Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 79-90

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: classfiable sememes and unclassficable semems. Classfiable sememes includes those sememes that are more

* 廈門大學計算機系
Department of Computer Science, Xiamen University , Xiamen，361005
Weifeng Su: waveletsu@263.net

useful in distinguishing a document's class from other documents. Unclassfiable sememes include those sememes that have similar appearance in all documents. Classfiable includes about 800 sememes. We used these 800 classficable sememes to build Classficable 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 provides 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: Classfiable Sememe, Vector Space, kNN, Text Representation, HowNet

1.引言

隨著因特網和其他在線資訊資源的迅猛發展，大量的資訊朝人們湧來，據統計美國每個上班人員平均每天收到 80 封電子郵件，當然裏面包含大量的無用的垃圾郵件，顯然如果對這些郵件逐一查看要花費很多時間，而且有可能激發某些病毒郵件從而破壞電腦系統，同樣的情景出現在許多辦公環境中，許多人希望能在許多的歷史或者別人送達的電子文本當中由電腦自動挑出自己最感興趣的內容。

文本過濾是自動分挑出有用的文本的一種很重要的方法。文本過濾是指從大量的源資訊中過濾出那些最符合用戶需求的資訊傳送給用戶，而跨語種文本過濾是指源資訊中包含多種語言（比如英語、漢語等），或者某個文本中就含有多種語言，從中過濾出用戶所需要的文本，過濾出的文本可能也是多種語言的。在沒有國界的因特網上，跨語種過濾出所要的資訊就顯得更爲重要。在把大量的資訊送給用戶之前過濾掉那些用戶不感興趣的東西，這比在有條件後，翻譯成某種語言過後再進行過濾更能省掉用戶大量的精力和時間，跨語種過濾系統對於那些對需要這一語種的資訊而又對該語言掌握得不好的用戶特別重要。

在跨語種文本過濾方面，人們已經摸索出了許多方法來實現不同語種之間的相互轉換形式。最初人們是提出一種基於控制辭彙的方法[TRANSLIB 1995]，即把文本表示成一些固定的詞，用戶的需求也表示成這些固定辭彙，然後進行匹配。這個方法最大的缺陷是辭彙必須在可管理的範圍之內，而一旦辭彙超出可管理的範圍，則其召回率和精確率則迅速下降，而且如何把文本表示爲辭彙目前也沒有一個很好的方法。

接著又有人提出基於字典的方法[L. Ballesteros 1996]，就是編輯一本多語字典把某種語言的文本表現形式通過翻譯表示成另一種語言的表現形式，從而使那些單種語言上的文本過濾技術可以應用於多語言的文本過濾，這個方法理論上是有可能的，但是有兩個方面的原因卻限制了它的應用。首先是一詞多義的現象，在翻譯中一個詞可能翻譯成幾個意思，若幾個意思全都採用則大大降低了精確率，若採用某一個意思，則有可能降低召回率，或者根本就選擇錯誤而導致召回率極低。第二是一義多詞的現象，由於不同的作者可能用不同的詞來表達同一個意思而導致召回率下降。

本文提出一種新的思路，我們不從詞這一級來分析概念，而是把詞所包含的概念進行分解，再對分解過後的概念進行分析，從而得到文本的主題和性質。其實類似的思想

在一些自然學科當中經常用到，比如我們分析某種物質的性質時，我們經常從其構成的分子或原子水平的性質進行分析然後再得到物質的性質。

2. 過濾模型的系統結構

我們採用的技術主要是向量空間模型，即文本表示成爲向量空間中的一個向量，向量空間的優點是將文本內容轉換成易爲數學處理的向量方式，使得各種相似運算和排序成爲可能。因此 ，在文本檢索、文本過濾和文本摘要等方面獲得廣泛應用，取得了良好效果。本文所提出的基於向量空間的文本過濾模型可以用於對中文和英文的文本進行過濾。其基本思想是首先利用用戶所提供的材料來獲取用戶的模板，然後利用用戶模板來判斷某一文件是否與用戶模板相近。

我們採用了董振東先生所研製的《知網》[董振東 等]，該系統帶有 53000 個中文片語和 57000 英語單詞。《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的常識知識庫，《知網》採用義原來表示概念，義原是最基本的、不易於再分割的意義的最小單位，我們設想所有的概念都可以分解成各種各樣的義原。董振東先生提取出了 1500 多個義原，並用它們的組合來表示世上所有的概念，比如它是這樣注釋“扭虧爲盈”的：

DEF=alter|改變,StateIni=InDebt|虧損,StateFin=earn|賺。

即是指“扭虧爲盈”是一種“改變”，其起始狀態是“虧損”，最終狀態是“賺”。

把概念分解成爲義原可以極大限度地解決一義多詞的問題，比如“電腦”、“電腦”、“computer”這三個詞，在《知網》裏均定義爲“computer|電腦”，這樣我們就可以把它們視爲概念等同的三個詞語。其實從這個層面上來理解，我們可以把某個詞的中英文意思同樣看作是一個一義多詞的一種形式，這樣只要解決好了排岐的問題，我們並不需要特殊處理就可以解決跨語種的問題，因而從義原這一個層面上來說我們的方法可以說是一種與語言無關的方法。

