

Document relevance calculation based on Lexical cohesion with structure analysis¹

Zhao Yuming, Liu Bingquan, Wang Xiaolong

School of Computer Science and Technology, Harbin Institute of Technology, Harbin
150001, China

Email: {ymzhao, liubq, wangxl}@insun.hit.edu.cn

Abstract

This paper explores the feasibility of constructing a document relevance calculating model based on lexical cohesion with structure analysis. In this model, by extracting the semantic-relative word clusters in documents according to the lexicon cohesion principle, documents are formalized in expressions which are composed of lexicon chains with structure information. And based on this kind of representation, document relevance calculation is substituted by semantic distance calculation of lexical chains. The feasibility of this novel approach has been examined by experiments conducted on Chinese Library Classification (CLC) dataset. The results show that the method makes good use of the background knowledge of ordinary users, and it is an effective method for relevance calculation of documents.

Keywords

Document relevance calculation, lexical cohesion, lexical chain, Bag of Words,

1. Introduction

The concept of “relevance” has been of great importance in the research field of information processing. From 1930s to present, a lot of research work has been done on it, and the history could be divided into three stages: in the first stage, from 1930s to 1950s, “relevance” was discussed and understood from the point of view of “system”; in the second stage, from 1960s to 1970s, a lot of experimental studies (Cuadra C.A. 1967, Rees, A.M. 1967) were done to investigate the relevance judgment process, and frameworks of the concept of “relevance” were presented (Saracevic T. 1970, Cooper W S. 1971, Cooper W S. 1973); in the third stage, from 1980s to present, MacMullin (MacMullin S.E. 1984), Taylor (Taylor R.S. 1986), and Belkin (Belkin N.J. 1982) made studies which focused on

¹ Supported by National Natural Science Foundation of China (60435020, 60673037) and The National High Technology Research and Development Program of China (2006AA01Z197, 2007AA01Z172)

the views of users and sense-making, and Mizzaro (Mizzaro S. 1998) presented the famous “Mizzaro Model”, which introduces “time” as a influencing factor to relevance judgment. And now, an important phenomenon deserving our attention is that more and more researches are carried out on relevance feedback (Makoto 2000, Gu Z. 2004, Wu, H.C. 2006) to help improve the performance of information processing, especially of information retrievals.

Document relevance calculation, which provides the relevance degree between two documents automatically, is widely needed in text classification and clustering, relevance feedback of information retrieval, and many other fields of text processing.

Besides document relevance, there’s another concept, document similarity, which is also often mentioned in text processing. Document similarity calculation methods, such as the Vector Space Model, are always used as approximate methods for document relevant calculation. But, accurately, “similarity” is only a kind of relationship of “relevance”. For example, “serviceman” is similar with “soldier”, and they are also two relevant words. But we cannot say “soldier” is similar with “barrack”, although they are obviously relevant. So a method which can well describe the concept of “relevance”, not “similarity”, is needed to accomplish document relevance calculation.

Present methods of document relevance judgment could be classified into two kinds: “system-oriented relevance” and “user-oriented relevance”. For the methods of “system-oriented relevance”, “relevance” is considered as the connatural property of a system, and the methods focus on the inner mechanism of the system. For methods of “user-oriented relevance”, the interaction between users and systems are what they focused on. Although theories and methods about “user-oriented relevance” are ardently discussed, document relevance judgment methods of “system-oriented” are still widely applied in practical applications for their perspicuity and easily manipulated.

To calculate the relevance degree of documents, the first problem should be solved is how to represent the documents, and after that, the second problem is how to use these representations. The most popular approach of text representation is bag-of-word (BOW). In this representation, there is a dimension for each word, and a document is then encoded as a feature vector with word TFIDF weighting as elements. Despite of many other more sophisticated techniques(Lewis D.D. 1992) for text representation, so far the BOW method still can produce excellent results (Franca D. 2005). But document relevance judgment of human beings is not the simple management made at word-level, but a behavior taking place at semantic-level. So techniques, especially techniques of text representation, which involve more semantic information, are needed for document relevance calculation.

