以語意為基礎之網路犯罪資訊搜尋研究

On Semantic-Based Intelligent Crime Information Retrieval on the Internet

顏志平 Chih-Ping Yen
中央警察大學
電子計算機中心
peter@sun4.cpu.edu.tw

徐熊健 Shyong-Jian Shyu
銘傳大學
資訊管理研究所副教授
sjshyu@mcu.edu.tw

摘 要

檢警機關通常藉助入口網站的搜尋引擎，進行網際網路犯罪情報的搜集，然而這種搜尋引擎由於精確率及檢出率不高，所以往往回應許多不相關的網頁，致使偵辦人員需再耗費時間逐一過濾，相當不符效益，因此本文將運用智慧型的演算方式，來提高精確率及檢出率，以改善這個問題。

首先，利用語意場理論將詞與詞的同義關係，組織成語詞庫，建立起類似 WordNet 的階層式架構，同時使用這語詞庫，進行網頁內容的相似度比對。本文共推導五種相似度演算方式：包括「詞頻權重相似度」、「分類指數相似度」、「分類指數權重相似度」、「誤差校正相似度」、「詞頻權重重計」，並分別比較其間之優劣，擇出最佳的方法及推論出門檻值。

此外，本研究成果並與陳志誠於 1999 年之「網路上高精確率之犯罪資訊蒐尋系統」（稱為 e-Detective system）的研究計畫成果作比較，在以搜尋網際網路上「販售非法軟體」為例進行評估，實驗證明本文所建議之系統，其 F 測量值最佳達 0.5581，而前述系統最佳僅為 0.2376，顯然本研究系統效能較佳。

關鍵詞：搜尋引擎、資訊檢索、文件分類、相似度、精確率、檢出率、語意場、網路犯罪、電子偵探

Abstract

We usually search the Web with the help of search engines. Due to the imprecision of the search result, we often face the problem of too many pages recommended. The reason why

search engines response many irrelevant pages is that it just exactly matches the search word(s) user entered. In order to cope with the problem, we suggest the determination of similarities that should be associated with a knowledge base to a given topic. That will reduce the number of irrelevant pages significantly.

In this research we first apply to the theory of semantic fields in which a term (concept) forms a term database through its relationships to other concepts. Based on the term databases, we suggest several models to evaluate the similarity between search concepts and the contents of Web pages. They are the model of weighted terms (the modified vector space model), the model of classified weighted terms, and the exponential model of classified weighted terms. The latest one is designed based on to the Facet Analysis Method. We also evaluate the similarity with error correction and term reweighting. The approaches described in this paper are used to construct a search engine for discriminating Web pages advertising pirated compact discs (CDs) that are very difficult to be distinguished from the pages advertising legitimate CDs. We further determine an adequate threshold of term weights for our search purpose as a trade-off of recall and precision. Our search result compared with that of previous work shows the advantage of this approach.

Keywords: Search Engine, Information Retrieval, Text Classification, Similarity, Precision, Recall, Semantic Field, Cybercrime, *e-Detective*

1 緒論

網際網路上找尋資料，通常必須藉助入口網站的搜尋引擎如蕃薯藤[1]、奇摩Google[2]、Openfind[3]等，然而這些搜尋引擎由於精確率（precision）及檢出率（recall）不

[1] http://www.yam.com.tw
[2] http://www.kimo.com.tw
[3] http://www.openfind.com.tw

高，所以往往會回應太多不相關的網頁，例如回應網頁中含有同詞異義（homonymy）的情況，以「防火牆」而言，對於網路安全的「防火牆」，與消防安全所稱的「防火牆」就有極大不同，因此使用者必需再逐一過濾內容，運用上極為費時。

同樣地，檢警機關偵辦網路犯罪的人員，於搜尋引擎輸入關鍵詞（keyword），以搜尋網路犯罪資訊，亦會收到許多不相關之網頁，這些網頁可能偶然含有某些關鍵詞，或是同詞異義情形而被檢索到 [Lin et al. 1998] ，但是它們本身卻不含任何犯罪資訊。推究緣故，蓋因犯罪訊息往往是隱寓的，一篇網路犯罪的網頁文件中，可能全然不含任何與犯罪相關之詞語，例如：「徵求性伴侶」可能以「一夜情」稱呼；「販售非法軟體」可能以「大補帖」稱呼，因此一般搜尋引擎，只利用關鍵詞作比對是難以找到犯罪情報的，而如以「一夜情」、「大補帖」等語詞搜尋，則也僅能找到極少數的犯罪資訊。

有鑑於此，本研究首先從語言學的語意（semantics）觀點出發，並結合資訊檢索相似度（similarity）的相關理論，而提出相似程度演算方法，藉以提高精確率與檢出率；最後，使用大補帖網頁為例，並與陳志誠博士於 1999 年之「網路上高精確率之犯罪資訊蒐尋系統」（稱為 e-Detective system）的研究成果作比較，來評估及驗證我們建議的相似度演算方式較為優良。

2 文獻探討

本研究將運用不同學門，包括語言學（linguistics）、資訊檢索（information retrieval）等學科理論，藉以推導不同的相似程度演算。下面將就「資訊檢索技術」、「網路上高精確率之犯罪資訊蒐尋系統」及「犯罪網路語意分析」等文獻進行探討。

2.1 資訊檢索技術

資訊檢索係一門有關資訊之組織、架構、分析、儲存、搜尋以及傳播之科學[王朝煌 1998]。其理論及技術的發展已累積許多的成果，包括索引（indexing）技術、全文檢索（full-text retrieval）、自然語言處理（natural language processing）、跨語資訊檢索（cross-language information retrieval）、相關回饋（relevance feedback）、多媒體資訊檢索（multimedia information retrieval）及自動文件分類（automatic document classification）技術等 [Lesk 1995]。2.1.1 及 2.1.2 小節探討本研究所涉及之資訊檢索技術：自動文件分類及相關回饋，2.1.3 小節介紹檢索系統的績效評估。

2.1.1 自動文件分類技術

所謂自動文件分類，是將一群文件作類別區分，而要達成這樣的功能通常必須藉助資訊檢索技術，其技術方法甚多，包括布林（Boolean）模型、機率（probabilistic）模型、模糊數學（fuzzy set）模型、類神經網路（neural network）模型及向量空間（vector space）模型等，其中向量空間模型可說是最常用，以下將逐一說明，如何使用此模型，來進行文件分類 [Frants et al. 1997] 。

