文字探勘運用在自動摘要之研究

王台平、張琬婷、吳雅茹
真理大學資訊管理學系助理教授、真理大學資訊管理學系、真理大學資訊管理學系
tpwangau@gmail.com、lovepurple3@yahoo.com.tw、yearu118@yahoo.com.tw

摘要

現今網路文件普及，資料量越來越龐大。使用者在網路搜尋引擎做查詢動作後，在閱讀冗長的查詢結果時總會耗費大量的時間及人力。摘要是最足以代表整篇文章的重點，使用者可以在閱讀摘要後，再決定是否深入文章本身，因此好的摘要方法變得十分重要。

本研究談論的是單篇文章中，三種不同摘要方法的比較，分別是句子與全文的關聯方法、句子與主題的關聯方法，以及兩者共同考量的方法。本研究針對三種方法分別進行實驗。以新聞文章作為實驗資料，文章經過前處理的斷詞之後，再透過 TF 及 ISF 的內部運算以及計算相似度，最後選擇相似度高的句子為文章重要摘要句。這三種不同的摘要結果共同與人工摘要句子作比較，經過精確度和回應率的計算，實驗結果顯示句子與全文、主題兩者共同考量下，為較佳的摘要方法。

關鍵詞：自動摘要、相似度、TF*ISF。

1. 前言

文件自動摘要技術成為流行技術。文件摘要有非常多的方法可應用；如萃取句中的重要關鍵詞、重要段落之詞及指引詞如結論、歸納等。統計的方法在近年來受到重視，主要是以機器學習之理論利用已摘要完成之文章作為該方法之學習目標，將摘要過程轉換成句子分類之問題，並可利用貝式分類器完成摘要工作。Salton(1997)利用傳統 TFIDF 以外的觀點，即內文關聯(intra-document association)的角度，提出了 Global Bushy Path(GPB)方法，以文本中段落與段落的關鍵詞重複率，作為衡量段落間相似程度的依據。

本研究將利用全文代表及主題關聯的方式為基礎來實驗，分別透過 TF*ISF 及相似度的計算，並與人工摘要比較來做量測，希望得到最接近人工摘要之結果。

2. 文獻回顧與探討

2.1 自動摘要

蘇謖(民國 85 年)教授在中國圖書館學會會報第 56 期發表文章「自動摘要法」中，提到摘要的功能需有宣示功能（Announcement）、篩檢功能（Screening）、取代功能（Substitution）、回溯功能（Retrospective）；而其摘要類型可分為指示性摘要（Indicative Abstract），資料性摘要（Informative Abstract），評論性摘要（Critical Abstract），摘錄（Extract）。指示性摘要通常具有宣示功能與篩檢功能；資料性摘要則主要是具有取代功能；評論性摘要則比較特別，是要以摘要的形式對原文作一評論，這類型摘要的自動化處理非常困難；摘錄則是直接取文章的句子，很可能具有宣示、篩檢、以及取代的功能，其功能則視情形而定。而回溯功能，可查詢原始文件，則是四種類型摘都要具備的功能。

本研究是屬於指示型的摘錄，根據文章中的關鍵詞的詞頻，計算出 TFISF 後，再依本研究三種不同考量的摘要方法「句子與全文的關聯方法」、「句子與主題的關聯方法」，以及「句子與全文、主題共同考量」，利用公式計算其相似度後(Sholom M. Weiss, Nitin Indurkhya, Tong Zhany & Fred J. Damerau, 2005)，挑選相似度較高的幾個句子為文章摘要。

2.2 中文斷詞

中文斷詞的方法一般分為法則式(Heuristic rule-based methods)(J. H. Zheng & F. F. Wu., 1999)、統計式(Statistical methods)(Fan, C. K. & W. H. Tsai, 1988)、混合式(Hybrid methods)(Nie, J. Y., M. L. Hannan & W. Jin, 1995)。

