Study on the Optimization Design of the Subject Indexing Based on the Word-frequency Statistics

Huafeng Xie

Exploration & Production Research Institute, SINOPEC, Beijing 100083, China

E-mail: raobian@126.com

Fang Wu & Xuying Lu

Dezhou Vocational and Technical College, Dezhou 253034, China

E-mail: wfang421@163.com

Abstract

Based on the traditional word frequency statistical function, the new weighting function is established in this article, comprehensively combining with four important factors such as the weight value of subject words, the classes, the specificity, and the cohesion relation. This new method could standardize the indexing of the subject words of the official document, enhance the work efficiency, realize the automatic indexing, and reduce the mistakes because of personal factors. In addition, the program design and the implementation of the computer language of this method are also introduced in this article.

Keywords: Official documents, Subject words, Word frequency, Weight, Nonlinearity

1. Introduction

At present, there are many problems in the subject indexing of the official documents (Bun, K.K., 2002, P.73-82). First, the title would be segmented as many phrases to be the subject words. This indexing method which replaces the subject words language by the natural language and replaces the concept coordination by words violates the most prominent basic principle that the subject words must be standardized strictly, i.e. the concept coordination. Second, many subject words are not standard. For example, some subject words contained verbs and empty words, and even some abbreviations and symbols were selected as the subject words, which disobeyed the range of the subject words class control, i.e. the subject words could only adopt terms and noun phrases, and adjectives in special situation. Third, multivocal words were used for subject words. The polysemy of the subject words could easily induce ambiguity and fuzziness, disobeying the principle that the vocabulary and the concept must be corresponding one to one. Fourth, many man-made operations were not proper. Many persons who index the subject words in the official documents could not understand the attention of the subject words, and didn't know the structure relation among subject words, so the indexed subject words could not fully express the content of the topic, and the indexing meaning was not obvious. The selection of the official document subject concepts should be determined by the indexing demand and the practical situation of the concrete department. According to the property and task of the official document, the emphasis and the indexing depth of the subject indexing should be confirmed, and the most proper and important subject words should be selected based on that. Aiming at the deficiencies of the subject indexing in the official documents, based on the word frequency indexing and other advantages of other indexing methods, other measures are adopted to compensate the word frequency indexing, which could realize the automatic indexing of computer, enhance the indexing efficiency of the official documents, and reduce the mistakes because of man-made factors as more as possibly (Chen, 1990).

2. The method of word frequency statistics

The automatic retrieval of the subject words is to extract those characteristic words which could express the central content of the text, i.e. those comprehensively characteristic words. In the subject processing, it will be very nice that the system could simulate human thinking, make clear the sentence structure and syntactic structure of each sentence, and analyze their meanings, but at present, it still is being explored for the syntax analysis of the words and sentences in the article, and the practical effect is bad. But the official documents are standard, and the words at different positions are regulated clearly, for example, the size, the character style, and the color of the words in the title should perform relative regulations, and other parts such as the abstract and the text should also follow certain rules, and in this way, the weights of the position and the paragraph of the subject words should be confirmed based on that. This method is very simple and feasible, and though the word frequency is very important for the subject indexing, but it is still not enough to only use word frequency as the indexing standard. In fact, in the indexing of word frequency, only the small parts (high-frequency words) in the front are reliable. Aiming at these deficiencies in the traditional word frequency statistics, the word frequency will be the base to enhance the accuracy rate of the official document indexing, combining with the weight value design, the classification, the specialization, and the cohesion relation (Baayen, 2006, P.281-291).

2.1 Weighting coefficient

To exactly grasp the topic of the original article, it needs a set of reasonable weighting system. And this system should include following weighting factors.

(1) Word frequency

In the weighting processing based on the word frequency statistics method, the word frequency is one of the most important factors influencing the weighting coefficients. Generally, in same one article, the words which occur more times could better reflect the topic and the content (except for empty words) of the article than those words occur less times (Kang, 2006, P.1993-1994).

(2) Word position

The word position is another one important factor to determine the word importance. The position of words could be title, abstract, and text. Generally, the words in the title could better reflect the topic content of the article than the words in the abstract, and for those words in the abstract, the words in the head of the paragraph and conclusions are more important than the words in the text (Barnes, 1978, P.107-119 & Trotman A, 2005, P.243- 264).