This paper presents a document relevance calculation method based on lexical cohesion with structure analysis. It constructs lexical chains containing structure information to represent documents so that to make good use of the semantic features of them. And based on this representation, document relevance calculation is substituted by semantic distance calculation of lexical chains.

In Section 2 of this paper, lexical cohesion and the reason why it is adopted for document relevance calculation are introduced; in Section 3, the details of document relevance model based on lexical cohesion with structure analysis are described; in Section 4 the description about experiments and evaluations is given; and finally, the conclusion and the future work are presented in Section 5.

2. Lexical Cohesion and Document Relevance Calculation

2.1 Lexical Cohesion

Lexical cohesion, which is achieved through semantic connectedness between words in text, is a characteristic of natural language texts (Halliday M.A.K. 1976). A single instance of a lexical cohesive relationship between two words is usually referred to as a lexical link (Ellman 2000, Hirst 1997, Hoey 1991 and Morris 1991). And lexical chains, sequences of linked words, are normally used to realize lexical cohesion in text.

2.2 Lexical Cohesion in Document Relevance Calculation

To overcome the shortcomings of ordinary “system oriented” document relevance calculation methods which are widely used now, a method that carries out calculation at the semantic-level is needed.

In natural language texts, the presences of semantic relationships between words are great. For example, in a text which subject is “art”, there’re words such as “movie”, “producer”, “show”, “playwright”, “comedy”, “director”, etc., which have semantic relationships between each other obviously. Unfortunately, in the traditional methods, these words are treated and calculated separately, and the relationships between them are ignored. Usually, an individual of these words which describe the subject of the text excellently may not have a great enough TFIDF weight, and the importance of it may often be diluted with the other words that do not contribute to the subject so much. If the semantic relationships are considered and these related words are taken as a word cluster, it will be strong enough to defend against the interferer of insignificant words with great TFIDF weight, and the subject of the text will be better described. Based on these better representations, better relevance calculation results can also be expected. Detecting, investigating and making use of the semantic relationships between words is just what goes along with the theory of lexical cohesion. So theories, concepts and techniques of lexical cohesion can be adopted. And in this paper, we focused on mining the semantic information of texts based on lexical cohesion, and made it helpful for document relevance calculation.

3. Document Relevance Model based on Lexical cohesion with Structure analysis (DRMLS)

The document relevance model based on lexical cohesion with structure analysis presented in this paper is composed of three parts: Document Representation based on Lexical chain with Structure information (DRLS); Lexical chain Weight Calculation with Multi-features Fusion (LWCMF); Document Matching based on Lexical chain Relevance (DMLR). In this section, the details of these three parts are described.

3.1 Document Representation based on Lexical chain with Structure information (DRLS)

The function of DRLS is to make a document be represented by a set of lexical chains containing structure information through measuring semantic relationships between words in the document. To measure the cohesion relationships of words, Hownet is adopted as the knowledge resource.

Hownet is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in Chinese and English bilingual lexicons

(Dong Z.D. 2003). It focused on concepts rather than words. In Hownet, a word may have multi-concepts, and each concept is described by a record. And since the “DEF” item of a record, which describes the definition of the concept and the relationships such as Synonym, Hyponym, and Antonym, etc., is the nucleus of Hownet, DRLS measures the cohesion relationships of words mainly depending on it.

For word w_1 and word w_2 , the concepts they have are $\{DEF(w_1)_i | i=1,2,\dots,n,(n \geq 1)\}$ and $\{DEF(w_2)_j | j=1,2,\dots,m,(m \geq 1)\}$. $I(DEF(w_1)_i)_p$, $p \geq 1$, is a sememe (Dong Z.D, <http://www.keenage.com/>) of the word w_1 's concept $DEF(w_1)_i$.