法則式斷詞的作法是利用現有詞庫，擷取出文章中出現在詞庫的字詞。優點是可保留文章中較完整的語意、操作簡單；缺點則是若有新詞，未被詞庫包含，就無法被辨識。統計式斷詞優點是不須參考任何的詞典及詞庫，大部分是針對二字詞來做處理所以執行效率較高且可保留較多的詞組合，但若要將此法擴充的話其執行效率便會因此而降低。缺點是需要較大的空間儲存統計模式。混合式斷詞結合了上述兩種斷詞法的優點，但其缺點則是在詞庫及語料庫擴充時，其執行速率會愈來愈慢(王台平、古祐嘉、王海霞，2005)。

在本研究中我們採用法則式斷詞為主，使用中央研究院中文詞庫小組的 CKIP 中文斷詞系統(1999)。

2.3TF*IDF

字組成詞，詞組成句，句組成段落，段落組成文章。每個中文字都有它獨特的意思，而由字組成的詞是最能表達意思的。對於字詞的研究，早期有學者提出詞頻(Term Frequency, TF) (Salton & McGill, 1983)與逆向文件頻率(Inverse Document Frequency, IDF) (Spark Jones, 1972)的觀念，

一般文本探勘中 TF 是計算詞頻，其詞頻越高，代表該詞在文件中愈重要。IDF 是詞在文件集中的區隔能力，該詞在太少文件中出現或太多（甚至於所有文件集都出現該詞）都變成不重要了。因此 TF*IDF 的值就可確認該詞在文本中及文件集中均佔有重要或不重要的地位。但有一點非常重要，就是文本有長文與短文，長文中詞頻很自然的就會較高，短文則相反。因此一定要做正規化的動作，如此才能真正顯現出該詞的重要性，而不會被長文誤導，同時探勘出來的結果也會較理想。

IDF 一般多應用於以多篇文章組成的文件集中，主要是將一篇文章當作一份文件，而本研究將 IDF 的理論，應用在本論文之 ISF(Inverse Sentence Frequency)上，ISF 是將一個句子當作一份文件，計算出詞在句子集(也就是整篇文章)，出現太多或太少次，本研究主要是以句子而非文章為單位。

2.4 內文關聯法

Salton(1997)利用傳統TFIDF以外的觀點提出了三種內文關聯法，分別是GBP(Global Bushy Path)、DFP(Depth First Path)、SBP(Segmented Bushy Path)。其中GBP是在一致性和廣度上都比DFP和SBP好，GBP方法是先挑選文章中連結點較多的或權重值最高的段落，成為摘要的首段，首段挑出來之後，再從文章中挑選第二多連結點或權重值次高的段落當作第二個段落，以此類推。後續研究者(黃純敏、楊存一、邱立豐，2002)提出改以句子為施策對象，依據關鍵詞共同出現的頻率，利用Jacquard’s coefficient法則衡量句子間的相似度，句子的權值越高代表該句愈重要。

一般摘要方法均會考慮將段落或文件的第一句和最後一句給予加權，然而本論文為初步研究，因此並不考慮權重值，並將利用全文代表及主題關聯的方式為基礎來實驗，希望得到較好的結果。

2.5 摘要評估

自動摘要之目的就是要比傳統人工摘要更有效率，且節省成本，因此應該將自動摘要與人工摘要做比較，才可提供較具有信服力的比較結果(黃純敏、吳郁瑩，1999)。但是人工在進行評估的時候，一定會有主觀的因素存在，因此評估的結果也可能未必正確。而假如不使用這個方式做評估，似乎在技術層面找不到更好的評估方式。因此本研究最後還是將摘要結果與人工摘要的結果相互比較並進行評估。