(3) Specialization (the length of word)

In Chinese, long words always are used for reflecting concrete and hypogynous concepts, and short words are used for reflecting relatively abstract and epigynous concepts. For example, for the words including “technology”, “computer technology”, and “large computer technology”, their specializations strengthen gradually, but their generalities weaken gradually. The retrieval of the subject words should select those words with strong specialization as more as possible (long words generally should not exceed 10 characters).

2.2 Design of the weighting function

According to the weighting function proposed by Han Kesong & Wang Yongcheng (Pu, 2007, P.1394-1396 & Han, 2000, P.651-653), *W (w)* represents the weight of word, *Fre(w)*, *Pos(w)*, *Len(w)* and *Val (w)* respectively represent the word frequency, the position, the length, and the value of *w*. The weight value when the subject word *w* occurs at the *i*’th time is

$$W(w, i) = Pos (w) * Len (w) * Val (w) \quad (1)$$

$$W(w) = \sum(Fre (w) * W(w, i)) \quad (2)$$

$$Fre(w) = i/(i + 1) \quad (3)$$

According to the experiences, *Len (w) = a/(a+1)*, where *a* is the word length.

According to the statistical experience, the position factor *Pos (w)* is defined as the piecewise function, and when the word occurs in the text, its value is 1, and when it occurs in the title, its value is 5, and when it occurs in the abstract, its value is 3, and it occurs in the head or the end of the paragraph, its value is 1.5, and when the word occurs in the first paragraph and the ending paragraph, its value is 2.

3. Improved measures of the weighting function

3.1 Factor of the paragraph amount

The retrieval of the topic is to extract those characteristic words about the content of the full text, and one of prominent difference with local characteristic words is that these words would occur in many paragraphs of the text. Therefore, the weight value of the characteristic word should be the function of the number of the paragraph that these words occur. The length factor is determined by the length of the word, and when the word is composed by two Chinese characters, its value is 1, and when the word is composed by three Chinese characters, its value is 1.5, and when the word is composed by four to five Chinese characters, its value is 2, and when the word is composed by above five Chinese characters, its value is 3. That the length of one word is n times than the length of another word doesn't mean the importance of this word is n times than the importance of that word. Therefore, the word length factor should be considered in the nonlinear function, and the word length factor is defined as Len(w) = a/(a + 1), where a is the word length. Because of those characteristic words in the local or small range, the factor of the paragraph number should be added necessarily. According to the experience formula, the factor of paragraph number should be defined as the nonlinear function word, i.e. Pars (w) =m/(m + 1), where m is the paragraph number. And the formula (1) should be

W(w, i) = Pos(w) * Len(w) * Val(w) * Pars(w) (4)

3.2 Restatement cohesion relation and concept extension

In the official document drafting, secretaries often neglect the polysemy, the synonymy of multiple words, and the semantic associations among vocabularies, which could influence the retrieval and judgment of the topic of the text, and a series of words might be thought as the words with same concept in the restatement cohesion relation. The concept extension means to extend the exterior association of subject words, for example, "network", "LAN", "MAN", and "WAN" have the relationship of concept extension. The keywords set produced by the glossary cohesion and the concept extension for each subject word could be regarded as one concept unity, and if the words including a1, a2… aj denoting the concept C have the restatement cohesion relation, and at the same time, the words denoting this concept b1, b2… bk have the relationship of concept extension, so F(C) = j + k (F(C) denotes the occurrence probability of the concept C). The formula (3) should be Fre(w) =(j+ k)/(j + k + 1), and when computing the weighting values, this formula should be used to replace Fre(w) in the formula (2). To be simple and denote the difference with the formula (3), this formula should be denoted as

Fre(w) = i0/(i0 + 1) (5)

According to the formula (4) and the formula (5), the new weight value which is different with the weight value in the formula (2) could be computed, and Fre (w) and W(w, i) respectively represent the values in the formula (5) and the formula (4), i.e.

W(w) = Σ(Fre (w) * W(w, i)) (6)

3.3 Emphatic pattern

For those sentences containing the topic clue strings such as "the key is", "be aimed at", and "the main objective is", the subject words contained in them should be endowed bigger weight values. If the subject word q occurs in these emphatic patterns, the weighting formula should be

W(w, i)=Qval+W(q) (7)

"Qval" is the weight of the subject words occurring in these sentences, and "W(q)" is the weighting function of the formula (4). In some official document classes, to denote the importance of the subject words in the emphatic pattern, the value of "Qval" could be set as the weight of the subject words in the title.

