Cross-Lingual Text Filtering Based On Text Concepts And kNN

Li Shaozi
1. School of Computer,
National University of
Defence Technology
Changsha, P. R. China
2. Department of Computer
Science, Xiamen
University
Xiamen, P. R. China
szlig@xmu.edu.cn

Su Weifeng
Department of
Computer Science,
Hong kong
University
of Science and
Technology
Kowloon, Hong
Kong
weifeng@ust.hk

Li Tangqiu
Department of
Computer Science,
Xiamen University
Xiamen, P. R.
China
tqli@xmu.edu.cn

Chen Huowang
School of Computer,
National University of
Defence Technology
Changsha, P. R. China
chenhuowang@nudt.edu.cn

Abstract

This paper presents the model that can be used to filter the texts which the user is interested in from a large scale of source texts in Chinese or in English. Each text which the user is interested in can be represented as a vector in the vector space of classifiable sememes. The text to be sifted is represented as a vector too. The relevance of the text to the user can be measured by using the cosine angle between the text and its k nearest neighbor in the vector space. Experiments have been done and their results show that this scheme yields good results .

1 Introduction

With the rapid growth of the Internet and other networked information, there is an increasing need for reliable automatic texts filtering. As the Internet is boundless, Cross-Lingual Information Filtering(CLIF) system is also eager to access information which may be important to the user. Information filtering includes three subtasks: collecting information from information sources, selecting information which may interest the user and presenting the information to the user. Figure 1 depicts this subdivision.[5]

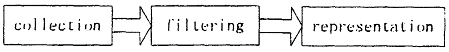

Figure 1 Information seeking task diagram

In this paper, we investigate how a cross lingual texts filtering model could be structured to acquire a user profile which enables it to distinguish between relevant and irrelevant documents in texts form on the Internet. This user profile is then used to accomplish the task of filtering documents from information sources automatically in English or in Chinese.

The task of texts filtering entailed: building a model of the features in the text predicting the relevance of the text to the user's interest. This asks for text representation which is discussed in section 3 and the user profile which is described in section 4. Techniques to compare between the text representation and the user profile to decide whether or not to keep the text and to notify the user are discussed in section 5. The architecture of the model is described in section 2. In section 6, report about some experiments is displayed to evaluate the value of this system. Conclusion about this model is made at the end of this paper. All techniques discussed below are suit to texts both in Chinese and in English. Some differences when handling information in Chinese and English will be discussed separately.

2 Multilingual texts filtering model

Figure 2 is the architecture of multilingual texts filtering model. First of all , the user provides sample texts as his interest to the system. The system analyses the sample texts by reducing them into sememes, a basic concept that will be described late. After calculating the sememes, we express the user's interest as a vector in the sememe vector space. And if a text is available, we analysis the text by reducing it into sememes and expressed it as a vector too. We calculate the similarity of the two vectors to show whether or not the text is relevant to the user's interest.

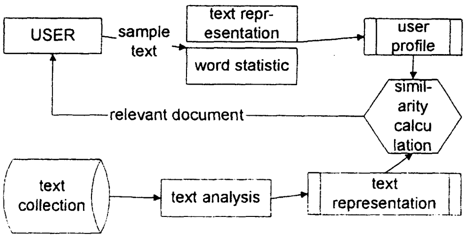

Figure 2:Texts filtering architecture

As mentioned above, this model calculates the similarity of a text to an user's interest based on sememes other than words. This model is based on HowNet[1], a knowledge database developed by Mr. Zhendong Dong, which comprises more than 53000 Chinese words and 57000 English words. In the philosophy of the HowNet, **sememe** refers to the smallest basic semantic unit that cannot be reduced further. Taken for instance, 'human being' , despite being a most complex concept composing a set of attributes, it can be regarded as a sememe. All concept can be reduced to the relevant sememes. Mr. Zhendong Dong has extracted a set of over 800 sememes that are used in HowNet and believed that all concepts in the world can be expressed by this set of sememes. If a word has several meaning, HowNet gives every concept a definition in its DEF entry by this set of sememes. Reducing a word to sememes can settle the problem synonymy of different words having the same concept which has puzzled us for several years with good results. Synonym is a well-known limitation of the word-based techniques that can make it difficult to find relevant documents. For example, word '电脑' and '计算机' are both explained as 'computer|电脑'. From this point of view, we can think Chinese words and English words are a kind of synonym too. We can take words '电脑', '计算机' and 'computer' as the same word without considering their different appearance. So we can handle texts in Chinese or in English without translation. Reducing the word into sememes is also useful for disambiguity which will be discussed in section 3.