3. 研究方法及架構

3.1 研究方法

本研究使用 CIRB030 資訊檢索測試集(中研院新聞語料庫提供)，隨機挑選其中文章，分別經過中文斷詞、斷句、計算 TF、計算 TF*ISF，再依本研究三種不同考量的摘要方法：「句子與全文的關聯方法」、「句子與主題的關聯方法」，以及「句子與全文、主題共同考量」分別計算出相似度後，比較其摘要後之結果。

3.2 研究步驟

Step 1

資料來源：隨機挑選 CIRB030 資訊檢索測試集內文章 1000 篇新聞。

Step 2

中文斷詞：使用中央研究院詞庫小組所維護的中文自動斷詞系統 1.0 版，進行文章的斷詞處理。

Step 3

以句為單位：使用本研究所寫的程式，將文章輸出為一句一個文字檔，檔名以 000、001…的順序排序。

Step 4

計算 TF：透過 TF 程式(陳俊達，2004)分別計算出各文章的詞頻(term frequency)。

Step 5

計算 TF*ISF：透過 IDF 理論的程式(陳俊達，2004)來計算 ISF，以句子為單位，將一個句子當做一份文件，計算句子集合的 ISF，並執行雜訊過濾的動作。

Step 6

相似度計算：使用相似度計算公式，分別計算「句子與全文的關聯方法」、「句子與主題的關聯方法」，以及「句子與全文、主題共同考量」的相似度。

Step 7

實驗結果分析：將自動摘要出之結果與人工摘要比較過後，進行結果分析。

Step 8

提出結論。

3.3 研究架構

本研究之研究架構圖如圖 1 所示。首先蒐集短篇文章，將之透過中研院 CKIP 斷詞程式，CKIP 後的文章透過斷句程式，使文章以句子為單位來進行

實驗，經過 TFISF 後，再計算各別相似度，最後進行摘要評估。詳細說明請參考 3.3.1 至 3.3.6 小節。

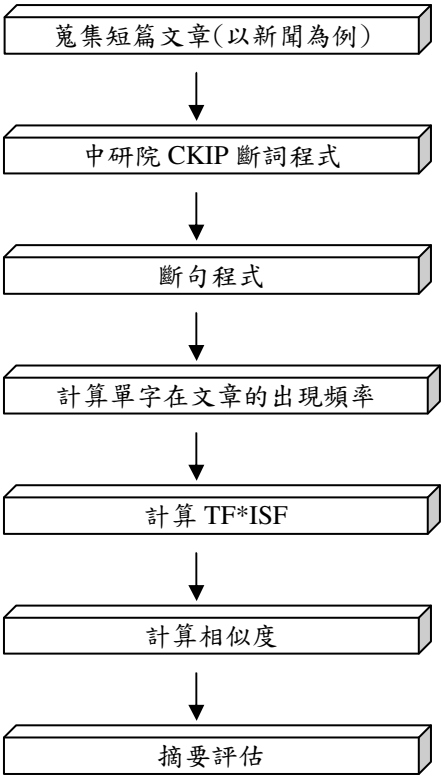

圖 1 研究架構圖

3.3.1CKIP

在文章原文中，將主題也視為一句，因此以人工作業在主題尾端加上句號，以利斷句。本研究使用中研院中文詞庫小組所提供之中文自動斷詞系統 1.0 版，將一篇文章的主題及文章內容貼入斷詞系統中，執行自動斷詞與標記，斷詞系統會從 10 萬目詞詞典與詞表中將詞作分類，保留需要的關鍵詞。