The lexical link restriction between two words in DRLS is defined as follows:

1. reiteration: $w_1 = w_2$;
2. collocation: $w_1 \neq w_2$, and $\forall i, j, R(I(DEF(w_1)_i)_p) \in DEF(w_2)_j$

$$\begin{aligned}
 R &= f_1 \circ f_2 \dots \circ f_m, \quad 1 \leq m \leq 5 & (1) \\
 f_v &\in \begin{cases} F \setminus \{Converse\}, & \exists! u < m, u < v, s.t. f_u = Converse \\ F, & else \end{cases} \\
 &1 \leq v \leq m, \\
 &F = \{Antonym, Converse, Synonym, hyponym\}
 \end{aligned}$$

In DRLS, a document is represented as

$$D_L = \{L(D)_1, L(D)_2, \dots, L(D)_n, n \geq 1\} \quad (2)$$

$L(D)_i$, $1 \leq i \leq n$, is the i th lexical chain of document D_L . It is defined as

$$L(D)_i = \{W(L(D)_i)_1, W(L(D)_i)_2, \dots, W(L(D)_i)_m, m \geq 1\} \quad (3)$$

$W(L(D)_i)_j$, $1 \leq j \leq m$, is the j th item of lexical chain $L(D)_i$. If $m > 1$, items in $L(D)_i$ must be connected based on the lexical link restriction defined above. And the item $W(L(D)_i)_j$ is described as

$$W(L(D)_i)_j = (word, wordweight, freq, (p_1, p_2, \dots, p_k), connect) \quad (4)$$

word is the spell of a word; *wordweight* is its TFIDF weight; *freq* is the frequency of the word's appearance in document D ; (p_1, p_2, \dots, p_k) , $k \geq 1$, is the sequence of the paragraph IDs that the word appears in; and *connect* is the number of the other words in lexical chain $L(D)_i$ that have lexical links to *word*.

connect is an important unit that describes the structure of lexical chain in DRLS. For the ordinary lexical cohesion theory, words are just made into a chain, and the structure of the chain and the weight of each word in the chain have not been cared about. Through filling the value of *connect* of each item in a lexical chain when it being constructed, DRLS gets the structure information of the lexical chains, which is greatly useful in LWCMF and DMLR.

After preprocessing (stop words wiping off, word segmentation and part of speech, etc.), the document is taken into DRLS as a sequence of words. The first word is treated as the first item of the first lexical chain $L(D)_1$. The word item description information of each part in formula (4) is filled in. For the next word, the relationship between the word item in existing lexical chain and itself is measured based on the lexical link restriction defined as before. If the relationship accords with the restriction, the word will be added into the same lexical chain as another word item, else a new lexical chain will be created, and the word will be accepted as the first word item of this chain. The processing of the other words in the sequence of the document is on the same analogy.

A portion of lexical chains of a document is shown in Figure 1.

L2
((表演,0.377021,3,(4,6,7),8), (剧场,0.260255,1,(5),12), (剧团,0.338647,2,(6),4), (舞台,0.338647,2,(5),11), (戏院,0.260255,1,(5),11), (演唱,0.260255,1,(4),12), (演出,0.377021,3,(5),8), (演员,0.377021,3,(5,6),12), (艺人,0.260255,1,(6),12), (剧种,0.260255,1,(7),10), (摄制,0.260255,1,(6),10), (文艺,0.338647,2,(4,6),11), (艺术,0.338647,2,(4,6),10), (作品,0.260255,1,(1),11), (故事,0.260255,1,(4),5), (经典,0.260255,1,(1),1))
L3
(动人,0.260255,1,(7),0)
L4
((成长,0.260255,1,(6),1), (发展,0.402033,4,(2,3,4,6),2), (吸收,0.260255,1,(4),1))

Fig 1. An example of lexical chains constructed by DRLS

3.2 Lexical chain Weight Calculation with Multi-features Fusion (LWCMF)

There're several features that influence the weights of lexical chains of a document (You W.J. 2003). Four features, which can measure the lexical chain's ability of describing the subject of the document, are selected and fused to be endowed to every lexical chain as its weight.