（1）文件向量表示

文件以數學向量表示，其方法是將文件集合中，所有 t 個不同語詞 $k_1, k_2, ..., k_t$，作為 t 維空間上座標軸的基底，而將每一個文件表示成 t 維空間上的一個向量，此向量中的每一個分量，代表是某一個特定語詞在該文件之份量。亦即：

$$\vec{d} = (w_{1,d}, w_{2,d}, ..., w_{i,d}, ... w_{t,d})$$

其中，$\vec{d}$ 代表文件 d 的向量表示，t 代表文件集合共有 t 個不同語詞，$w_{i,d}$ 代表文件 d 中語詞 k_i 的權重（weight）。

而欲決定上述權重，需考量兩個要素，一是語詞出現在該文件的詞頻（term frequency），以表示在該文件的相對重要性，其出現頻率愈高就愈重要；另一是語詞出現在整個文件集合的文件數目，稱反文件頻率（inverse document frequency），亦即該語詞愈少出現在別的文件中，就愈適合被用來與其它文件區分，代表對該文件愈重要。因此，根據上述兩個要素，我們可用其乘積的計算結果，來表示某一語詞在某一文件中的重要程度，此測量值稱為 $tf-idf$ [Salton 1973] [Salton 1975]，即權重之意，計算方式為

$$w_{i,j} = tf_{i,j} * idf_i$$

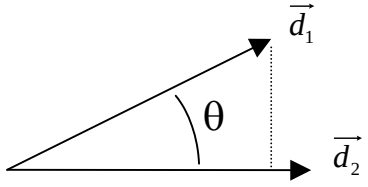
其中 $w_{i,j}$ 代表語詞 k_i 在文件 j 的權重，$tf_{i,j}$ 代表語詞 k_i 在文件 j 的詞頻，而 idf_i 代表語詞 k_i 的反文件頻率,並表示為

$$idf_i = \log \frac{N}{n}$$

N 代表文件集合的文件總數，n 代表文件集合中含有語詞 k_i 的文件數。

（2）文件相似度測量

為了要將一群文件分類至適當的類別，可用測量文件間的相似度來達成，一般常用的測量方法，就是前面所提的向量空間模型，此模型定義向量空間內兩文件 $\vec{d_1}$、$\vec{d_2}$ 的相似度為其投影量 cos(θ)



詳細的計算式如下

$$sim(d_1, d_2) = \frac{\vec{d_1} \bullet \vec{d_2}}{|\vec{d_1}| \times |\vec{d_2}|} = \frac{\sum_{i=1}^{t} w_{i,d_1} \times w_{i,d_2}}{\sqrt{\sum_{i=1}^{t} w_{i,d_1}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,d_2}^2}}$$

$sim(d_1, d_2)$ 代表文件 d_1、d_2 的相似度值，$\vec{d_1} \bullet \vec{d_2}$ 為兩文件的內積（inner product），$|\vec{d_1}|$、$|\vec{d_2}|$ 代表向量長度，w_{i,d_1} 代表文件 d_1 中語詞 k_i 的權重，w_{i,d_2} 代表文件 d_2 中語詞 k_i 的權

重。

（3）文件分類

基於向量空間模型，我們可以觀察到，屬於同一類別的文件，會聚集在一起形成叢集（cluster），在傳統文件分類方法中，即為每一叢集找出中心向量，以代表整個同類別的文件，當欲判斷一新文件之所屬類別時，祇要將其轉換為文件向量，並與每一叢集的中心向量計算相似度值，並將此新文件之類別歸屬於與其相似度最高之叢集。

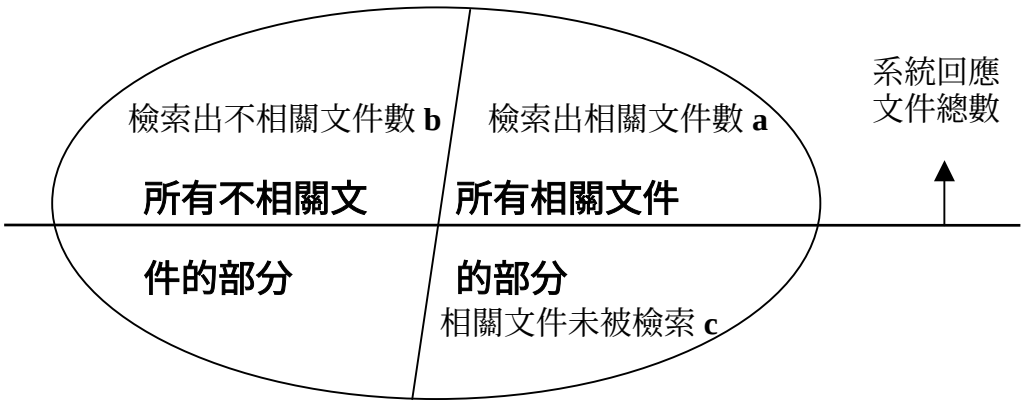
2.1.2 相關回饋

資訊檢索中，最受歡迎的查詢模式稱為「相關回饋」（relevance feedback），它被認為對提昇檢索之精確率及檢出率助益最大。進行的方式是使用者在前一階段查詢的回應文件中，挑選相關文件再回饋給系統即可，目的在找出更多相關的文件 [曾元顯 1997]，而其改善檢索的成效，主要是藉由兩大基本技術，「查詢擴展」（query expansion）及「詞頻權重重計」（term reweighting）的觀念。

所謂「查詢擴展」是將前次查詢所得到的相關網頁，重新擷取新詞後，再組合形成下一次新的查詢輸入，如此可不斷重覆進行多次，以改善系統效能。而詞頻權重重計技術，是將前次查詢所得到的相關網頁，從中找出符合於語詞庫的語詞，並改變語詞詞頻，如此作為下一次查詢比對時之依據，以改善系統效能 [Rocchio 1971]、[Ricardo 1999]。

2.1.3 檢索系統績效評估

如何判斷一個檢索系統之優劣，是資訊檢索領域中，一個相當重要的課題，目前最常使用的績效評估（evaluation）測量值，是以文件相關程度來判斷的精確率及檢出率，如圖 2-2，所謂精確率是指檢索到的文件中，相關文件的比例，檢出率則指相關文件被檢索出的比例 [Jones 1997] 、 [Rijsbergen 1979]。



$$\text{檢出率}=\frac{\text{檢索出相關文件數}}{\text{所有相關文件總數}}*100\%=\frac{a}{a+c}$$

$$\text{精確率}=\frac{\text{檢索出相關文件數}}{\text{系統回應文件總數}}*100\%=\frac{a}{a+b}$$

圖 2-2　精確率及檢出率

然而上述兩個測量值由於存在一種反比關係，因此許多研究學者，致力於找尋單一測量值，以用來正確評估檢索效能，其中 F 測量值（F measure）是常用的方式 [Shaw et al. 1997] ，它結合了精確率及檢出率，亦即計算兩者的調和平均數（harmonic mean），公式如下

$$F(j)=\frac{2}{\frac{1}{r(j)}+\frac{1}{P(j)}}$$

其中 $r(j)$ 表示系統檢索出相關文件數中，在第 j 個文件的檢出率， $P(j)$ 表示系統檢索出相關文件數中，在第 j 個文件的精確率，而 $F(j)$ 則是 $r(j)$ 及 $P(j)$ 的調和平均數，當 $r(j)$ 及 $P(j)$ 值愈大， $F(j)$ 值也愈高，據此以尋求最佳的精確率及檢出率。