3.4 Classification of the subject words

For most articles, it is ideal to use the computer to automatically classify these documents. There are many computer classifications. However, the official documents are different with other literatures, and when even before the secretary drafts the document, he should know which class that the official document belongs to, and the subject indexing should be completed before the document is sent, i.e. before the document is flowed. The official documents could be classified by the medical class, the mechanical class, and the scientific technology class. According to the classification of the official documents, the subject words belonging to this class should be abstracted as more as possible, and other words should be dealt with the minimum weights, which could effectively improve the indexing.

First, when the official document only belongs to one official document class GL1, the words which all belong to this class in the official document should be weighted, and the weights of those subject words which don't belong to this class should be reduced. The indexing words obtained by the secretary when he drafts the official document should be endowed by higher weight values, so the subject words except for GL1 are difficult to enter into the selected indexing words set, and the quality of subject indexing could be guaranteed.

Second, when the official document belongs to multiple official document classes (above two classes) (GL1, GL2 and GL3), only the subject words belonging to GL1, GL2 and GL3 should be indexed.

Third, for the setting of the class weight, after the official documents are classified, the subject words should be redistributed by weight values according to the new attributes after classification, based on the original weighting function. Taking 6 official document classes (from GL1 to GL6) and four subject words (from W1 to

W4), a matrix M could be established as seen in Table 1, and the left column of the matrix includes all words occurred in certain one sample article. The default value in the matrix is 0, and Mij denotes whether the word Wi belongs to the class GLj, and if it belongs to the class GLj, the default value is 1, or else, it is 0. As seen in the matrix M, W1 belongs to GL2, and W2 belongs to the class GL1, and W3 could occur not only in the class GL2, but also in the class GL4, and W4 belongs to four classes (GL2, GL3, GL5 and GL6).

The matrix P is established from the matrix M (seen in Table 2), where Pij= mij / $\sum_{j=1}^{n} mij$. When the value of Pij is more close to 1, the monopoly that Wi belongs to the class GLj is strong. As seen in the table 2, W1, W3 and W4 all belong to the class GL2, but W1 has more monopoly for the class GLj, and then A3 and A4 (Li, 2009). In the situation with same word frequency, the word with stronger monopoly should rank in front, and the weighting formula could be obtained based on that. For the word Wi,

W(Wi)=Hval + Mij* W(W, i) + Pij * C (8)

Where, C is the constant, Hval is the weight of the subject word in the title, and W(W, i) is the weighting function of the formula (7).

When certain one word occurs in certain one specific official document class, to show the monopoly of the subject words in the official document, the importance should amount to the subject words in the title. At the same time, to distinguish the subject words with less weights (subject words occur in many official document classes at the same time), the formula Pij= mij / $\sum_{j=1}^{n} mij$ could be used, and when the subject words only belong to certain one class, Pij=1, and when they belong to above two classes at the same time, Pij=1/n (n is the amount of the official class number).

4. Design of the subject indexing and the computer implementation

4.1 Feudal autocracy blocked the political modernization

Based on the intelligence linguistics principle, the method in this article uses the classification codes (categories word) to control the subject words, and uses the subject words to control the keywords, and establishes the knowledge base based on the subject words. In fact, this kind of knowledge base is an expert knowledge system, including Chinese Library grammar bank, Chinese table base, classification codes, subject words comparison base, the thesaurus base, and the keywords base. Based on these knowledge bases, the integration of indexing and searching of classification language, topic language, and natural language could be realized, and the automatic indexing and the automatic classification can be implemented.

The concrete implementation includes following approaches.

(1) Abstract the information in the document and establish the knowledge base.

(2) According to the control identifier of the document, and take out the text content of various indexing sources from the document, and put them into the database for the reserve of different fields.

(3) Segmentation. Use the self-made stop word base or half-stop word base with thousands words to set up the segmentation symbol for the indexing source text character string, and segment long character strings into many shorter substrings.

(4) Abstract keywords. Abstract the worlds from the indexing source text character strings by the keywords base. The abstracting process is implemented by the positive maximum matching method to ensure that the words with long word length and more special concept could be abstract first.

(5) Based on the function of the thesaurus base, change the indexing words from the keywords to the subject words, and implement the statistics and ranking of the word frequency weight values for those indexing words, and then complete the topic indexing of webpage.

(6) Based on the function of the classification codes-subject words comparison base, use the word similarity degree matching algorithm to turn the topic words (or word strings) to corresponding classification codes or class name word, and then complete the classification indexing of webpage (seen in Figure 1).

4.2 Computer implementation

This module is implemented in VB.net, and part function codes are seen as follows.