3.3.2 以句為單位

本研究使用斷句程式，可將 CKIP 後的文章斷成一句一個文字文件檔案，順序為 000.txt、001.txt.....。

斷句完之後，為求品質較高的實驗結果，句子低於八句的文章將不列入實驗中，並以人工作業過濾並將之刪除。

3.3.3TF 與 ISF

(1) Term Frequency(TF)：計算單字在某文件的出現頻率

$$TF_{ij} = \frac{n_j}{n_{all}}$$

代表單字 j 在文件 i 的出現頻率，其中

n_j ：表示單字 j 在文件 i 的出現次數。

n_{all} ：表示文件 i 所有具意義的總詞類。

執行 TF 程式，將文件所在磁碟設為 D 槽，再輸入要分析的文件總數(有幾句)，以名詞+動詞的形式出現，並予以正規化處理。

(2) Inverse Sentence Frequency(ISF)：計算逆向句子頻率

$$ISF_j = \log_2 \frac{N}{sf_j}$$

代表單字 j 在所有句子裡出現頻率的倒數

N ：代表所有句子的總數

sf_j ：代表單字 j 有出現過的句子總數

本研究以 IDF 之理論及工具，運用在本研究的 ISF 上面，以一個句子當作一份文件，並執行 TF*IDF 程式，將文件所在磁碟設為 D 槽，再輸入要分析的文件總數（有幾句），予以正規化處理，並取樣所有關鍵字來做分析。

3.3.4 相似度計算

將使用句子與全文的關聯方法、句子與主題的關聯方法、以及兩者共同考量的方法來做相似度計算。

相似性比對為將每一句轉換成關鍵詞之向量。

$$s_1 = \{(t_1, w_{11}), (t_2, w_{21}), (t_3, w_{31}), (t_4, w_{41}).....\}$$
$$s_2 = \{(t_1, w_{12}), (t_2, w_{22}), (t_3, w_{32}), (t_4, w_{42}).....\}$$
$$s_3 = \{(t_1, w_{13}), (t_2, w_{23}), (t_3, w_{33}), (t_4, w_{43}).....\}$$
$$\vdots$$

接著利用下列公式計算相似度(Sholom M. Weiss, Nitin Indurkhya, Tong Zhany & Fred J. Damerau, 2005)：

$$Sim(s_i, s_j) = \cos(s_i, s_j)$$
$$= \frac{w_{1,i} \times w_{1,j} + w_{2,i} \times w_{2,j} + w_{3,i} \times w_{3,j} + ...}{|s_i| \times |s_j|}$$
$$|s_i| = \sqrt{\sum_{k=1}^{m} w_{ki}^2} \qquad |s_j| = \sqrt{\sum_{k=1}^{m} w_{kj}^2}$$

3.3.5 自動摘要結果

Morris et al.(1992)認為摘要的長度至少應該是原文長度的 20-30%，才能適切的表達原文所要傳述的意思。所以本研究在計算完相似度後，選出佔文

章前 30%句相似度高的句子作為人工摘要。例如：文章共有 10 句，則取出相似度排名前三高的句子作為其文章之摘要。

3.3.6 評估方法

本研究使用 F-Measure 的評估方式包含回應率(recall radio)與精確度(precision radio)(Christopher Manning & Hinrich Scütze, 1999)，將摘要出之結果使用人工量測。

TP：代表應被選出的摘要句子被選出來之句子總數

FP：代表不應被摘要出來的句子被選出來之句子總數

FN：代表應被摘要出的句子未被選出之句子總數。

(1) 回應率(recall radio)

以人工量測摘要出來的句子作為基數，將三種方法所分別擷取的摘要句子與其相比較，可計算出三種方法摘要出來的句子正確率。根據上列之敘述，回應率為『TP/(TP+FN)』。

(2) 精確度(precision radio)

以三種方法摘要出來的句子作為基數，作法與回應率相似。根據上列之敘述，精確度為『TP/(TP+FP)』

(3) F-Measure

以 R 代表回應率、P 代表精確度，則 F-Measure 的公式為：

F-Measure = 2PR / (P+R)

有時候可能是回應率好、精確率差，也可能相反，很難判斷出成效。所以本研究使用 F-Measure 公式衡量結果的好壞。當回應率與精確度的值愈高時，F-Measure 的值也會愈高，這表示其摘要的結果越好。