2.2 網路上高精確率之犯罪資訊蒐尋系統

國科會於民國 88 年委託研究一個「網路上高精確率之犯罪資訊蒐尋系統」的先導計畫 [陳志誠 1999] 。該計畫分析出網路犯罪語詞及其詞頻權重，並建置成語詞庫，藉由比對網頁的語詞及累計其詞頻權重，以計算每一網頁的總權重值，最後經由統計精確率及檢出率，來推論、驗證屬於網路犯罪網頁的門檻權重值 [Chen 2000]，該計畫完成下列工作：

（1）測試集與樣本

為檢驗、評估檢索系統效能，通常必須在一個規範化的測試環境中進行實驗，目前國外已有一些機構，建置了許多標準測試集（collection set），如歐美的 CACM、MEDLARS、TIPSTER、TREC、AMARYLLIS、日本的 NTCIR 等，而且這些測試集也趨向多主題領域及大型化發展，像美國的 TREC、法國的 AMARYLLIS、日本的 NTCIR 均屬之。但對於中文而言，尚無一個公認的標準測試集，因此該計畫根據一般入口網站（portal site）分類目錄內的網址，蒐集 53,240 個網頁，總儲存空間大小約 6.8G，以作為系統績效評估的測試集來源。另外，此研究自刑事警察局破獲的網路犯罪案件裡，選取大補帖樣本網頁 20 個，並從中擷取相關的語詞，及統計語詞出現的個數及詞頻（各語詞累積的次數，與所有語詞累積次數的比值）百分比，以作為比對大補帖網頁之語詞

來源。

（2）門檻值及網頁權重值

該研究從實驗結果中，首先決定一個門檻值（threshold），作為判斷是否為網路犯罪之網頁；接著，凡是欲比對的網頁，經語詞比對及網頁權重值計算後，只要網頁權重值大於門檻值就認定為可疑網頁，例如大補帖網頁門檻值，經實驗結果為 13.8，而網頁權重值之計算公式如下：

$$w = \sum_{i=1}^{t} n_i * f_i$$

其中，w 代表比對網頁的權重值，t 表語詞庫中不同語詞的個數，n_i 代表比對網頁中出現語詞庫第 i 個語詞的個數，f_i 表語詞庫中第 i 個語詞的詞頻。

上述研究提出一個網頁經過語詞比對後，以出現語詞次數與其詞頻的乘積，來表示為網頁權重值，並從網頁權重值大小，判斷是否為欲搜尋標的的網頁，並實際測試的結果顯示，此系統對於網頁內容及網站分類的判斷上，其精確率均較一般蒐尋引擎為高。我們認為這方式的缺點是同一語詞若重複出現在比對網頁中，就有權重值偏高的現象。例如圖 2-1 的比對網頁範例，語詞「CD」出現 6 次，則「CD」詞頻百分比（該研究指出為 10.24）乘上出現 6 次，結果為 61.44，大於門檻值 13.8，然此網頁並不是搜尋標的（非大補帖網頁）。

圖 2-1　比對網頁範例（資料來源：http://yesup.asiainf.com/chinese/music/），其中語詞「CD」出現 6 次

上述同一語詞重複出現所造成的缺點，我們將於下一章設計相似程度演算方式，而予以改善，並用數學模式給予較嚴謹的定義。

2.3 網路犯罪語意分析

傳統語意學通常把詞看作是語意的最小單位，但是當代語意學則有許多不同看法的理論提出 [何三本 1995] 。如「語意成分」（semantic features）理論是把詞看作是許多對立的「語意成分」組合；並置理論（collocational theory）認為某些詞與某些詞間是有一些相關聯繫的，像「夜」、「晚上」、「黑暗的」就常並置來搭配使用；語意場（semantic fields）理論，則指透過詞與詞的關係所形成的一個語意範圍，這些關係包括同義關係（synonym）、反義關係（antonym）及上、下義關係（hypernym、hyponym），例如在「動物」這個概念下，貓、狗、馬、虎、象等詞，即能構成一個同義的語意場，「冷的、熱的」、「溫的、涼的」各為一對反義關係，能構成一個「冷的、熱的、溫的、涼的」反義的語意場，而「花」與「玫瑰」、「玉蘭」、「牡丹」、「桃花」、「鬱金香」等，存在下義關係，即「花」是上義詞，其它種類的花是下義詞，這說明語意結構是個多層級架構，它反映了詞之間的等級關係。

目前，應用這語意場概念，最知名的莫過於 WordNet [Miller 1990]，它是以階層式架構來組織同義詞，打破傳統的扁平式分類，許多研究者亦紛紛依此方式，從現有扁平式的同義詞典中，拉起類似 WordNet 的階層式架構，以建立語詞庫，作為相關研究之用 [Farreres et al. 1998]。同樣地，如「色情網站」、「販賣大補帖」等犯罪，我們可以分析、歸納其相關文件中那些語詞是同義關係，以形成不同的同義語意場，並賦予各語意場上義詞，如此便能清楚、整體地描述這些犯罪。以「販賣大補帖」的語詞為例 [顏志平 2001] ，我們可從達成犯罪目的之標的物「媒體技術」、透過何種「交易方式」、如何吸引購買者的「廣告行銷」、所採用「數量單位」，來分析其上義詞及其所屬的相關語詞，其分類結果如表 2-1。

表 2-1　大補帖之上義詞及其所屬的相關語詞

上義詞	語　意　場（同義關係）
媒體技術	MP3、VCD、DVD、金片、藍片、原版、容量、光碟、對拷、燒錄、挑片、錄製、音樂、光碟、破解、影片、大碟、CD…
交易方式	付款、郵購、訂單、匯款、郵局、代收、E-mail、來信、郵筒、交貨、郵遞、區號、索取、訂購、回信、電話、郵差、賣…
廣告行銷	麵店、保證、退貨、優惠、便宜、精選、自選、專輯、麵攤，泡麵、珍藏、要快、試聽、不看、可惜、片名、最新…
數量單位	單價、每張、一套、全套、單片、中文版、片裝、元、價錢、幾、原版、片、時、天、日期、滿滿、每碗、破解版、集、MOD…