Further more, we divided this set of more than eight hundred sememes into two group: sememes that are useful in classifying the relevance and sememes that are useless in classifying the relevance. We called the former group of sememes **classifiable sememes(C.S.)** and the latter group of sememes **unclassifiable sememes**. Most unclassifiable sememes are so frequently appear in the explanation of concepts, such as location, time, fact and or so, that if they were used in classifying, it would mislead us that most of the texts are somewhat alike. Classifiable sememes play important role in the process of filtering in our system.

3. Text Representation

In the techniques described later in this paper, we used the **vector space paradigm** in which the text is represented as a vector [3]. Assume some vector $\vec{D}$ where each element is d_i is a classfiable sememe in the HowNet. Each document then has a vector representation $\vec{V}$,where element v_i is the weight of the sememe d_i for the document. If the document does not contain d_i then v_i =0. The term *feature* refers to an element of the vector V for a document and so is denoted by v_i .

Not all words that appear in that document are reduced to sememes to be features, only those that are expected to contribute significantly to the estimation of the latent structure and to the classification of the document as relevant and irrelevant. That is to say, be it in Chinese or in English, main words should be selected from the document. Statistical criteria is used to decide which words contribute significantly. Because of the word's sense ambiguity, something should be done to disambiguity the main word. Thus the text representation can be summarized as below:

1) Judging the text is in English or in Chinese. If the text is in Chinese, it should be divided into words. If part of the text is in English and part of it in Chinese, the part in Chinese should be divided into words.

2) Part-of-Speech tagging. Part-of-Speech tagging involves selecting the most likely sequence of syntactic categories for the words in the text.

3) Removing all words belonging to the following categories: determiners(e.g. *a, the, all*), prepositions or subordinating conjunction(e.g. *in, to, o f*), pronouns (e.g. *I yours*), the infinite marker *to*, modal verbs(e.g. *would, must*) and coordinating conjunction(e.g. *and*). As a result. only the **main words** like nouns, verbs, adjective, adverbs in the text are left denoted as **W(s,w,c)**, in which **w** is the word ,**s** is the serial number and **c** is the category of the word. For example, (12,think,verb) means the *twelfth* word is *think* and this word is a *verb* after removing words mentioned above. Words that appear in the text very few can also be removed away. Thus the text is reduced to the main word list.

4) Word sense disambiguity. The basic idea is to estimate the probability of the sense of a word *w* relative to a window of words in **W** centered on *w*. Given a window size of n centered on a word *w*, the words in the windows are indicated as follows:

$$w_1 w_2 \quad w_{n\,2} w w_{n\,2+1} \quad w_{n-1}$$

For each sense of word *w* in HowNet, give it a initial weight *k* . Change the weight of the senses of word described as figure 3. In this module, we have taken into account the following relationship between classifiable sememes of other words in this window and the classifiable sememes of *w*:

a. the same classifiable sememe
b. material-product
c. agent-event
d. patient-event
e. instrument-event
f. location-event
g. time-event
h. event-role
i. concepts co-relation

As a result, in most cases, the actual meaning of the word *w* is emphasized by adding its weight. After that, we should unify the weight of the senses of word as follow:

$$wt(sense_i) = \frac{weight(sense_i)}{\sum_i weight(sense_i)}$$

where *i* is the serial number of the meaning of *w*.

```
$W_I$:Word I in the window except $w$
$S_{I\,J}$:No.J DEF of $W_I$
$CS_{I\,JK}$:No.K Classifiable sememe of $S_{I\,J}$
$WS_J$:No. J DEF of $w$
$WCS_{JK}$:No.K classifiable sememe of $WS_J$
Weight($WS_J$):Weight of No. J DEF of $w$

**Input**: $w_1 w_2 \quad w_{n\,2} w w_{n\,2\cdot 1} \quad w_{n-1}$
**Initialize** Weight($WS_J$)=1
For I=1 to n-1 // *Every word in the window*
                         *except w*
  For J=1 to (Number of DEF of $W_I$)
      For K=1 to ( Number of **C.S.** of $S_{I\,J}$ )
         For M=1 to (Number of DEF of $w$)
             For O=1 to (Number of **C.S.** of $w$)
           If $CS_{I\,JK}$ has $WCS_{JK}$ relation
                                  mentioned ablove
                THEN
                   Weight($WS_J$)=Weight($WS_J$)+
                                  1
           Endif
          Endfor
          Endfor
      Endfor
   Endfor
Endfor
```