(1) Function: OpenWordfile. It is used to open the subject word table, and take the length and the word property of each topic word out. According to the establishment rule of the topic word table, each topic word and its word property occupy one line, and the word property is the last two bytes, so the content of the last two bytes on each line should be the word property of the word, and the surplus is the topic word.

```
“Open FilePath For Input As #FileNum                'Open the file of topic word
Do While Not EOF(FileNum)
Line Input #FileNum,Word(Linecount)              'Take out one line and save it into the array each time
'Following codes are to take out the topic words block by block and corresponding length and word property of
these words
Wordquality(Linecount)=Right(Word(Linecount),2)
Word(Linecount)=Left(Word(Linecount), LenB(Word(Linecount))-2
Wordlen(Linecount)=Len(Word(Linecount))
Linecount= Linecount+1
Loop
Close #FileNum”
```

(2) Function: EdocWord. It is used to open the official documents which should be indexed. To prevent the document is too large and overflow, this method takes the topic words out block by block, and each block (16K) uses the array word(i) in the function OpenWordfile. This array is used to store the topic words in the topic word table. And then the function CountStrings is used to count the word frequency of each topic word in the official document one by one.

```
“FileNum = FreeFile
Open TextStr For Binary As #FileNum             'Open the official document
Do While Not EOF(FileNum)                              'Take out all contents of the text if it doesn’t end, and each
block is 16K
Filestr = Space(16 * 10 ^ 3)                     'Distribute the space of 16K
Get #FileNum, , Filestr                     'Take out the content of the document
Strback = Strback + Filestr          'Put the last 16 bytes in the former block in the front of the latter block to
prevent omitting the topic words
'Store the times that each topic word occurs in the whole document
ReDim Preserve Time(0 To Linecount - 1) As Integer
For i = 1 To Linecount - 1
'Transfer the function CountStrings to compute the times that each topic word occurs
Uses the function “CountStrings” to count the time of each subject word
Time(i) = Time(i) + CountStrings(Strback, Trim(Word(i)))
Next i
Loop
Close #FileNum”
```

Then, take out the attributes of the topic words such as the position and the paragraph where the toptic words are located according to the design, and use the formula (4), the formula (5), the formula (6), the formula (7), and formula (8) to compute the weights of the topic words, and finally complete the topic indexing.

5. Conclusions

Aiming at the problems in the official document indexing and combining with the characteristics of the official documents, the word frequency statistics and the non-linear weighting function are used to obtain the weights and implement the subject indexing according to different weights. The subject words table is stored by the document form, not the database, which is very advantageous from speed, cost, and maintenance. After the weighting function is improved, the accuracy rate, the standard-ability, and the efficiency of the subject indexing all will be enhanced. Though the effect is satisfactory through practical test, but this method in the article still has some problems. First, what rules that the experienced values and the nonlinear function should follow have not be confirmed, and they should not depend on the statistical result from limited data. Second, the subject word table could be established according to relative rules of the General Office of the State Council of China, and the tables such as the association knowledge table and the emphatic pattern table which would influence the weight value of the subject words should be studied first in the future.

References

Baayen R.H. & Lieber R. (2006). Word Frequency Distributions and Lexical Semantics. *Computers and the Humanities.* Volume 30, Number 4. P.281-291(11).

Barnes, C.I. & Constantini, L. and Perschke, S. (1978). Automatic Indexing Using the SLC2 ⅡSystem. *Information Processing and Management.* No.14(2). P.107-119.

Bun, K.K., Ishizuka, M. (2002). Topic extraction from news archives using TF*PDF Algrithm. *The Third International Conference on Web Information Systems Engineering.* Singapore, P.73-82.

Chen Yesho. (1990). Booth's Law of Word Frequency. *Journal of the American Society for Information Science*. No.4.

Christoher S, Yubin D, Teck E. (2002). Using Statistical and Contextual Information to Identify Two- and Three-Character Words in Chinese Text. *Journal of the American Society for Information Science and Technology.* No.53(5). P.365-377.

Han, Kesong & Wang, Yongcheng. (2000). A Non-linear Term Weighting Method Used for Subject Distillation. *Journal of the China Society for Scientific and Technical Information.* No.12. P.651-653.

Jin H, Wong K. (2002). A Chinese Dictionary Construction Algorithm for Information Retrieval. ACM Transactions on Asian Language Information Processing, 2002, 1(4): 281-296.