另外，使用上述語意場理論究竟有什麼好處呢？我們將藉由物件導向分析（object-oriented analysis）的「概念」（concept）來說明。所謂「概念」是一種我們對事

物的特別看法及了解，也就是當我們形成了某一個概念，相對地，就能相應出的所屬事物 [Martin 1992]，例如我們有了一個「國會」的概念，就能相映出有關之事物。同時，不同的概念可以組合，以形成新的概念，而且有愈多的概念，就能愈清楚描述所指的事物，所以當我們又加入新的概念「現任」、「女性」，則所指「國會」、「現任」、「女性」就更清楚知道是描述在國會中的女性人物了。因此，所謂上義詞不就是一個「概念」的意思，當上義詞愈多就愈能夠清楚描述出所指的對象。

隨後將於下一節，引用本節「語意場」、「概念」的這些觀點，設計我們所建議之相似度計算。

3 相似度計算

資訊檢索系統通常使用相似度的測量值，來鑑別兩文件的相關程度，因此相似度的計算方式，將攸關系統的效能。本章將分為五小節，前三小節描述研究過程中，逐步推導相似程度的最佳計算方法。最後二小節描述前面最佳相似程度計算下，如何再提昇精確率與檢出率的輔助方法。

第 3.1 節應用 Gerard Salton 基本向量空間模型的相似度運算概念 [Salton 1983]，我們稱為「詞頻權重相似度」（similarity of term weight），以賦予 2.2 節「網路上高精確率之犯罪資訊蒐尋系統」計畫 [陳志誠 1999]，一個較嚴謹的數學定義，經由標準化計算，改善該研究計畫多個同一語詞重複計數的問題。

第 3.2 節描述前小節相似度計算方法的缺點，即非搜尋標的的比對網頁中，少數卻有高詞頻的語詞，會造成相似度值依然偏高的現象。因此建議採用「分類指數相似度」（similarity of classification with exponent）計算，本方法是先將語詞庫的語詞，依不同之「上義詞」關係作分類，再利用此語詞庫檢測比對網頁，統計各分類中語詞出現的種類個數，藉以計算網頁相似度，改善精確率與檢出率。

第 3.3 節描述前小節分類指數相似度亦有其缺點，即不同網頁但相似度值一樣時，無法區別出含有較高語詞詞頻的網頁，應有較高之相似度值。因此建議加入詞頻權重觀念，本研究稱為「分類指數權重相似度」（similarity of classification with exponential weight），以辨別出含有高語詞詞頻應有較高相似度。

第 3.4 節描述具「歧義性」（ambiguity）的這一類網頁（如大補帖與正版唱片、軟體網頁），仍會被系統檢索出來，也就是合法之唱片及軟體網頁，仍會計得極高相似度值。因此，建議找出合法網頁與非法網頁間之特徵語詞，同樣形成一個含有多個上義詞的誤差校正語詞庫，再利用誤差校正（error correction）相似度計算，以有效鑑別（discriminating）合法與非法網頁的情形。

第 3.5 節描述從檢索到的相關網頁裡，將出現在語詞庫的語詞擷取出來，並重新計算語詞庫的詞頻權重（term reweighting），藉以提昇精確率與檢出率。

3.1 詞頻權重相似度

在 2.1 節該蒐集一些網路犯罪的樣本（sample）網頁，如大補帖、網路色情等，並由專業的網路犯罪偵查人員，就樣本網頁擷取出網路犯罪相關的語詞（term）。另外，我們將予以統計詞頻（frequency）及計算詞頻權重（term weight），且依此建置成不同的網路犯罪語詞庫。下面將用數學模式給予較嚴謹的定義。

我們以 S 表示語詞庫，其語詞集合 $S=\{k_1,k_2,...,k_t\}$，其中 k_i 為第 i 個語詞，$1\le i\le t$，以 $freq_{i,S}$ 表示語詞庫 S 中語詞 k_i 的詞頻，而 $f_{i,S}$ 則是語詞庫 S 中語詞 k_i 的詞頻權重，表示如下：

$$f_{i,S}=\frac{freq_{i,S}}{freq_{\max,S}} \tag{3.1}$$

$$freq_{\max,S}=\max\{freq_{j,S}\mid 1\le j\le t\}$$

其中 $freq_{\max,S}$ 指語詞庫 S 中的最大詞頻，所以詞頻權重 $f_{i,S}$ 的範圍應是介於 0 與 1 間（$0< f_{i,S}\le 1$）。如表 3-1，是以大補帖樣本網頁所建成的語詞庫（僅列出部分），語詞「CD」擁有最大的詞頻 24，所以詞頻權重為 1（=24/24）。

表 3-1　大補帖網頁語詞庫，其相關之語詞、詞頻及詞頻權重

語詞	詞頻	詞頻權重	語詞	詞頻	詞頻權重
CD	24	24/24=1.0000	原版	4	4/24=0.1667
音樂	16	16/24=0.6667	熱門	3	3/24=0.1250
訂購	13	13/24=0.5417	金片	3	3/24=0.1250
精選集	12	12/24=0.5000	補麵	2	2/24=0.0833
購買	10	10/24=0.4167	中文版	2	2/24=0.0833
貨款	7	7/24=0.2917	……	……	……
燒錄	5	5/24=0.2083	每碗	1	1/24=0.0417

我們以 P_r 表示第 r 個比對網頁，用 P_r 代表比對網頁 P_r 內含有上述語詞庫中語詞的集合，$P_r\subseteq S$，當我們比對輸入的網頁時，則是應用 Gerard Salton 基本向量空間模型的相似度運算概念，以求得一新相似度值，並藉由此相似度值的大小，判斷何者最為相似。我們將語詞庫 S 內的所有 t 個語詞，表示為 t 維空間上座標軸的基底，並令語詞庫 S 的所有語詞之詞頻、詞頻權重分別表示為向量

$$\overrightarrow{F_S}=(freq_{1,S},freq_{2,S},...,freq_{i,S},...,freq_{t,S})$$
$$\overrightarrow{W_S}=(f_{1,S},f_{2,S},...,f_{i,S},...,f_{t,S})$$

而比對的網頁 P_r，其對應於語詞庫 S 的所有語詞詞頻、詞頻權重表示為向量

$\overrightarrow{F_{P_r}} = (freq_{1,P_r}, freq_{2,P_r}, ..., freq_{i,P_r}, ..., freq_{t,P_r})$

$\overrightarrow{W_{P_r}} = (f_{1,P_r}, f_{2,P_r}, ..., f_{i,P_r}, ..., f_{t,P_r})$

