基於文本概念和 kNN 的跨語種文本過濾

Cross-Language Text Filtering Based on Text Concepts and kNN

蘇偉峰*, 李紹滋*, 李堂秋*, 尤文建*

Weifeng Su, Shaozi Li, Tanqiu Li, Wenjian You

摘　要
本文介紹一個可以從中文或英文大量的資訊中過濾出用戶的興趣所在的文檔的模型，用一簇可分義原向量空間的向量來表示用戶所感興趣的文本，然後把需要處理的文本也表示成一個可分義原空間中的一個向量，在向量空間中與 k 個最相近的向量進行計算，從而決定是否將該文本呈現給用戶。實驗證明，這是一個比較好的過濾方法。

關鍵字：可分義原，向量空間，kNN，文本表示，知網

Abstract
The WWW is increasingly being used source of information. The volume of information is accessed by users using direct manipulation tools. It is obviously that we’d like to have a tool to keep those texts we want and remove those texts we don’t want from so much information flow to us. This paper describes a module that sifts through large number of texts retrieved by the user.

The module is based on HowNet, a knowledge dictionary developed by Mr. Zhendong Dong. In this dictionary, the concept of a word is divided into sememes. In the philosophy of HowNet, all concepts in the world can be expressed by a combination more than 1500 sememes. Sememe is a very useful concept in settle the problem of synonym which is the most difficult problem in text filtering. We classified the set of sememes into two sets of sememes: classifiable sememes and unclassifiable semems. Classifiable sememes includes those sememes that are more...
useful in distinguishing a document’s class from other documents. Unclassifiable sememes include those sememes that have similar appearance in all documents. Classifiable includes about 800 sememes. We used these 800 classifiable sememes to build Classifiable Sememes Vector Space (CSVS).

A text is represented as a vector in the CSVS after the following step:

1. text preprosessing: Judge the language of the text and do some process attribute to its language.
2. Part-of-Speech tagging
3. keywords extraction
4. keyword sense disambiguation based on its environment by calculating its classifiable sememes relevance with it’s environment’s classifiable sememes. We add the weight of a semantic item if there are classifiable sememes the same as classifiable sememe in the its environment word’s semantic item. This is not a strict disambiguation algorithm. We just adjust the weights of those semantic items.
5. Those keywords are reduced to sememes and the weight of all keywords ‘s all semantic items ‘s classifiable sememes are calculated to be the weight of its vector feature.

A user provides some texts to express the text he interested in. They are all expressed as vectors in the CSVS. Then those vectors represent the user’s preference. The relevance of two texts can be measured by using the cosine angle between the two text’s vectors. When a new text comes, it is expressed as a vector in CSVS too. We find its k nearest neighbours in the texts provided by the user in the CSVS. Calculating the relevance of the new text to its k nearest neighbours and if it is bigger than a certain valve, than it means it is of the user’s interest if smaller, it means that it is not belong to the user’s interesting. The k is determined by calculated every training vector its neighbours.

Information filtering based on classifiable sememes has several advantage:

1. Low dimentional input space. We use 800 sememes instead of 10000 words.
2. Few irrelevant feature after the keyword extraction and unclassifiable sememes’s removal.
3. Document vector’s feature’s weight are big.
We made use of documents from eight different users in our experiments. All these users provided texts both in Chinese and English. We took into account the user’s feedback and got a result of about 88 percent of recall and precision. It demonstrates that this is a success method.

**Keywords:** Classifiable Sememe, Vector Space, kNN, Text Representation, HowNet

1. Introduction

With the rapid development of the Internet and other online information resources, a large amount of information is flowing towards people every day. According to statistics, each working person in the United States receives about 80 email messages a day, and among them there are a large number of useless spam messages. Obviously, if you go through these messages one by one, it will take a lot of time, and there is also a possibility of triggering certain virus emails that may destroy the computer system. Similarly, this situation occurs in many office environments, where many people hope to automatically extract the most interesting content from a lot of historical or received email messages by the computer.

Text filtering is an important method to automatically identify useful documents. Text filtering is the process of filtering out those documents most relevant to the user's needs and sending them to the user. Cross-language text filtering refers to the situation where the source information contains multiple languages (such as English, Chinese, etc.), or only one document contains multiple languages, and filtering out the documents that are most relevant to the user. The filtered documents may also be in multiple languages. In the unbounded Internet, cross-language filtering is particularly important. In the process of sending a large amount of information to the user, filtering out those things that are not of interest to the user before sending it is more efficient than filtering out those things after translating them into a certain language, which can save a lot of user's time and effort. Cross-language filtering systems are particularly important for users who need information in one language but do not understand the language well.