3.2.1 Feature Selection

Feature 1: Length of lexical chain

A longer lexical chain covers more context of the document. That means the topic of this chain is more similar to the subject of the document. So the length and the weight of a lexical chain are in direct proportion.

$$F_{n1} = \frac{\sum_{s=1}^{i_n} freq(W(L(D)_n)_s)}{\max(\sum_{s=1}^{i_1} freq(W(L(D)_1)_s), \sum_{s=1}^{i_2} freq(W(L(D)_2)_s), \dots, \sum_{s=1}^{i_m} freq(W(L(D)_m)_s))} \quad (5)$$

m is the amount of the lexical chains that document D has.

Feature 2: Weight of word items

Words which have greater TFIDF weights always act as more important characters in a document. And lexical chain which has greater average word item weight should also act as a more important role.

$$F_{n2} = \frac{\sum_{s=1}^{i_n} \text{wordweight}(W(L(D)_n)_s)}{i_n} \quad (6)$$

i_n is the amount of word items that lexical chain $L(D)_n$ has.

Feature 3: Area that lexical chain covers

A lexical chain is likely to focus on a sub-subject of a text if the components of it appear only in a small area. The larger the area that a lexical chain covers the better it describes the subject of the whole text. The number of the different paragraphs that words in a lexical chain appear in is used to describe this feature.

$$F_{n3} = \text{difpnum}(L(D)_n) / \text{pnum}(D) \quad (7)$$

$\text{difpnum}(L(D)_n)$ is the number of the different paragraphs that words in lexical chain $L(D)_n$ appear in; and $\text{pnum}(D)$ is the total of paragraphs that document D has.

Feature 4: Structure of lexical chain

The concept of lexical “chain” in DRLS is not restricted to the concept “chain”, which is defined in the field of data structure. In a lexical chain of DRLS, which has two or more than two components, a word item has at least, but not only one semantic relationship (the same thing that has been talked as lexical link in 3.1) to another word item. And the lexical chain, whose components have more semantic links with each other, is semantically tighter inside, and it is more meaningful for a document.

$$F_{n4} = \frac{\sum_{s=1}^{i_n} \text{connect}(W(L(D)_n)_s)}{i_n \times (i_n - 1)} \quad (8)$$

3.2.2 Feature fusion:

Logistic regression is a popular technique for modeling the effects of one or more explanatory variables on some binary-valued outcome variables. In LWCMF, relevant and not relevant are used as outcomes, and the observations, relevant values obtained by calculating respectively based on those 4 features that are talked above, are described as $\vec{x}_{fn} = \{x_{fn1}, x_{fn2}, x_{fn3}, x_{fn4}\}$. And the probability that they indicate success is given as

$$q(\vec{x}_{fn}) = \frac{e^{g(\vec{x}_{fn})}}{1 + e^{g(\vec{x}_{fn})}} \quad (9)$$

and,

$$g(\bar{x}_{jn}) = \beta_0 + \sum_{j=1}^4 \beta_j x_{jnj} \quad (10)$$

In LWCMF, based on the criteria of Chinese Library Classification² (CLC4), 240 documents of 6 classes, A, E, J, Q, TM, and U, from the data used in TC evaluating of the High Technology Research and Development Program (863 project) in 2003 and 2004, are made into 3600 pairs, and these documents pairs are adopted as training data of logistic regression for feature fusion. Based on the parameters, $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, obtained by logistic regression, the weight of a lexical chain $L(D)_n$ is calculated as

$$lweight(L(D)_n) = \frac{e^{(\beta_0 + \beta_1 F_{n1} + \beta_2 F_{n2} + \beta_3 F_{n3} + \beta_4 F_{n4})}}{1 + e^{(\beta_0 + \beta_1 F_{n1} + \beta_2 F_{n2} + \beta_3 F_{n3} + \beta_4 F_{n4})}} \quad (11)$$

3.3 Document Matching based on Lexical chain Relevance (DMLR)

3.3.1 Relevance Calculation of Lexical chains

As it is discussed above, a word item in a lexical chain of DRLS which has at least two components may have semantic relationships to more than one another word items in the same chain. Actually, the lexical chain here we talk about is a graph composed of word items as nodes and semantic relationships as edges. We just follow the tradition of lexical cohesion theory and still call the representation structure in DRLS as “lexical chain”.