其中 $freq_{i,P_r}$ 指第 r 個比對網頁中語詞 k_i 的詞頻（僅統計出現在語詞庫 S 中的語詞），f_{i,P_r} 指第 r 個比對網頁中語詞 k_i 的詞頻權重，表示如下：

$$f_{i,P_r} = \frac{freq_{i,P_r}}{freq_{\max,S}}$$

同時，若 $freq_{i,P_r} > freq_{\max,S}$ 則令 $f_{i,P_r} = 1$。前述詞頻權重計算方式，雖應用了基本向量空間模型的概念，但使用上並不採用 2.2.1 節所說明的 idf，因為我們將語詞庫視為整個文件集合，亦即文件集合中的文件總數只有 1，所以在任何語詞之 idf 均為 0 的情形下，我們並沒有使用它。在 3.5 節本研究將以詞頻權重重計方法校正，校正本假設可能造成之誤差。

接著在語詞庫中的語詞間均無關係下，令語詞庫詞頻權重向量 $\overrightarrow{W_S}$，與比對網頁詞頻權重向量 $\overrightarrow{W_{P_r}}$ 的夾角為 θ，則其相似度即定義為 cos(θ)

$$sim(S, P_r) = \cos(\theta) = \frac{\overrightarrow{W_S} \bullet \overrightarrow{W_{P_r}}}{\left|\overrightarrow{W_S}\right| \times \left|\overrightarrow{W_{P_r}}\right|} = \frac{\sum_{i=1}^{t} f_{i,S} \times f_{i,P_r}}{\sqrt{\sum_{i=1}^{t} f_{i,S}^{\,2}} \times \sqrt{\sum_{i=1}^{t} f_{i,P_r}^{\,2}}} \qquad (3.2)$$

其中 $\left|\overrightarrow{W_S}\right|$ 及 $\left|\overrightarrow{W_{P_r}}\right|$ 分別表示語詞庫與比對網頁的向量長度，所以 $sim(S, P_r)$ 即是一個標準化的相似度，當 cos(θ) 愈大，兩向量夾角 θ 就愈小，則我們認為比對網頁與語詞庫所要表徵的某一特性網頁（如利用大補帖網頁所建構而成的語詞庫）愈接近。我們以範例 1 說明相似度的計算。

範例 1：

假設語詞庫 S 含有 10 個不同的語詞及其詞頻，分別為 CD（24 個）、音樂（16 個）、訂購（13 個）、精選集（12 個）、購買（10 個）、貨款（7 個）、燒錄（5 個）、原版（4 個）、熱門（3 個）、每碗（1 個），且依前述方法，每一個維度代表一個不同的語詞，因此本語詞庫即可表示為 10 維空間座標上的向量，其語詞詞頻向量如下：

$\overrightarrow{F_S}$ =(24, 16, 13, 12, 10, 7, 5, 4, 3, 1)　而 $freq_{\max,S}$ =24 所以詞頻權重向量為

$\overrightarrow{W_S}$ =(24/24, 16/24, 13/24, 12/24, 10/24, 7/24, 5/24, 4/24, 3/24, 1/24)

=(1.0000, 0.6667, 0.5417, 0.5000, 0.4167, 0.2917, 0.2083, 0.1667, 0.1250, 0.0417)

若有兩個比對網頁 P_a、P_b 之詞頻向量為

$\overrightarrow{F_{P_a}}$ =(2, 1, 0, 0, 0, 0, 0, 0, 0, 1)　　$\overrightarrow{F_{P_b}}$ =(2, 0, 1, 0, 0, 0, 0, 0, 0, 1)

詞頻權重向量為

$\overrightarrow{W_{P_a}}$ =(2/24, 1/24, 0, 0, 0, 0, 0, 0, 0, 1/24)=(0.0833, 0.0417, 0, 0, 0, 0, 0, 0, 0, 0.0417)

$\overrightarrow{W_{P_b}}$ =(2/24, 0, 1/24, 0, 0, 0, 0, 0, 0, 1/24)=(0.0833, 0, 0.0417, 0, 0, 0, 0, 0, 0, 0.0417)

則相似度分別為

$$sim(S,P_a)=\frac{(1\times 0.0833)+(0.6667\times 0.0417)+(0.0417\times 0.0417)}{\sqrt{2.3351}\sqrt{0.0104}}\approx 0.7235$$

$$sim(S,P_b)=\frac{(1\times 0.0833)+(0.5417\times 0.0417)+(0.0417\times 0.0417)}{\sqrt{2.3351}\sqrt{0.0104}}\approx 0.6829$$

亦即 $sim(S,P_a)>sim(S,P_b)$，我們從而推測網頁 P_a 較網頁 P_b 更相似於搜尋標的。上述方法雖有標準化性質，能改善 2.1.4 節多個同一語詞重複計數問題，但仍有其缺點，即當一比對網頁中，只要包含有任一個高詞頻的語詞，則會立即使相似度值亦變得很大，例如圖 3-1 之比對網頁（以 P_g 表示），其詞頻權重向量為

$\overrightarrow{W_{P_g}}$ =(1/24, 0, 0, 0, 0, 0, 0, 0, 0, 0)=(0.0417, 0, 0, 0, 0, 0, 0, 0, 0, 0)

圖 3-1　比對網頁 P_g 範例（資料來源：http://www.cdnet.com.tw/），其中語詞「CD」出現 1 次

則相似度為

$$sim(S,P_g)=\frac{(1\times 0.0417)}{\sqrt{2.3351}\sqrt{0.0017}}\approx 0.6566$$

結果顯示，當比對網頁中一旦出現有語詞「CD」，其基本的相似度就高達 0.6566，然此網頁不是搜尋標的（非大補帖網頁）。相對地，當一比對網頁中，雖然包含許多低詞頻的語詞，但相似度值仍舊很小，例如圖 3-2 之比對網頁（以 P_q 表示），包含有語詞「訂購」、「燒錄」、「原版」，其詞頻權重向量為

$\overrightarrow{W_{P_q}}$ =(0, 0, 1/24, 0, 0, 0, 1/24, 2/24, 0, 0)=(0, 0, 0.0417, 0, 0, 0, 0.0417, 0.0833, 0, 0)

圖 3-2　比對網頁 P_q 範例（資料來源：http://buy.faithweb.com/），其中語詞「訂購」出現 1 次、「燒錄」出現 1 次、「原版」出現 2 次

則相似度為

$$sim(S, P_q) = \frac{(0.5417 \times 0.0417) + (0.2083 \times 0.0417) + (0.1667 \times 0.0833)}{\sqrt{2.3351}\sqrt{0.0104}} \approx 0.2823$$

結果顯示，雖然比對網頁 P_q 中包含「訂購」、「原版」、「燒錄」等語詞，為搜尋的標的（大補帖網頁），但其相似度仍祇有 0.2823，遠小於前述非大補帖網頁 P_g。