我們把義原繼續分爲兩類：可分義原和不可分義原。把義原分爲可分義原與不可分義原是從以下兩方面考慮的：

- 某一義原若是在某一類主題的文檔當中出現頻度越高，則認爲該義原與這一類主題越有關係。
- 某一義原若是在語料庫中所有的文檔當中出現的頻度越高，則認爲該義原在區分主題的作用越差。

因此不可分義原是指那些比較常見，沒法用來指出該概念一些特有的性質的義原，而可分義原則是指那些能表示該概念的重要的可與別的概念相區分的義原。若我們不排除掉不可分義原，則由於不可分義原的較高的出現頻率，就有可能誤導我們。可分義原在本系統中占著重要的地位。

可分義原和不可分義原的粗略判定方法如下：

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，若某個詞有岐義，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爲不可分義原，其餘的義原定爲可分義原。

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，文本表示爲向量空間中的一個向量。向量空間表示爲 $\vec{D}$ ，而每一個分量 d_i 是知網中的一個可分義原，那文本就表示成向量$\vec{V}$ ，其分量 v_i 爲對應於 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辭彙是關鍵字彙，還有，由於辭彙的岐義，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。對於中文文本包括詞的切分、詞性標注，而對於英文文本，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、段尾出現的詞語也可以增加其權重。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。過多的岐義會損害我們向量表示該文本的效果，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詞的義原對該詞爲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，著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無以下情況：

a.有相同可分義原，

b.材料-成品關係，

c.施事/經驗者/關係主體-事件關係，

d.受事/內容/領屬物等-事件關係，

e.工具-事件關係，

f.場所-事件關係，

g.時間-事件關係，

h.事件-角色關係，

i. 相關關係。

如果其上下文的某個關鍵字當中有個可分義原與該詞的某個意思的某一可分義原有上述關係，則增加該意思的權重。

在 **W** 中，對某個詞 *w*，在以其爲中心的窗口寬度爲 n 的字串表示爲：

$$w_1w_2...w_{n/2}ww_{n/2+1}...w_{n-1}$$

對於 *w* 在知網中的每一個意思，賦予一個初權 k，調節詞 *w* 每一個意思的權值的方法的僞代碼演算法 1 所示

演算法 1 ：詞的義原的權值的調節

W_I—窗口中除去 *w* 的第 I 個詞

S_{IJ}—窗口中除去 *w* 的第 I 詞的第 J 個意思

CS_{IJK}--窗口中除去 *w* 的第 I 詞的第 J 個意思的第 K 個可分義原

WS_J—詞 *w* 的第 J 個意思

WCS_{JK}--詞 *w* 的第 J 個意思的第 K 個可分義原

Weight(WCS_J)—詞 *w* 的第 J 個意思的權值

```
FOR I=1 TO n-1     //對於窗口中除了 w 外的每一個詞
  FOR J=1 TO   (WI的意思數目)
    FOR K=1 TO   ( SIJ的可分義原數目)
      FOR M=1 TO   (詞 w 的意思數目)
        FOR O=1 TO   (WSJ的可分義原數目)
          IF CSIJK 與 WCSJK 有上述關係 THEN Weight(WSJ)= Weight(WSJ)
                                              +1
          ENDIF
        ENDFOR
      ENDFOR
    ENDFOR
  ENDFOR
ENDFOR
```

由此，詞語的那些與上下文相關的意思都通過增加權值而得到加強，當然我們還要對此進行歸一化處理，其歸一化的公式如下所示：

$$wt(WS_i) = \frac{Weight(WS_i)}{\sum_i Weight(WS_i)} \qquad (1)$$

其中 i 是該詞的意思的序號。

文本表示成一向量。在經過了關鍵字提取和排岐之後，我們把這些關鍵詞根據其義原權值按照知網裏的單詞定義分解成爲義原，並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演算法

V_K— 向量中的分量的值

$SM_{I\,JK}$—第 I 個關鍵字第 J 個意思的第 K 個可分義原

Weightof(SM)—某個可分義原的標量值

$wt(S_{I\,J})$－第 I 個關鍵字第 J 個意思的權值

給向量的每個分量值賦初值 0

```
FOR I:=1 TO    (關鍵字的數目)
   FOR J=1 TO    (第 I 個關鍵字的意思總數)
      FOR K=1 TO (第 I 個關鍵字第 J 個意思)
           Weightof(SM_K)= Weightof(SM_K)+wt(S_IJ)
      ENDFOR
  ENDFOR
ENDFOR
```