In the field of cross-language text filtering, people have explored many methods to achieve the mutual transformation of different languages. Initially, a method based on controlled vocabulary was proposed [TRANSLIB 1995], which converted each text into a fixed set of words, and the user's needs were also represented in fixed terminology. Then, matching was done. This method has the drawback that the vocabulary must be kept within a manageable size, and as soon as the vocabulary exceeds the manageable size, the recall and precision will decrease rapidly, and there is currently no good way to represent text as a vocabulary.

Next, people proposed the dictionary-based method [L. Ballesteros 1996], which compiled a multilingual dictionary to translate one language text into the other language text form, thus enabling single-language text filtering methods to be applied to multilingual text filtering. Theoretically, this is possible, but there are two aspects that limit its application. Firstly, the phenomenon of a word having multiple meanings, which can lead to a decrease in precision if all meanings are used, or a decrease in recall if only one meaning is used, or even a complete failure if the wrong meaning is chosen. Secondly, the phenomenon of a single word having multiple meanings, which leads to a decrease in recall among different authors using different words to express the same meaning.

This paper proposes a new idea. Instead of analyzing at the word level, we decompose the meaning contained in the text into a series of concepts, and then analyze the decomposed concepts to obtain the themes and characteristics of the text. In a similar way...
2. 过滤模型的系统结构

我们在一些自然学科中常用到，比如我们分析某种物质的性质时，我们经常从其构成的分子或原子水平的性质进行分析然后得到物质的性质。

2. 过滤模型的系统结构

我们的技术主要是向量空间模型，即文本表示成为向量空间中的一个向量。向量空间的优缺点是将文本内容转化为易于数学处理的向量方式，使得各种相似度计算和排序成为可能。因此，在文本检索、文本过滤和文本摘要等方面得到广泛应用，取得了良好效果。本文所提出的基于向量空间的文本过滤模型可以用于对中文和英文的文本进行过滤。其基本思想是首先利用用户所提供的材料来获取用户的模板，然后利用用户模板来判断某一文件是否符合用户模板。

我们采用了董振东先生所研究的《知网》[董振东 等]，该系统带有53000个中文词语和57000个英文单词。《知网》是一个以汉英语和英汉语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。《知网》采用义本来表示概念，义原始是基本的、不易再分割的单位，我们设想所有的概念都可以分解为各种各样的义原。董振东先生提取出了1500多个义原，并用它们的组合来表示所有的概念。比如它是这样注释“扭亏为盈”的：

DEF=alter|改变,StateIni=InDebt|盈,StateFin=earn|胜。

即是指“扭亏为盈”是一种“改变”，其起始状态是“盈”，最终状态是“胜”。

把概念分解成义原可以最大限度地解决一义多词的问题，比如“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。比如把“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。比如把“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。

把概念分解成义原可以解决一义多词的问题，比如“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。比如把“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。比如把“电脑”、“计算机”、“computer”这3个词，在《知网》中均定义为“computer|电脑”，像这样我们就可以把它视作或概念的义原。我们把这些概念分解成义原可以解决一义多词的问题。
從語料庫中隨機抽取 500 篇各類的文章，在對這些文章進行分詞之後，把詞分解成相應的義原，並對這些義原進行統計，若某個詞有歧義，則把該詞的所有意義進行歸一併將其義原加入統計，設定一高通篩檢程式，對於義原的統計值高於某個值的列為不可分義原，其餘的義原定為可分義原。

本文下面所用到的技術對於中文文檔和英文文檔同時適用，若是有不同的地方則會分別指出。

3. 過濾模型的設計和實現

3.1 文本表示方法

我們採用的技術是向量空間模型，文本表示為向量空間中的一個向量。向量空間表示為 \( \vec{D} \)，而每個分量 \( d_i \) 是知識中的一個可分義原，那文本就表示為向量 \( \vec{V} \)，其分量 \( v_i \) 為對應於 \( d_i \) 的值，若文本中沒有包含 \( d_i \)，則 \( v_i = 0 \)。

然而並非文件當中所有的詞都用於構造文本向量，只有那些最能代表文件所要表達的意思的詞也就是關鍵字彙可被用來構造向量。我們可以採用統計的方法來決定哪些詞彙是關鍵字彙，還有，由於詞彙的岐義，我們也要作一定程度上的排岐。文本表示方法可歸納如下:

文本預處理。對於中文文檔包括詞的切分、詞性標注，而對於英文文檔，則只進行詞性標注。

關鍵字提取。在英語文本中去除所有屬於下列的單詞：冠詞（如 a, the, an）、介詞或連接主句和從句的副詞（如 in, to, of）、情態動詞（如 would, must）和連接詞（如 and）等，在中文文本中去除所有的虛詞，這樣在文本中就剩下主要的詞像名詞、動詞、形容詞和副詞，形成關鍵字序。我們也可以給各種詞性的詞賦予不同的權值來表示它們不同的重要性，一般而言，名詞要賦以最大的權值。對於那些在標題、首段、末段、段首、段尾出現的詞彙也可以增加其權重。我們也可以設一個閥值，把那些出現頻率低於該頻率的詞去除。

關鍵字概念排岐。過多的歧義會損害我們向量表示的效果，尤其當某個詞在該文本當中佔有比較重要的地位時。排岐的基本思想是根據上下文文詞的義原對該詞的某一意思進行概率統計，其主要思想是：在一篇文章當中，某個詞會對上下文的用詞產生影響。通過上下文可以判定某個詞的意思從而進行排岐，在本模型下，着重考慮其上下文當中其他的關鍵字的義原與該詞的義原有以下情況：

a. 有相同可分義原，
b. 材料-成品關係，
c. 施事-受事關係，
d. 事件-領屬物等-事件關係，
e. 工具-事件關係，
苏伟峰

f. 场所-事件关系，
g. 时间-事件关系，
h. 事件-角色关系，
i. 相关关系。

如果其上下文的某个关键字当中有多个可分义原与该词的某个意思的某一可分义原有关，则增加该意思的权重。

在 W 中，对某单词 w，在以其为中心的窗口宽度为 n 的字符串表示为：

W/W2...Wn/2WWn/2+1...Wn/

对于 w 在知识中的每一个意思，赋予一个初始 k，调节词 w 每一个意思的权重的方法的伪代码算法 1 所示

算法 1：词的义原的权重的调节

W1—窗口中除去 w 的第 1 个词
S1J—窗口中除去 w 的第 1 词的第 J 个意思
CS1JK—窗口中除去 w 的第 1 词的第 J 个意思的第 K 个可分义原
WSJ—词 w 的第 J 个意思
WCSJK—词 w 的第 J 个意思的第 K 个可分义原
Weight(WCSJ)—词 w 的第 J 个意思的权重

FOR I=1 TO n-1 //对于窗口中除 w 外的每一个词
    FOR J=1 TO (W1的意思数目)
        FOR K=1 TO (S1J的可分义原数目)
            FOR M=1 TO (词 w 的意思数目)
                FOR O=1 TO (WSJ的可分义原数目)
                    IF CS1JK 与 WCSJK 有关则 Weight(WSJ) = Weight(WSJ) + 1
                    ENDIF
                ENDFOR
            ENDFOR
        ENDFOR
    ENDFOR
ENDFOR

由此，词语的那些与上下文相关的意思都通过增加权重而得到加强，当然我们要对此进行归一化处理，其归一化的公式如下所示：

\[
wt(WS_j) = \frac{Weight(WS_j)}{\sum_i Weight(WS_i)} \quad (1)
\]

其中 i 是该词的意思的序号。
文本表示成向量：在經過了關鍵字提取和排歧之後，我們把這些關鍵詞根據其義原權值按照知識的單詞定義分解成義原，在去除了不可分義原之後，我們採用演算法 2 中的方法計算各可分義原，文件就表示成了可分義原空間中的一個向量。

演算法 2 把一個文件表示成可分義原空間的一個向量

3.2 用戶模板表示
首先用戶提供 m 篇其所感興趣的文檔，爲了增加用戶興趣的文本在向量空間中的密度，一般要求 m>50，採用上文所述的方法把這些文本表示爲可分義原空間中的向量，這些向量就成了代表該用戶興趣的示例，我們稱其爲用戶示例。在進行文本過濾時，我們就是從用戶示例中找出 k 個與正在過濾的文本最爲鄰近的向量作爲鄰居向量進行分析。