3.2 分類指數相似度

上一節相似度計算的方法有其缺點，即一個非搜尋標的的比對網頁中，祇要含有少數高詞頻的語詞，其相似度值就會變大，甚至大於搜尋標的的網頁，我們在本節中將找出改善的方法。

3.2.1 建立階層式詞語庫

一般而言，一個網頁內容，均有作者所欲傳達的主題，而圍繞在這主題下，就有與主題相關（relevance to a subject）的一些描述，即 2.3 節所稱之「概念」，若其它網頁有愈多與此網頁一樣的相關描述，就愈可能與此網頁相似 [Martin 1992]。是故，本研究語詞庫將依不同主題的概念，作同義詞的分類，其方式是參酌語意場理論 [何三本 1995]，及類似 WordNet 階層式的同義詞庫，將不同之上義詞關係作一般化的分類，來建立本研究語詞庫的階層關係。

以大補帖網頁為例，當一個網頁欲達成販售大補帖目的，其網頁內容描述，應含括犯罪目的之過程中，所有或部分要件，因此我們必需一般化大補帖語詞庫中的語詞，以輪廓出這些要件，亦即找出共同語詞的上義詞。透過網頁的內容分析，我們歸納出大補帖網頁應包含有標的物的「媒體技術」、透過何種「交易方式」、如何吸引購買者的「廣告行銷」、所採用「數量單位」，最後再將所有語詞，依此四個上義詞、在網頁的那段主題描述及參考 WordNet 作分類及歸屬，如圖 3-3，而表 3-2 是所有上義詞及其所包含的語詞、詞頻。

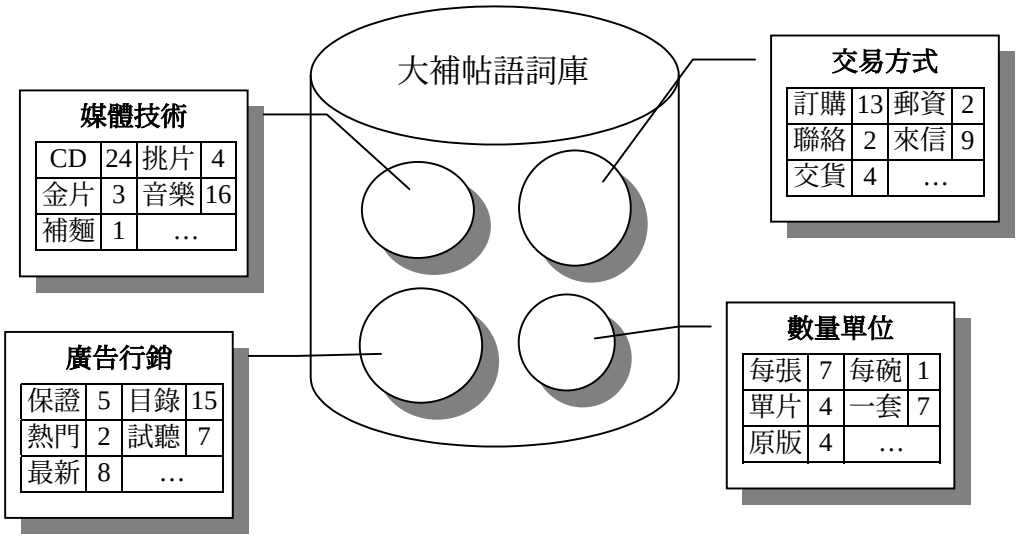


圖 3-3　大補帖語詞庫的上義詞一般化

表 3-2　經上義詞一般化的大補帖語詞庫

上義詞	語詞	詞頻	語詞	詞頻	語詞	詞頻	語詞	詞頻	語詞	詞頻
媒體技術	補麵	1	可夾檔	1	光碟容量	3	金綠片	1	音樂	13
	MP3	23	光碟	33	巫毒卡	4	挑片	4	CD	24
	VCD	165	光碟片	6	金片	3	負荷	3	容量	9
	備份	3	無法讀取	3	電影	11	燒錄	9	麵麵	1
	讀取	7	對拷	5	撥放程式	1	藍片	1	新麵	2
	測試	7	錄製	3						
廣	專輯	102	保證退貨	2	畫質有保障	1	影片教學	1	心動不如馬上行動	1
	HOT	38	要買要快	3	絕不能錯過	1	熱門	2	送禮自用兩相宜	1

廣告行銷	專輯	102	保證退貨	2	畫質有保障	1	影片教學	1	心動不如馬上行動	1
	HOT	38	要買要快	3	絕不能錯過	1	熱門	2	送禮自用兩相宜	1
	不看可惜	3	重量級軟體	1	超動感	1	編號	17	沒問題	3
	不看會後悔	1	限制級	1	超強	2	自選	21	更新	2
	中文電腦字幕	1	便宜又大碗	1	超清析	1	燒錄軟體	2	舞曲音樂選輯	1
	完整目錄	1	資料庫	1	量身定做	1	選輯	30	名稱	1
	片子	7	選輯	1	搶購	1	選錄	1	通通	1
	片名	3	倍數	1	新增	2	馬上	1	蓋不受理	3
	包退包換	1	值得收藏	1	保證	5	隨你挑選	1	最新	8
	未滿十八歲	1	作業規定	3	會員	5	應有盡有	1	歌曲容量	3
	目錄	15	原聲帶	22	概不退還	5	舊雨新知	1	歌詞	3
	特價	5	恕不服務	1	試試	2	雙重火力	1	舞曲	6
	列表	1	特惠價	2	試聽	3	穩定	1	精選輯	24
	好片	1	特價風暴	2	資料片	6	麵店	2	無碼	1
	收藏	5	索取	11	資料缺一	3	寄錯概不退還	2	畫質	1
	有問題	4	免費更換	3	隔天馬上寄出	1	便宜	2	電影原聲帶	19
	自由挑選	1	來看看哦	2	電腦修正	1				
交易方式	E-MAIL	2	來信	9	郵費	2	包裹	1	訂製	6
	一次限購	1	來信索取	4	郵資	13	另加郵資	3	訂購	10
	郵寄	1	呼叫器	1	郵遞區號	3	外縣市	1	資料	1
	公平交易	1	姓名	8	傳真機號碼	1	本區	6	需知	1
	付款	2	直接交貨	1	意者回信	1	交易	3	訂購辦法	1
	出售	2	訂單	7	詳細地址	1	交貨	4	時間地點	1
	出貨	3	訂單內容	3	電子信箱	1	方式	2	價目	1
	全套購買	3	寄	1	告知	1	地點	1	真實姓名	4
	回信	3	賣	2	地址	6	任你選	1	售價	2
	反映	1	行動電話	1	電話	13	聯絡	2	郵局代收	5
	含郵	5	安全可靠	1	預先通知	1	貨到付款	2	註明	2
	貨款	10	郵筒	2						
數量單位	MOD	7	專業中文版	1	全套	3	中文商業版	1	破解版	3
	光碟完整版	2	專業版	2	完整版	2	片裝	38	動畫版	2
	V8 版	1	備份版	2	完整破解版	2	加長版	1	商業版	1
	一套	7	單片	4	每片	7	正式版	21	國際中文版	1
	中文正式版	10	集	21	每張	7	光碟版	1	珍藏版	3
	中文版	2	元/片	1	每碗	1	英文正式版	1	原版	4
	滿滿	3								