本研究於人工量測時，總計 6 人參與量測工作，實屬於小樣本，為證實量測結果之精確度及回應率為有效值，本研究使用小樣本檢定來增加實驗結果的說服性。

使用一般小樣本檢定之 T 檢定，需使用 t 分配及自由度來找出合乎檢定之範圍；本研究共有 6 人來量測精確度與回應率，假設其精確度分別為 80、78.2、79.4、81、80.6、81.1，如我們用 80%來檢定即代表結果確定有 80%。設定信心度如 95%，找出自由度及 n-1 如 6 位專家(樣本數 n=6)，則自由度為 n-1=5=(6-1)，查統計表(t 分配)，t=0.025，自由度為 5，可得數值 2.5706，其值代表 6 位專家的量測對精確度 80%做檢定，檢定值在(-2.5706,2.5706)之間代表通過檢定。

依上述方法檢定回應率與精確度通過後，再以 F-Measure 值來做實驗結果比較。

4. 實驗步驟與結果

4.1 實驗資料

本研究隨機挑選 CIRB030 資訊檢索測試集內文章 1000 篇文章。進行到斷句前處理時，將不到八句的文章進行過濾刪除後，共 700 篇文章繼續進行實驗。

4.2 實驗設計

由於本研究為探討「句子與全文的關聯方法」、「句子與主題的關聯方法」，以及「句子與全文、主題共同考量」這三種方法下，自動摘要之結果好壞。因此將進行三組實驗並針對其結果做比較。

實驗一：

以 700 篇文章進行實驗，相似度計算時，以句子與全文考量之下做計算。

實驗二：

以 700 篇文章進行實驗，相似度計算時，以句子與主題考量之下做計算。

實驗三：

以 700 篇文章進行實驗，相似度計算時，以句子與全文、主題兩者共同考量之下做計算。

4.3 實驗步驟

Step1：隨機挑選 CIRB030 資訊檢索測試集內文章 1000 篇新聞。

Step2：利用中文自動斷詞系統，進行文章的斷詞處理。

Step3：利用程式，將文章斷成一句一個文字檔，不到八句的刪除，剩下 700 篇

Step4：透過 TF 程式計算文章的詞頻(term frequency)。

Step5：透過 IDF 理論之程式計算 ISF，計算關鍵詞在每句中出現的次數。

Step6：分別開始進行實驗一、實驗二、實驗三。

Step7：針對實驗檢定結果以 F 值做評估，比較其實驗結果。

4.4 實驗結果

(1) 實驗一：

以 700 篇文章進行實驗，相似度計算時，以句子與全文考量之下做計算，總計 6 人參與量測工作。實驗一量測結果如表 1 所示。

表 1「句子與全文的關聯方法」量測結果

值 / 人	precision	recall
1	0.5400	0.6279
2	0.6026	0.7109
3	0.5666	0.6812
4	0.4890	0.5982
5	0.5255	0.6486
6	0.5397	0.6837

實驗一檢定程序：

如表 1 之 6 人的量測，對精確度 54%及回應率 66%做檢定：

■ 精確度 T 檢定

t = [0.5439 - 0.55] / [0.0383 / sqrt(6)] = -0.3910。得檢定值為-0.3910，其數值落在(-2.5706, 2.5706)之間，代表通過檢定。在 95%的信心度之下，精確度為 54%是有效的。

■ 回應率 T 檢定

t = [0.6584 - 0.65] / [0.0417 / sqrt(6)] = 0.4941。得檢定值為 0.4941，其數值落在(-2.5706, 2.5706)之間，代表通過檢定。在 95%的信心度之下，回應率為 66%是有效的。

(2) 實驗二：

以 700 篇文章進行實驗，相似度計算時，以句子與主題考量之下做計算，總計 6 人參與量測工作。實驗二量測結果如表 2 所示。

表 2「句子與主題的關聯方法」量測結果

值 / 人	precision	recall
1	0.4871	0.5247
2	0.6111	0.5654
3	0.4797	0.5359
4	0.5102	0.5681
5	0.5605	0.5213
6	0.4827	0.4768

