comprehensive disease prevention and control measures." and so on. Because the automatic indexing words above topics and natural language search statement are "potato, cultivation techniques, planting techniques, precocious, detoxification, Heilongjiang Province, mechanization, disease, comprehensive prevention and control measures" and "potato, cultivation technology, planting, technology. " According to the degree of overlap between the two indexing terms will have above order. Another example: Search statement was "practical agricultural technology database" only was retrieved two documents in CNKI database. That is, "the WEB retrieval of agricultural practical technology database", "the regional agricultural practical technology database under the network environment". If we use the natural language retrieval method provided by this paper, we can find out the 50 documents. Because it is not a complete word with the search statement to find, but with its indexing word "agriculture, practical technology, database," a "or" operation to retrieve. Of course, so find out more literature, this is why this method will not result in a leak.

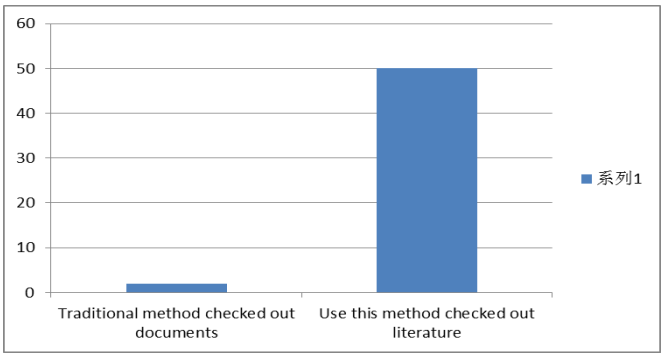


Figure 2 Document detection number comparison

However, sometimes, check out the document number is too much, in order to save time for the user to browse the documents, the maximum number of documents and the maximum number of views in the literature will be set up. Literature amount exceeds this value no longer that shows the document content. The documents are sorted according to the similarity. In order to meet the same degree of similarity in the literature, the order is also random. In addition, using this method to retrieve must use the same indexing system and thesaurus database indexing and index, in order to ensure the accuracy of the document similarity. In the requirements of retrieval statement, both keywords (keywords), a phrase, or spontaneous statements is the same, the method is applicable.

## 7 Conclusion

This article provides a natural language search method based on automatic indexing is a method of natural language retrieval application method easier to achieve. This method has some advantages: （1）Simple user operation, the search condition (statement) requiring a simple and different level of users can operate; （2）System implementation easier. Systems involved in the automatic indexing algorithm module, indexing data module and so on, most of them have more sophisticated software, it is easier to integrate them together into a useful natural language search system.（3）Made some technical standardization process in natural language processing, such as between words of "use, on behalf of, divided, the Senate" and other relations, given the appropriate indexing terms, to improve the documentation of precision and recall.（4）The introduction of the concept of literature similarity. The high degree of similarity literature in the front, the user feels natural language search improved precision. In summary, this method is a simple, practical and ideal natural language search methods.

## Acknowledgment

Funds for this research was provided by Technology Innovation Project of Chinese Academy of Agricultural Science.

## References

1.Tang Yanli, Lai Maosheng. Study of the Application of Ontology in Natural Language Information Retrieval[J]. New Technology of Library and Information Service.2005,120(2):33-36
2.Geng Qian,Tang Yanli. Web Information Oriented Natural Language Retrieval[J]. Information Science, 2004,22(7):845-849
3.Zhang Qiyu.Impact of Various Elements on Retrieval Effectiveness in Natural Language Retrieval[J]. Information Studies:Theory & Application, 1997,(5)：257-259.
4.Chen Guanghua. Information retrieval queries of natural language processing [DB/OL].www.cmgt.ntut.edu.tw/cimet/92active/ntut/EC/pages/pdf/200200023_MainFile.pdf, 2006-04-20
5.Jia Jia，Song Enmei，Su Huan. Research on Assessment of Answer Quality in Social Q&A Platform. Journal of Information Resources Management, 2013,(2):19-27.
6.Geng Qian, Lai Mao-sheng. The Pivotal Issues and Implementation of Natural Language Information Retrieval[J]. Information Science, 2007,25(5):733-741.
7.Wang Dan,Yang Xiaorong. Study on Elimination Method of Ambiguous Words in Chinese Automatic Indexing[J]. Library and Information Service, 2014,58(5):93-97.
8.He Xin，Wang Wan-wu. Research and Application of Chinese Word Segumentation Technical Based on Natural Language Information Retrieval[J]. Information Science, 2008,26(5):787-797.