3.2.2 相似度

若以S 表示語詞庫中語詞之集合，則每一上義詞即為集合S 之一個分割（partition），記之為$\{S_1, S_2, ..., S_i, ..., S_m\}$，$m$ 為語詞庫之上義詞個數，S_i 為第i 個子集合，$1 \le i \le m$，亦即這些非空S_i 的聯集就等於S

$$S = \bigcup_{i=1}^{m} S_i$$

且對任何相異的S_i 及S_j 而言，其交集等於空集合

$$S_i \cap S_j = \phi \quad for\ S_i \neq S_j$$

同樣地，對於一個比對網頁 P_r 內，出現含有語詞庫 S 的語詞集合為 P_r，$P_r \subseteq S$，亦可

$$sim(S,P_r) = SIM(S_1,P_{r_1}) + SIM(S_2,P_{r_2}) + ... + SIM(S_i,P_{r_i}) + ... + SIM(S_m,P_{r_m})$$
$$= \sum_{i=1}^{m} SIM(S_i,P_{r_i}) \qquad (3.3)$$

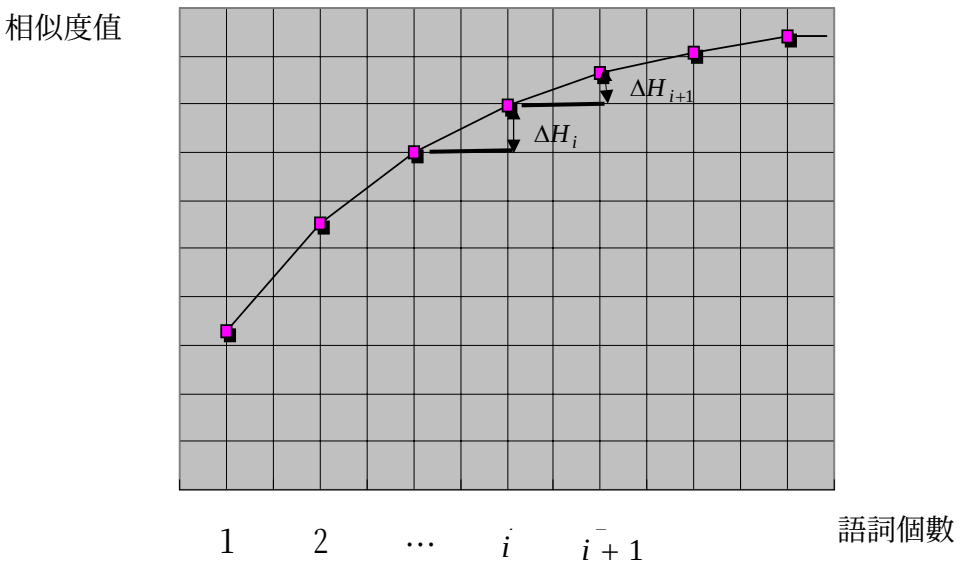
依 S 的上義詞分類方式，作集合分割並記為 $\{P_{r_1}, P_{r_2}, ...P_{r_i}, ..., P_{r_m}\}$。基於此，將語詞庫 S 與輸入比對網頁 P_r 的相似度計算，應可看待為兩者間，其對應類別的相似度值總和。

接著定義一個 u_i，$0 < u_i \leq 1$，以表示其各類別的重要程度，因此(3.3)式，改寫為(3.4)式是標準化的表示，至於 $SIM(S_i,P_{r_i})$，要如何求得？我們由上一節可知，更多概

$$SIM(S,P_r) = \frac{\sum_{i=1}^{m} SIM(S_i,P_{r_i}) * u_i}{\sum_{i=1}^{m} u_i} \qquad (3.4)$$

念即表示愈多的上義詞，它愈能清楚、明確辨認出所指的事物，反過來說，要愈明確指出特定事物，也就要愈多的上義詞。因此，若所要比對網頁中的語詞，在愈多不同的上義詞子集合中出現，我們推測其有較高的可能為大補帖網頁；而同一語詞出現的次數，對相似度的貢獻不必是線性（linear）的關係，我們認為以指數（exponent）函數的關係，應可更貼切地表現語詞出現次數對相似度的影響，即某一上義詞中的語詞，一旦出現在比對網頁，則該網頁所求得的相似度值，應大於第二次語詞出現所增加的相似度值。如圖 3-4 所示，第 i 次出現對相似程度的貢獻 ΔH_i，應大於第 $i+1$ 次出現所增加的程度 ΔH_{i+1}。

圖 3-4　以指數函數關係表示在特定上義詞分類中，不重覆之



語詞個數與相似度值的關係，其中 $\Delta H_i > \Delta H_{i+1}$

據此，我們就可將語詞庫 S 第 i 個分類 S_i，與比對網頁 P_r 的第 i 個分類 P_{r_i} 之相似度值表示為

$$SIM\ (S_i, P_{r_i}) = 1 - e^{-n_i}$$

$$n_i = \left| S_i \cap P_{r_i} \right| \qquad (3.5)$$

(3.5)式中 n_i 代表子集合 S_i（即第 i 個上義詞），與比對網頁子集合 P_{r_i} 交集的語詞個數。而(3.4)式即可進一步改寫為

$$SIM(S, P_r) = \frac{\sum_{i=1}^{m} SIM\ (S_i, P_{r_i}) * u_i}{\sum_{i=1}^{m} u_i} = \frac{\sum_{i=1}^{m} (1 - e^{-n_i}) * u_i}{\sum_{i=1}^{m} u_i} \qquad (3.6)$$

我們在範例 2 中，用上述方法重新計算範例 1 的相似度值

範例 2：

令 S ={ {CD,音樂,燒錄}媒體技術,{精選集,熱門}廣告行銷,{購買,訂購,貨款}交易方式,{原版,每碗}數量單位 }

即 S 的分割分別為

S_1 ={CD,音樂,燒錄}　　S_2 ={精選集,熱門}

S_3 ={購買,訂購,貨款}　　S_4 ={原版,每碗}

且設 S_1 至 S_4 的類別重要程度 u_i 均為 1。另比對網頁 P_a 有語詞「CD、音樂、每碗」、P_b 有語詞「CD、訂購、每碗」，其不重覆的語詞集合為