3.3 文本相似度的計算
至此，文本已表示成可分向量空間中的一個向量，兩個文本的相似度可以通過公式（1）中的余弦值表示，其值越大，則表示這兩個文本的主題越相似，我們認爲它越是用戶所感興趣的文本。
\[ S_i = \sum_{m} s_i (\cos (a_i)) \]  (3)

其中

\[ S(x) = \begin{cases} 
0 & \text{当} x < h \text{ 时} \\
\frac{x}{h} & \text{当} x \geq h \text{ 时}
\end{cases} \]

在所需過濾的所有文本當中，我們可以根據 \( S_i \) 來進行相關度排序反饋給用戶，也可以設一閾值 \( t \)，當某文本與用戶需求的相關度大於 \( t \) 時則認爲該文本符合用戶需求，把文本按相關度大小的順序返回給用戶，把低於該值的所有文本去除或存在某處以備用戶在有空時處理。我們可以把用戶的回饋考慮進去，若用戶認為幾乎所有我們所過濾出的文件都是他所感興趣的，則我們可調低 \( t \) 值，反過來，若有很多文本不符合用戶的興趣，則我們調高 \( t \) 值。

3.4 文本類別的歸類

我們採用 kNN 的方法。首先我們訓練的時候，我們把這些已經分好類的按是否為用戶的需要全部按上述方法表示成可分義原向量空間的向量，對一新進來的一個新的文本，我們採用上面的方法轉化為可分義原向量空間中的空間向量，假設為 \( d \)，從中找出 \( k \) 個與其最為鄰近的向量，然後檢查這 \( k \) 個已經確定好類別的向量的類別作爲這個向量的類別。這 \( k \) 個向量的權重可以通過其與 \( d \) 的相近程度進行賦值。

kNN 是一個基於範例的學習法，其主要的計算量是從向量空間中找出 \( k \) 個最接近的鄰居時間複雜度為 \( O(L^2N) \)，其中 \( L \) 是可分向量空間的可分義原數目，\( N \) 爲可分向量空間中的訓練文本的數量。

\( k \) 值的確定方法：

我們主要採用登山法來確認 \( k \) 值。在訓練文本全部表示成向量空間的向量以後，按下面演算法進行計算：

```
算法 3 kNN 中的 k 的計算演算法
biggestequal:=0
bigestk :=0;
給向量的每個分量值賦初值 0
FOR k:=（一個>1 的小整數）TO （一個大整數）
k:=0;
FOR I=1 TO （訓練文本的數目）
  對於第 I 個訓練文本，計算 k 個最近鄰居，並利用 k 個鄰居的類別判定
  第 I 個文本的類別，如果相等，則 km:=km+1;
ENDFOR
```
If km>biggestequal then
Begin
biggestequal:=km;
bigestk:=k;
end;
ENDFOR

4. 過濾模型的實驗結果及實驗分析

我們獲得了八個用戶的實驗資料，這八個用戶都提供了自己所感興趣的內容相近的中英文文本各60篇作為相關文本，另外提供100篇其他內容的文書作為干擾文書，其中中英文各500篇，對於每個用戶，我們使用從其所提供的相關文書隨機抽取中英文文書各30篇構造其用戶模板，其餘的相關文書與干擾文書混雜一起構成了測試集，我們就想從其中過濾出那些相關文書。

我們使用了兩個參數來評價我們的模型：召回率和精確率。召回率是指我們過濾出的相關文書占所有相關文書的比率，精確率是指在我們所有過濾出的文書當中，相關文本所占的比率，一般而言，召回率上升，則精確率會下降，而精確率上升，則召回率會下降。

表1就是我們實驗的結果，結果表明用該方法進行過濾的方法效果非常好，精確率很高，在實際應用當中，我們還可以把用戶反饋的情況考慮進去，形成可根據用戶的興趣改變而把改變用戶模版向量從而改變選擇的文本的自適應系統。

|       | User 1 | User 2 | User 3 | User 4 | User 5 | User 6 | User 7 | User 8 | Average |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| 召回率(%) |        |        |        |        |        |        |        |        |         |
| English | 88.7   | 90     | 90     | 89     | 86     | 87     | 92     | 91     | 89.2    |
| Chinese | 86.6   | 91.5   | 86     | 85     | 84     | 87     | 90.6   | 90     | 87.6    |
| 精確率(%) |        |        |        |        |        |        |        |        |         |
| English | 86     | 88.6   | 85     | 88.7   | 87.5   | 88.5   | 84.7   | 90     | 88.5    |
| Chinese | 82     | 85.4   | 85     | 87.6   | 84.2   | 86.3   | 88.6   | 86.8   | 87.5    |