Figure 3: Adjust the weight of the senses of word

5) Text representation as a vector. In the end, all words of the text are reduced to sememes. After removing unclassifiable sememes, the system calculates the weight of classifiable sememes of the text in the space of classifiable sememes in figure 4.

```
$KW_I$:No.I keyword
$S_{I\,J}$:No.J DEF of Keyword $KW_I$
Weight($WS_{I\,J}$):Weight of No. J DEF of $KW_I$

For I:=1 to (Keyword number)
   For J=1 to ($KW_I$'s DEF number)
      For K=1 to ($S_{I\,J}$'s **C.S.** number)
         Weightof($SM_K$)=Weightof($SM_K$)+
                             Wt($WS_{I\,J}$)
      Endfor
   Endfor
Endfor
```

Figure 4: Calculation of weights of feature of the texts in the vector space of **C.S.**

As a result, the text is represented as a vector in the vector space of classifiable sememes.

4. Cross-lingual Information Filtering

Once both the document and the user profile are represented as a vector, the relevance of a text to the user can be measured by using the cosine angle between the user vector and the text vector as below:

$$\cos(a)=\frac{(V_{user},V_{text})}{|V_{user}||V_{text}|} \qquad (1)$$

Where (V_{user},V_{text}) is the inner product of user vector and the text vector, $|V_{user}|$ and $|V_{text}|$ are the length of user vector and text vector.

We use kNN method to decide whether a text is needed: For an new text, we express it as a vector in the classifiable sememe space. In all the user's instance, we calculate k nearest neighbor by equ. (1). And then we calculate the similarity between the k nearest neighbor and the new text by equ.(2)

$$Si=\sum_{i=1}^{k} S^2(\cos(a_i)) \qquad .(2)$$

in which

$$S(x)=\begin{cases}0 & \text{if } x<h \\ x & \text{if } x>=h\end{cases}$$

The bigger the *Si* is, tne more reievant the text with the user's interest is . We can give a experiential value to decide whether or not the text is relevant to the user's interest. If the cosine angle between the two vectors is bigger than the value, we think that the text is of the user's interest and we present it to the user. Otherwise we get rid of the text or we keep it for a future browse. We can adjust the value of the valve taking into account of the user's feedback. If almost all files we presented to the user are of the user's interest, we should decrease the value. On the contrary, if many files we presented to the user are not of the user's interest, we should increase the value.

The value of k can be determined by the algorithm in figure 5:

```
biggestequal:=0
bigestk: =0;
FOR k:=2 TO a big integer
 km:=0;
   FOR I=1 TO the number of train instance
         For the No. I train instance,calculate
       the k nearest neighbor. If the same kind
       of text as No. I train instance,
       km:=km+1
    ENDFOR
    If km>biggestequal then
            Begin
                 biggestequal:=km;
                 bigestk:=k;
            end;
   ENDFOR
```

Figure 5:Determine the value of k

The advantage of reducing the concept to sememe and the reason why the module works so well in texts filtering is described as below.

Monosemantemic sememe: Every sememe is monosemantemic , and after the words' reducing to sememes, synonym have the same sememe. That is to say, no synonym sememe can be found which would improve the ratio of recall greatly.

Low dimension input space: Since all concepts are reduced to classifiable sememes, we calculate the similarity in the vector space of classifiable sememes. We only have to deal with about four hundred classifiable sememes other than more than 100000 words. It is obviously that low dimensional will greatly increase the ratio of recall. What is more, it will greatly cut the complexity of calculation the cosine angle between the vector of text and the user's profile.

Few irrelevant feature. After the extraction of main words , word disambiguity and the removal of unclassifiable sememes, most sememes that left are of the most importance to the text. Very few irrelevant feature were left to disturb our attention.

Document vector's feature's weight are big. Because the concepts of words are reduced to the sememes, the sememes that represent the text's main idea's weight may be much bigger than those irrelevant sememes which would diminish the influence of irrelevant sememes which may greatly increase the ratio of precision.