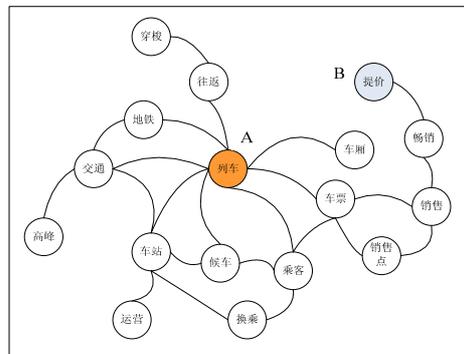


Fig 2. An example showing the word items of different status in a lexical chain

Word items in a lexical chain may have different status, as it is described in Figure 2. The central word item “A”, which has more semantic relationships than the marginal word item

²The guideline of the evaluation on Chinese Text Classification in 2004.
http://www.863data.org.cn/english/2004syllabus_en.php

“B”, has more important status in this lexical chain, since the central word item describes the semantic subject of the lexical chain more powerful.

In DMLR, whether two lexical chains are relevant depends on whether there is at least one pair of word items, which are from these lexical chains respectively, having cohesion relationship. Since different word item has different status in a lexical chain, different cohesion pair composed of different word items also has different meaning to the relevant degree of two lexical chains.

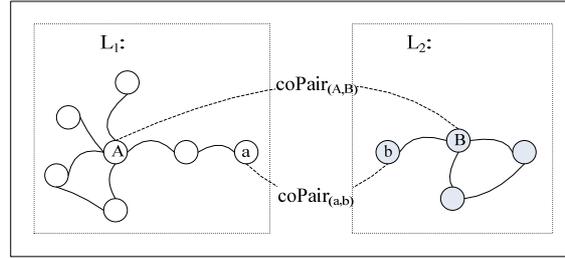


Fig 3. An example showing the cohesion pairs of different status word items

As it is shown in Figure 3., for the relevance calculation of lexical chain L_1 and L_2 , the cohesion pair composed of word item “A” and “B” is more meaningful than the cohesion pair composed of word item “a” and “b”, because word item “A” and “B” have more important status. So the description of word item status can be used to analyze the relevance of two lexical chains quantitatively.

In DMLR, it is calculated as

$$Lrel(L(D_j)_i, L(D_i)_s) = \frac{2 \sum_{0 \leq q \leq |L(D_j)_i|, 0 \leq p \leq |L(D_i)_s|} connect(W(L(D_j)_i)_q) \times connect(W(L(D_i)_s)_p)}{\sum_{n=1}^{|L(D_j)_i|} connect(W(L(D_j)_i)_n)^2 + \sum_{m=1}^{|L(D_i)_s|} connect(W(L(D_i)_s)_m)^2} \quad (12)$$

and

$$connect(W(L(D_j)_i)_k) = \frac{connect(W(L(D_j)_i)_k)}{\sum_{h=1}^{|L(D_j)_i|} connect(W(L(D_j)_i)_h)} \quad (13)$$

is used to describe the status of a word item in a lexical chain, and $connect(W(L(D_j)_i)_k)$ is the corresponding $connect$ in formula (4).

3.3.2 Matching Algorithm of Documents

Algorithm: Calculate the relevance degree between two documents

Input: $D_i = (L(D_i)_1, L(D_i)_2, \dots, L(D_i)_n, n \geq 1)$, $D_j = (L(D_j)_1, L(D_j)_2, \dots, L(D_j)_m, m \geq 1)$

Output: $Drel(D_i, D_j)$ (the relevance degree between D_i and D_j)

Step 0: For each $L(D_i)_k$ in D_i

Filter out lexical chains which $lweight(L(D_i)_k) < \lambda$,

s.t. $D_i' = (L(D_i)_1, L(D_i)_2, \dots, L(D_i)_p, 1 \leq p \leq n)$

For each $L(D_j)_l$ in D_j

Filter out lexical chains which $lweight(L(D_j)_l) < \lambda$,

s.t. $D_j' = (L(D_j)_1, L(D_j)_2, \dots, L(D_j)_q, 1 \leq q \leq m)$

Step 1: For each $L(D_i)_k$ in D_i'