Kang, Kai, Lin, Kunhui & Zhou, Changle. (2006). New Text Categorization Method Based on the Frequency of Topic Words. *Journal of Computer Applications.* No.8. P.1993-1994.

Li, Jia. (2009). The Study on Core Word's Roles in Co-word Clustered Analysis. *Journal of Intelligence.* Dec 2009. Vol.28, No.12.

Pu, Qiang, Li, Xin, Liu, Qihe & Yang, Guowei. (2007). Study on General Extracting Method of Web Topic Text. *Journal of Computer Applications.* No.5. P.1394-1396.

Secretariat of the General Office of the State Council of China. (1997). *The State Council Official Keywords List of China.* Beijing. Dec 1997.

Trotman A. (2005). Choosing document structure weights. *Information Processing and Management.* No.41. P.243- 264.

Xu, Dezhi & Liu, Yijing. (2010). A Content Analysis Method Used for Ontology Ranking. *Application Research of Computers.* Jun 2010. Vol.27, No.6.

Table 1. Matrix M

	GL1	GL2	GL3	GL4	GL5	GL6
W1				1		
W2	1					
W3				1		1
W4	1	1	1	1	1	1
…						

Table 2. Matrix P

	GL1	GL2	GL3	GL4	GL5	GL6
W1				1		
W2	1					
W3				1/2		1/2
W4	1/6	1/6	1/6	1/6	1/6	1/6
…						

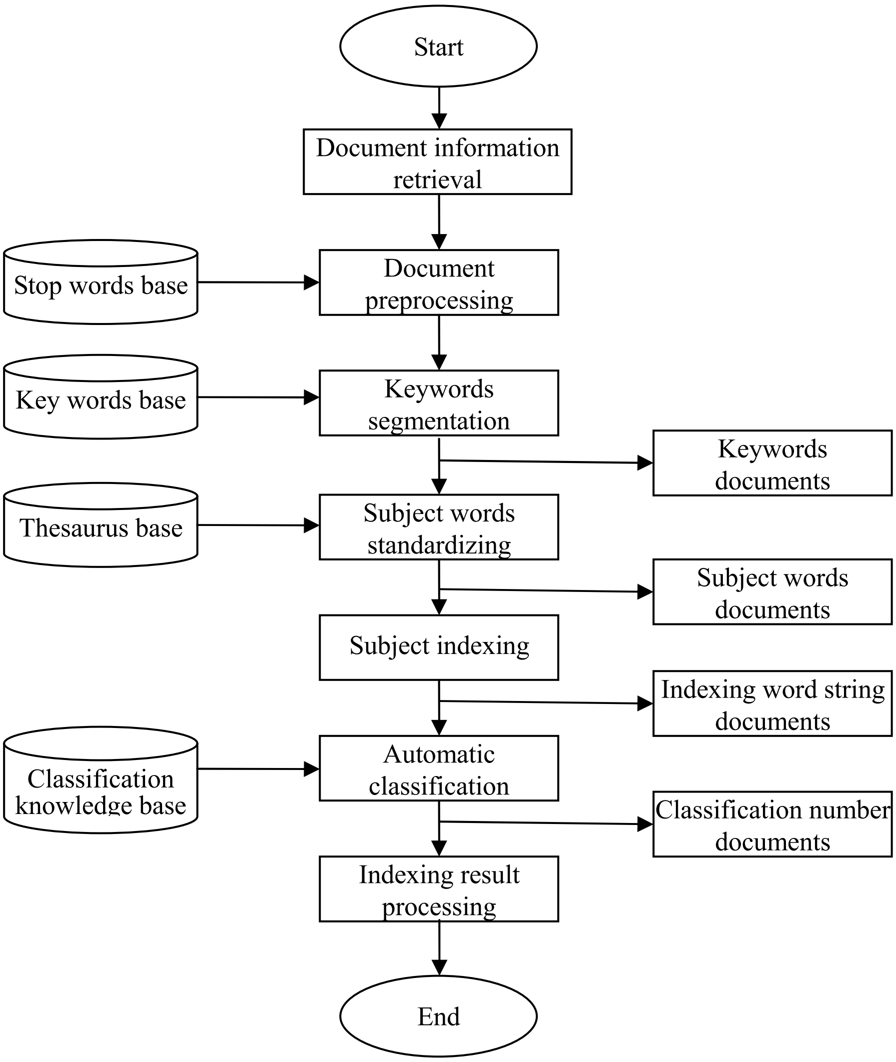

Figure 1. Flow Chart of the Automatic Indexing and Classification of Official Documents