5. Experiments

We made use of documents from eight different users. Each of them were asked to provided 80 documents in English and 80 documents in Chinese in the field which is of his interest as positive document and 200 documents in other field as negative documents in which 100 of them are in English and 100 in Chinese. It should be noted that each of them could provide those documents quite easily. For each user, we made use of ten English documents and ten Chinese documents in his field as training set to form the *profile vector* for the user as described earlier. The positive documents that are not belong to the training set and the negative documents are mixed to form a test set in which we use the system to select the positive document .

We evaluate the performance of the two techniques using the measures of recall and precision. These are information retrieval measures rather than machine learning measures of performance. Recall is the ratio of relevant document identified as relevant by the system by the number of relevant documents present. Precision is the ratio of the number of relevant documents identified as relevant, to the total number of document presented to the user.

Table 1 shows the results of the experiment. The average precision and recall is about ninty percent. It shows that the filtering accuracy of this technique is quite good. In practice, the performance also increases if additional feedback was given by the user and are added into the system.

	Recall(%)		Precision(%)	
	Chinese	English	Chinese	English
User 1	88.7	86.6	86	88.7
User 2	90	91.5	88.6	90
User 3	90	86	85	90
User 4	89	85	88.7	89
User 5	86	84	87.5	86
User 6	87	87	88.5	87
User 7	92	90.6	84.7	92
User 8	91	90	90	91
Average	89.2	87.6	88.5	86.6

Table 1: Recall and Precision point of eight user using this method

We have represented the user's as a vector in the classifiable sememe vector. Compared with this, the kNN has advantage below:

1. If the user has several aspect of interest, they can be represented as different group of vector in the vector space which suit the user's interest.
2. If the user provides enough relevant text, it is more efficient to determine a new text's relevance.
3. If the user change its interest, it is easier for him to change without historical impact.

Our experiments also demonstrate these advantages as in figure 6 and figure 7.

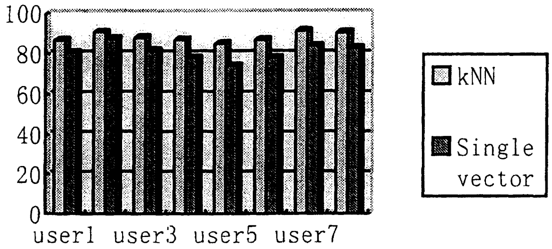

Figure 6:the recall point of kNN and single vector

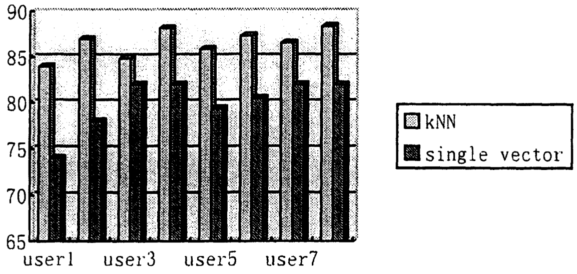

Figure 7:the precision of kNN and single vector

6. Conclusions

This paper describes a model that learns a profile of the user's preferences automatically in order to notify the user when relevant information becomes available and get rid of those that are not relevant to the user's interest. It provides both theoretical and empirical evidence that this module works very well for texts filtering. The theoretical analysis concludes that this module acknowledges the particular properties of text: (a) Monosemantemic sememe, (b) Low dimensional input space, (c) Few irrelevant feature ,and (d) Document vector's feature's weight are big. The experimental result shows that this module consistently achieves good performance on texts filtering. This makes it a useful and promising method for filtering from information sources.

7. Acknowledgement

Our thanks go to Prof. Zhengdong Dong for his help on HowNet. This project is supported by the National '863 High-Tech Program of China(2001AA114110) and major project of Science-Technology Plan of Fujian ,China(2001H023, 2001J005).

References

[1]Zhendong DONG and Qiang DONG. HowNet. http://www.keenage.com/html/index.html.
[2]A.T.Armapatzis and Th.P. van der Weide and C.H.A.Koster and P.van Bommel. Texts Filtering using Linguistically-Motivated Indexing Terms, http://citeseer.nj.nec.com
[3]Anandeep S.Pannu and Katia Sycara. A Learning Personal Agent for Texts Filtering and Notification, http://citeseer.nj.nec.com
[4]James Allen, Natural Language Understanding . The Benjamin/Cumming Publishing Company, Inc.
[5]Douglas W.Oard and Gary Marchionini, A Conceptual Framework for Texts Filtering, http://citeseer.nj.nec.com
[6]Thorsten Joachims, Texts categorization with support vector machines: Learning with many relevant features. http://citeseer.nj.nec.com