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，我們認爲他們是越相近鄰居：

$$\cos(a) = \frac{(V_{user}, V_{text})}{|V_{user}||V_{text}|} \qquad (2)$$

其中 (V_{text1}, V_{text2}) 是指用戶向量和文本向量的內積，$|V_{text}|$ 表示文本向量的標量。

在文本過濾當中，我們採用了 k 個最近鄰居（kNN）的方法：對於某一輸入文本 s，按照上面所述的方法將其表示爲可分義原空間的向量，在用戶示例中，利用公式（2）挑選出 k（k<<m）個與之最相近的鄰居文本，根據公式（3）計算它與這 k 個文本的相似程度 *Si*，其值越高，則我們 r 認爲它越是用戶所感興趣的文本。

$$Si = \sum_{i=1}^{k} S^2(\cos(a_i)) \tag{3}$$

其中

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

在所需過濾的所有文本當中，我們可以根據 *Si* 來進行相關度排序反饋給用戶，也可以設一閥值 *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，若用戶認爲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爲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爲可分義原向量空間中的空間向量，假設爲 d，從中找出 k 個與其最爲鄰近的向量，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 個最近的鄰居時間複雜度爲 ***O(L*N)***，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：

演算法 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)
   km:=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他所感興趣的內容相近的中英文文本各 60 篇作爲相關文本，另外提供 1000 篇其他內容的文本作爲干擾文本，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本隨機抽取中英文文本各 30 篇構造其用戶模板，其餘的相關文本與干擾文本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本。

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本占所有相關文本的比率，精確率是指在我們所有過濾出的文本當中，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板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. **相關分量值較大**：比如在一篇病人上醫院去看病的文本裏，可能會會出現許多類似“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有率。
3. **干擾項較少**：經過了關鍵字提取、詞語排岐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之下明顯較小。

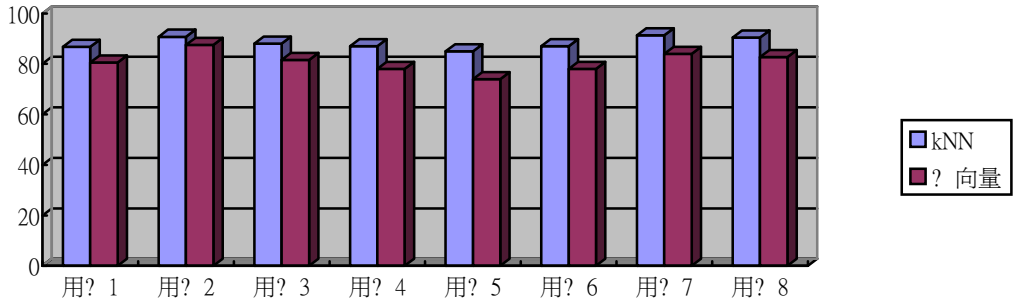
在我們以前的工作當中，我們把用戶表示成爲**一個**向量，並以用戶向量與文本向量的夾角來表示文本與用戶的相關性，而採用了 kNN 技術，可在以下這些方面體現出其優勢：

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，平均向量的改變較爲遲緩，並且在這個過程當中也有較大的誤差。

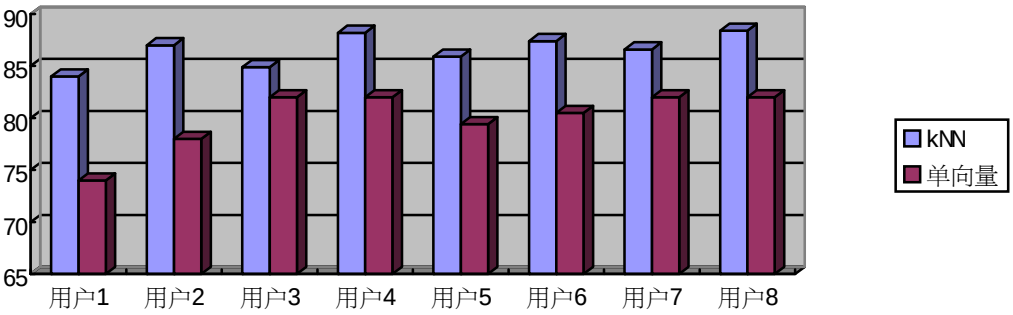
而 kNN 則恰恰相反，

1. 若用戶有比較廣泛的興趣，則在向量空間當中形成不同簇的向量，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簇的向量，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。



图*圖1 kNN 和單向量來表示用戶需求的召回率的比較*



图*圖2 kNN 和單向量來表示用戶需求的精確率的比較*

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 retieval.” *Proc. Of the 7 th Int. DEXA Conference on Database and Expert Systems Applications* ,1996.

董振東、董強 《知網》 http://www.keenage.com/html/index.html

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

張月傑、姚天順 ＜基於特徵相關性的漢語文本自動分類模型的研究＞《小型微型電腦系統》，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 Laguage 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 asbasis forhigh-precision textclassification.” *ACM Trams-actions on Information System*, vol. 12, No 3, July 1994

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

蘇偉峰、李紹滋、李堂秋、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