For each $L(D_j)_l$ in D_j'

If $L(D_j)_l$ relevant to $L(D_i)_k$

Then

$$Lrel(L(D_j)_l, L(D_i)_k) = \frac{2 \sum_{\substack{0 \leq s \leq |L(D_j)_l| \\ 0 \leq t \leq |L(D_i)_k|}} contweight(W(L(D_j)_l)_s) \times contweight(W(L(D_i)_k)_t)}{|L(D_j)_l| \sum_{s=1}^{|L(D_j)_l|} contweight(W(L(D_j)_l)_s)^2 + |L(D_i)_k| \sum_{t=1}^{|L(D_i)_k|} contweight(W(L(D_i)_k)_t)^2} \quad (14)$$

Else

$$Lrel(L(D_j)_l, L(D_i)_k) = 0$$

Find $\mu_k = \max(Lrel(L(D_j)_l, L(D_i)_k))$

Step 2: Compute the relevance degree between D_i and D_j with the improved method based on DICE coefficient (Sun J.J. 2004)

$$Drel(D_i, D_j) = Drel(D_i', D_j') = \frac{2 \sum_{1 \leq s \leq p, 1 \leq t \leq q} \mu_s \cdot Lweight(D_i)_s \cdot Lweight(D_j)_t}{\sum_{s=1}^p Lweight(D_i)_s^2 + \sum_{t=1}^q Lweight(D_j)_t^2} \quad (15)$$

4. Evaluations and Analysis

4.1 Experiment Setting

The experiments of DRMLS are conducted on the data that used in TC evaluating of the High Technology Research and Development Program (863 project) in 2003 and 2004, and the supplementary data from Internet, based on the criteria of Chinese Library Classification² (CLC4). The performance of document relevance calculation is inspected by the performance of document classification. And the class identifications made by human

beings are taken as the evaluating measure.

In order to demonstrate the effectiveness of DRMLS, two experiments were carried out:

Test1: From each of the 6 classes, “A: Marxism”, “E: military affairs”, “J: art”, “Q: bioscience”, “TM: electric technology” and “U: traffic”, 60 documents are selected, and they are divided into two sets, “I” and “II”, and each of it is composed of $30 \times 6 = 180$ documents. Documents in sets “I” and “II” are made into $180 \times 180 = 32400$ test pairs. For a document in set “I”, if its average relevant degree to the documents belonging to one of the categories in set “II” is greater than to the documents of other categories, it will be identified as belonging to this category. And if this category is the same one that the document is selected from, the relevance calculation on it is considered as correct, otherwise, it is wrong.

Test2: With the same data in Test1, the KNN classification algorithm based on DRMLS is used to identify the categories of documents in set “I”. And its performances on different k values are inspected.

4.2 Results

To verify the technical soundness of DRMLS, Test1 compares the results of DRMLS with the traditional outstanding documents relevance calculating algorithm, VSM based on BOW. The precisions, recalls and F-measures of DRMLS and BOW which are obtained in Test1 are shown in Figure 4.