實驗二檢定程序：

如表 2 之 6 人的量測，對精確度 51%及回應率 53%做檢定：

■ 精確度 T 檢定

t = [0.5052 - 0.50] / [0.0306 / sqrt(6)] = 0.4193。得檢定值為 0.4193，其數值落在(-2.5706, 2.5706)之間，代表通過檢定。在 95%的信心度之下，精確度為 51%是有效的。

■ 回應率 T 檢定

t = [0.5320 - 0.55] / [0.0342 / sqrt(6)] = 1.2857。得檢定值為 1.2857，其數值落在(-2.5706,2.5706)之間，代表通過檢定。在 95%的信心度之下，回應率為 53%是有效的。

(3) 實驗三：

以 700 篇文章進行實驗，相似度計算時，以句子與全文、主題兩者共同考量之下做計算，總計 6 人參與量測工作。實驗三量測結果如表 3 所示。

表 3「句子與全文、主題共同考量」量測結果

值 / 人	precision	recall
1	0.5136	0.5836
2	0.5699	0.6816
3	0.6563	0.7795
4	0.6643	0.6158
5	0.6169	0.5759
6	0.6962	0.6455

實驗三檢定程序：

如表 3 之 6 人的量測，對精確度 62%及回應率 65%做檢定：

■ 精確度 T 檢定

t = [0.6195 - 0.6] / [0.068 / sqrt(6)] = 0.7039。得檢定值 0.7039，其數值落在(-2.5706, 2.5706)之間，代表通過檢定。在 95%的信心度之下，精確度為 62%是有效的。

■ 回應率 T 檢定

t = [0.6469 - 0.65] / [0.7674 / sqrt(6)] = -0.0098。得檢定值-0.0098，其數值落在(-2.5706, 2.5706)之間，代表通過檢定。在 95%的信心度之下，精確度為 65%是有效的。

通過檢定之三組實驗結果如表 4 所示：

表 4 實驗一、實驗二、實驗三之實驗結果

	實驗一	實驗二	實驗三
precision	54%	51%	62%
recall	66%	53%	65%
F-Measure	59.4%	52%	63.5%

5. 結論

本研究分別探討句子與全文的關聯方法、句子與主題的關聯方法，以及兩者共同考量的方法。為

求更好的摘要方法，在透過以上三組實驗過後，可以觀察到三種不同的摘要結果經過精確度和回應率的計算，共同與人工摘要句子作比較，最後實驗結果如圖 2 所示。其中實驗三之 F-Value 達 63.5%，即使其回應率較實驗二低了 1%，但精確度與其他兩種實驗結果比較過後亦為較高。綜觀來看，實驗三所顯示的結果是三種方法中較好的方法。這也代表著，當句子與全文作比較並共同考慮與主題的相關性，其選出之摘要句較接近人工摘要句，顯示句子與主題及全文的關係密不可分，在此三種方法中為較佳的摘要方法。

本研究所探討的三種方法，對於多語文(中英文)超文件自動摘要與評估(黃純敏、楊存一、邱立豐，89 年至 90 年)做比較，使用 GBP 方法後之平均精確度為 40.29%，平均回應率為 49.33%。本研究三種方法對其數值都有所改善，特別是實驗三中，句子與全文、主題共同考量的方法對 GBP 方法之精確度改善了 21.71%，而回應率改善了 15.67%。