表1 使用該方法的八個用戶的召回率和精確率

我們可以從以下幾方面來分析這個過濾模型產生較好結果的原因：

1. 低維分析空間：所有的概念都被分解成義原，只須在可分義原空間中計算相似程度，這樣我們就只要計算600個左右的可分義原而不是100000個左右的中英文
單詞，如此降低維數可極大地提高召回率，還有，可以降低計算複雜度。

2. 相關成分值較大：比如在一篇病人上醫院看病的文本裏，可能會出現許多類似“病人”、“醫生”、“醫院”、“治療”等片語，這些詞都包含有“醫治”等義原，從而使“醫治”這個義原分量的值比較大，這樣就能突出本文的所要講述的內容主要是關於醫療這一方面的，這有助於提高精確率與召回率。

3. 千擾項較少：經過了關鍵字提取、詞語排斥和不可分義原的去除後，所剩下的義原大多與文本有重要的聯繫，而與文本相關度較少的其他分量的值相比較之下明顯較小。

在我們以前的工作當中，我們把用戶表示成一個向量，並以用戶向量與文本向量的夾角來表示文本與用戶的相關性。而採用了 kNN 技術，可在以下這些方面體現出其優勢：

1. 首先對於某一個用戶可能有比較廣泛的興趣，則取其平均向量可能會導致比較大的誤差。

2. 對於同一個領域，不同體裁的文章其在向量空間當中也有可能有較大的差距，取平均向量也會造成較大的誤差。

3. 如果用戶興趣產生變化，平均向量的改變較為遲緩，並且在這個過程當中也有較大的誤差。

而 kNN 則恰恰相反，

1. 若用戶有比較廣泛的興趣，則在向量空間當中形成不同類的向量，就可有不同的鄰居。

2. 對於同一領域而不同體裁的文章，也可在向量空間中形成不同類的向量，構成不同的鄰居。

3. 若用戶興趣發生變化，只要再次提供新的所興趣的文本，在向量空間當中幾乎不會受舊的向量的影響，且可保留舊的向量以備另用。

其優勢可在圖 1 和圖 2 體現出來。
基於文本概念和kNN的跨語種文本過濾

5. 結束語

從網路資訊服務需求出發，我們認爲有必要對資訊源的資訊進行過濾。本文提出了一個在可分義原空間中採用向量空間模型的方法進行文本過濾的模型，理論和實驗均表明，該模型具有比較好的過濾效果，從速度和服務性能上達到了較好程度。

在模型的實現過程中，我們發現把這種方法與關鍵字的方法相結合在相當程度上會提高過濾的性能，這將是我們下一步研究的目標。

參考文獻

TRANSLIB. “Advanced Tools for Accessing Multilingual Library Catalogues.” Technical Report, Deleveralbe D.1.4: Evaluation of Tools. Knowledge S.A., June 1995.
L. Ballesteros, W. B. Croft. “Dictionary-based methods for cross-lingual information retrieval.” Proc. Of the 7th Int. DEXA Conference on Database and Expert Systems Applications, 1996.

Douglas W. Oard, Gary Marchionini, “A Conceptual Framework for Text Filtering.” http://citeseer.nj.nec.com

张月杰、姚天顺 <基于特征相关性的汉语文本自动分类模型的研究>《小型微型计算机系统》, 1998年第8期

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

Anandeep S. Pannu and Katia Sycara. “A Learning Personal Agent for Texts Filtering and Notification.” http://citeseer.nj.nec.com

James Allen, Natural Language Understanding. The Benjamin/Cumming Publishing Company, Inc.

Thorsten Joachims. “Texts categorization with support vector machines: Learning with many relevant features.” http://citeseer.nj.nec.com

Douglas W. Oard and Nicholas DeClaris. “On Automatic Filtering of Multilingual.” http://citeseer.nj.nec.com

Ellen Riloff and Wendy Lehnert. “Information extraction as basis for high-precision text classification.” ACM Transactions on Information System, vol. 12, No 3, July 1994

Eui-Hong(Sam) Han, George Karypis and Vipin Kumar. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. http://citeseer.nj.nec.com

蘇偉峰、李紹滋、李堂慶、尤文建 <可分義原向量空間中的跨語種文本過濾模型>《自然語言理解與機器翻譯》2001年