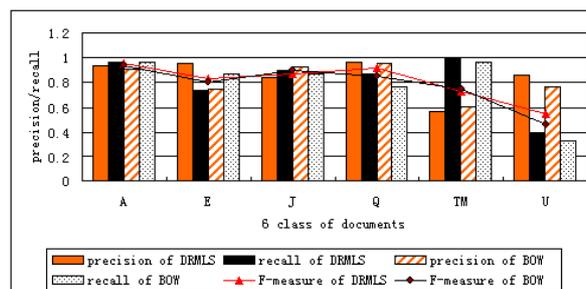


Fig 4. Precisions, recalls and F-measures of DRMLS and BOW

In Test1, for documents of the 6 classes, the average F-measure of classification based on BOW is 78.22%, and the average F-measure of classification based on DRMLS is 80.55%. For the criteria of Chinese Library Classification² (CLC4), the semantic distances between subjects of categories are greatly different, and the inner cohesions of different categories are also greatly different. So the performances of DRMLS varies on documents of different classes (For documents of classes “A”, “E”, “J” and “Q”, the average F-measure of classification based on DRMLS is 89.11%, and for documents of classes “TM” and “U”, the average F-measure is only about 63.41%). Since the inner cohesions between documents of class “TM” and class “U” are weaker, we adjust m in Formula (1) and make $1 \leq m \leq 7$, which relax the restriction of lexical link between words. The results are shown in Figure 5.

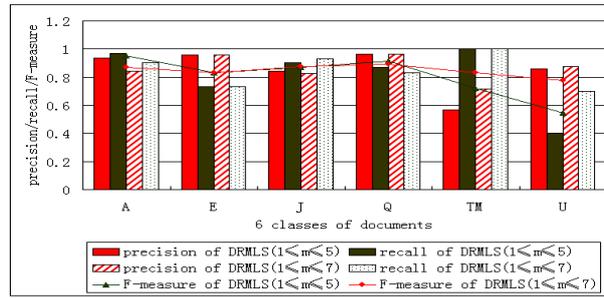


Fig 5. Precisions, recalls and F-measures of DRMLS on different m

The results of Test2, the F-measures of “A”, “E”, “J”, “Q”, “TM” and “U”, and the average F-measures of them, obtained by the KNN classification method based on DRMLS while different k values are selected, are shown in Figure 6.

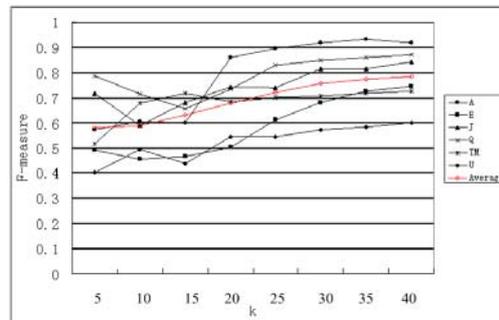


Fig 6. F-measures of DRMLS on different k values

The x-axis is the value of selected k of KNN, and the y-axis is the value of F-measure. The results show that with the increasing of k value, the average F-measure also increases. While the selected k values are less than 30, the increasing rate of average F-measure is much greater than the rate while the k values are more than 30. According to the experiment data set in Test2, the maximum k value that can be selected and is meaningful to documents classification is 180. And in fact, when the k value is defined as 180, the results in Test2 are just the results of DRMLS that are obtained in Test1.

In Figure 4, for class "E" and "J", the recall or precision of BOW is higher than DRMLS. And after adjust the value range of m , in Figure 5, the recall or precision of BOW is higher than DRMLS in class "A", and "Q", instead of class "E" and "J". It can be seen that the adjusting of m is what induced this change. The aim of adjusting m is to adapt the different inner cohesions of different document classes. And it can be conjectured that the difference of inner cohesion may be the reason why the the recall or precision of BOW is slightly higher than DRMLS in some class. A proper value of m , obtained from a method with more substantial theoretical foundation, can be expected to improve the performance of DRMLS. And it will be studied in our future work.

5. Conclusions

In this paper, Document Relevance Model based on Lexical cohesion with Structure analysis (DRMLS) is explored. Depending on the experiments results, it can be considered that DRMLS is an effective relevance calculation method.

DRMLS makes representations of documents with concepts which are described by the cohesions of words, instead of representing with words simply. And comparing with the ordinary lexical cohesion method, the structure information of a lexical chain, to be exact, the information about different weights of words in a lexical chain, is considered to improve the performance of document relevance calculation. With adjusting the restriction of lexical link between words, lexical chains with different inner cohesions among their elements can be constructed, so that DRMLS is enabled to adapt to the relevance calculation of documents with different inner cohesions. As an approach from “system-oriented relevance” to “user-oriented relevance”, DRMLS covers the knowledge background of ordinary users with the help of Hownet, and describes the semantic information of documents ulteriorly.