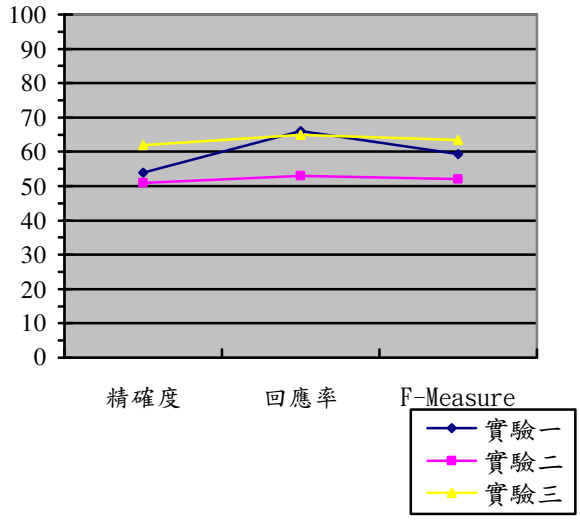

圖 2 三種實驗結果之折線圖

6. 致謝

致謝中央研究院 CKIP 中文詞庫小組所提供的 CKIP 中文斷詞系統、中華民國計算語言學會及中央研究院詞庫小組所提供的 CIRB030 資訊檢索測試集。

參考文獻

[1] 黃純敏、楊存一、邱立豐，多語文(中英文)超文件自動摘要與評估，行政院國家科學委員會補助專題研究計畫成果報告，個別型計畫，NSC 89-2416-H-224-053，89 年 8 月 1 日至 90 年 7 月 31 日。

[2] 黃純敏、楊存一、邱立豐，2002，TFIDF 與 GBP 方法於重要句子擷取績效評估，第十三屆國際資訊管理學術研討會，中華民國資訊管理學會主辦，淡江大學資訊管理學系承辦。

[3] 黃純敏、吳郁瑩(1999)，「網路中文文件自動摘要」，台灣區網際網路研討會 TANET’99，國立中山大學承辦。

[4] 中文自動斷詞系統 1.0 版，中研院詞庫小組，1999.11.1

[5] 程式來源:IDF v4.0，陳俊達，2004。

[6] 程式來源:term frequency v3.0，陳俊達，2004。

[7] 蘇誤，「自動摘要法」，中國圖書館學會會報第 56 期（民國 85 年 6 月），頁 41-47。

[8] 王台平、古祐嘉、王海霞，相對雜訊過濾法－以混合式技術改善文件聚類之精確度，第三屆演化式計算應用研討會暨 2005 機會探索國際工坊，真理大學，12/10/2005。

[9] Christopher Manning and Hinrich Scütze, “Foundations of Statistical Natural Language Processing”,The MIT Press, Cambridge, Massachusetts, London, England, 1999.

[10] Fan, C. K. and W. H. Tsai, “Automatic Word Identification in ChineseSentences by the Relaxation Technique,” Computer Processing of Chinese and Oriental Languages, Vol. 2, No. 4, pp. 33-56, 1988.

[11] J. H. Zheng and F. F. Wu. “Study on segmentation of ambiguous phrases with thecombinatorial type,”Collections of Papers on Computational Linguistics.Tsinghua University Press, Beijing, pp. 129-134, 1999

[12] Morris, A. H., Kasper, G., and Adams, D.(1992). The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance.Information Systems Research 3(1),pp.17-35.

[13] Nie, J. Y., M. L. Hannan and W. Jin, “Combining Dictionary, Rules and Statistical Information in Segmentation of Chinese,” Computer Processing of Chinese and Oriental Languages, Vol. 9, pp. 125-143, 1995.

[14] Salton, G., McGill, M., "Introduction to Modern Information Retrieval,"McGraw-Hill,1983.

[15] Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997). Automatic Text Structuring and Summarization. Information Processing & Management，33(2),193-207.

[16] Sholom M. Weiss, Nitin Indurkhya, Tong Zhany and Fred J. Damerau, “Text Mining-Predictive Methods for Analyzing Unstructured Information”, Chapter4.4.3-Consine Similarity,page 91-92, 2005.

[17] Spark-Jones, K.. “A statistical interpretation of term s specificity and its application in retrieval”, Journal of Documentation,1972,28(5):111–121.