In the future, based on DRMLS, further research on the relevance calculation between documents of different languages will be carried out.

Acknowledgements

We thank Kunpeng Zhu and Yongdong Xu for discussions related to this work. This research was supported by National Natural Science Foundation of China (60435020, 60673037) and The National High Technology Research and Development Program of China (2006AA01Z197, 2007AA01Z172).

References

- Cuadra, C. A. and Katter, R. V., 1967, Experimental studies of relevance judgments: final report, System Development Corporation, vol. 1: Project Summary, NSF Report No. TM-3520/001/00.
- Rees, A. M. and Schultz, D. G., 1967, A field experimental approach to the study of relevance assessments in relation to document searching. vol. 1: Final Report, NSF Contract No. C-423.
- Saracevic, T., 1970, The concept of “relevance” in information science: a historical review, In T. Saracevic ed. *Introduction to Information Science*, pp. 111-151.
- Cooper, W. S., 1971, A definition of relevance for information retrieval, *Information Storage and Retrieval*, 7(1), pp. 19-37.
- Cooper, W. S., 1973, On selecting a measure of retrieval effectiveness, part 1, The subjective philosophy of evaluation, *Journal of the American Society for Information Science*, 24(2), pp. 87-100.
- MacMullin, S. E., and Taylor, R. S., 1984, Problem dimensions and information traits, *The Information Society*, 3(1), pp. 91-111.
- Taylor, R. S., 1986, *Value-added processes in information systems*, Norwood, NJ: Ablex Publishing Corporation.
- Belkin, N. J., Oddy, R. N., and Brooks, H. M., 1982, Ask for information retrieval. *Journal*

- of Documentation, 38(2), pp. 61-71.
- Mizzaro, S., 1998, How many relevances in information retrieval?. Interacting with Computers, Elsevier, Netherlands, June, 10(3), pp. 305-322.
- Makoto, I., 2000 Relevance feedback with a small number of relevance judgements: Incremental relevance feedback vs. document clustering. Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), Jul 24-Jul 28, Athens, Greece, p 10-16.
- Gu, Z., and Luo, M., 2004, Comparison of using passages and documents for blind relevance feedback in information retrieval. Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 482-483.
- Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L., 2006, Probabilistic document-context based relevance feedback with limited relevance judgments. Proceedings of the 15th ACM Conference on Information and Knowledge Management, CIKM, pp. 854-855.
- Lewis, D. D., 1992, An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92), New York, US., pp. 37-50.
- Franca, D., and Sebastiani, F., 2005, A Analysis of The Relative Hardness of Reuters-21578 Subsets: Research Articles. Journal of the American Society for Information Science and Technology, vol. 56, no. 6, pp. 584 – 596.
- Halliday, M. A. K., and Hasan R., 1976, Cohesion in English, Longman, London.
- Ellman, J., Tait J., 2000, On the generality of thesaurally derived lexical links, Proceedings of 5th JADT, pp. 147–154.
- Hirst, G., and St-Onge D., 1997, Lexical chains as representation of context for the detection and correction of malapropisms, WordNet. An electronic lexical database, MIT Press, Cambridge, MA, pp. 305–332.
- Hoey, M. , 1991, Patterns of lexis in text, Oxford: Oxford University Press, pp. 45-47.
- Morris, J., and Hirst G., 1991, Lexical cohesion computed by thesaural relations as an indicator of the structure of the text, Computational Linguistics, 17(1), pp. 21-48.
- Dong, Z. D., and Dong, Q., 2003, The Construction of the Relevant Concept Field, Proceedings of the 7th Joint Session of Computing Language (JSCL03), pp. 364-370.
- Dong, Z. D., and Dong, Q., <http://www.keenage.com/>
- You, W. J., Li, S. Z., and Li, T. Q., 2003, A text filtering module based on lexical chain, Application Research of Computers, pp . 9:32-35.
- Sun, J. J., and Cheng, Y., 2004, Information retrieval technology, Beijing: Science Press, 10, pp. 341-377.