P_a ={ {CD,音樂}媒體技術,{每碗}數量單位}

P_b ={ {CD}媒體技術,{訂購}交易方式,{每碗}數量單位}

即 P_a 及 P_b 的分割分別為

P_{a_1} ={CD,音樂}　$P_{a_2} = \phi$　$P_{a_3} = \phi$　P_{a_4} ={每碗}

P_{b_1} ={CD}　$P_{b_2} = \phi$　P_{b_3} ={訂購}　P_{b_4} ={每碗}

經(3.6)式之相似度計算為

$$SIM\ (S, P_a) = \frac{\sum_{i=1}^{m} (1 - e^{-n_i}) * u_i}{\sum_{i=1}^{m} u_i}$$

$$SIM\ (S, P_b) = \frac{\sum_{i=1}^{m} (1 - e^{-n_i}) * u_i}{\sum_{i=1}^{m} u_i}$$

$$= \frac{(1 - e^{-\left|S_1 \cap P_{a_1}\right|}) + (1 - e^{-\left|S_2 \cap P_{a_2}\right|}) + (1 - e^{-\left|S_3 \cap P_{a_3}\right|}) + (1 - e^{-\left|S_4 \cap P_{a_4}\right|})}{4}$$

$$= \frac{(1 - e^{-\left|S_1 \cap P_{b_1}\right|}) + (1 - e^{-\left|S_2 \cap P_{b_2}\right|}) + (1 - e^{-\left|S_3 \cap P_{b_3}\right|}) + (1 - e^{-\left|S_4 \cap P_{b_4}\right|})}{4}$$

$$= \frac{(1 - e^{-2}) + 0 + 0 + (1 - e^{-1})}{4} \approx \frac{0.7442}{4} \approx 0.4740$$

亦即 $SIM(S,P_a) < SIM(S,P_b)$， P_b 這網頁是較 P_a 網頁更相似大補帖網頁，這與 3-1 節同樣的例子結果卻迥然相異，而下一節實驗將對更多網頁做實際測試，以證明(3.6)式優於(3.2)式。

本小節之分類指數相似度之計算源於「面向分析法」(facet analysis method)[Chen 2002]，當以一個物件 A 作基準來衡量另一物件 B 時，B 的性質(property)可以用量化方式表示。亦即將 A 之性質分成幾個面向(facet)，每一面向佔有一定之權重，且總權重為 1。在每一面向中可能有一個或若干個構成元素(element)，每一構成元素亦賦予一定之權重，而物件 B 之性質，可以由其所擁有 A 之構成元素的權重表示。

3.3 分類指數權重相似度

若另有一比對網頁 P_c，其不重覆的語詞集合為

P_c ={ {CD,燒錄}媒體技術,{每碗}數量單位}

相似度為

$$SIM(S,P_c) = \frac{\sum_{i=1}^{m}(1-e^{-n_i})*u_i}{\sum_{i=1}^{m}u_i}$$

$$= \frac{(1-e^{-|S_1 \cap P_{c_1}|}) + (1-e^{-|S_2 \cap P_{c_2}|}) + (1-e^{-|S_3 \cap P_{c_3}|}) + (1-e^{-|S_4 \cap P_{a_4}|})}{4}$$

$$= \frac{(1-e^{-2}) + 0 + 0 + (1-e^{-1})}{4} \approx 0.3742$$

經(3.6)式相似度的運算， $SIM(S,P_a)$ 等於 $SIM(S,P_c)$。然而 P_a 應較 P_c 來的更相似於大補帖網頁，因為(3.6)式只取語詞出現的個數，不看各個語詞的權重，但是 P_{a_1} 的「音樂」卻較 P_{c_1} 的「燒錄」要有較高的詞頻權重，所以 P_{a_1} 的媒體技術類相似度 $SIM(S_1,P_{a_1})$ 應大於 P_{c_1} 的相似度 $SIM(S_1,P_{c_1})$，最後總計各分類相似度 $SIM(S,P_a)$ 會大於 $SIM(S,P_c)$。

因此我們應把比對網頁計總的詞頻權重考量進來，以改善判別力；圖 3-5 之指數函數的曲線趨勢，呈現隨著所計得的詞頻權重而變動的結果。當權重愈高者，採用之指數函數曲率愈大，權重愈低者，採用之指數函數曲率則愈小。

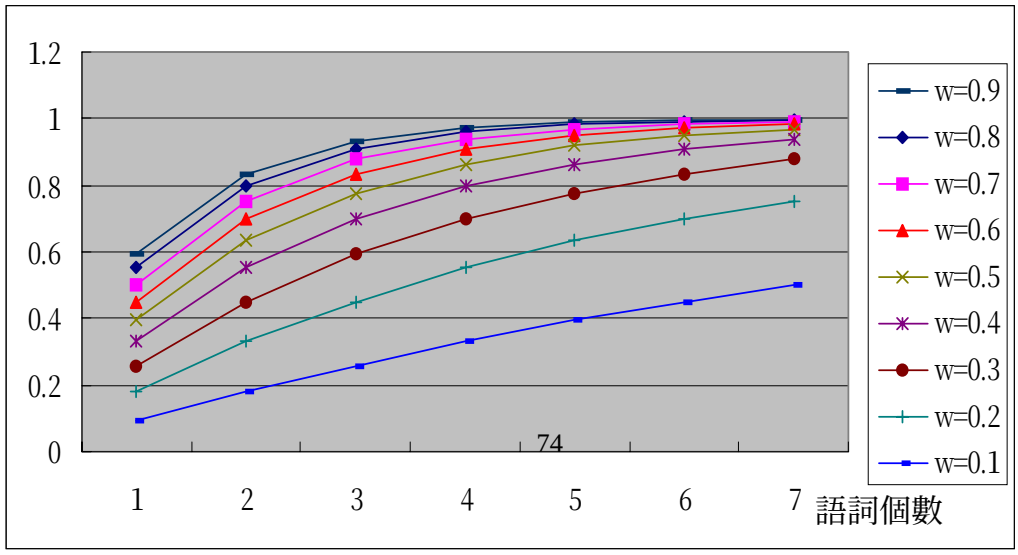


圖 3-5 以權重決定使用之指數函數曲率，其中 $0 < w < 1$ 本圖例

示 9 種不同權重的情形

如上所述，我們將各分類計總的詞頻權重加入(3-6)式後，表示為

$$SIM\ (S, P_r) = \frac{\sum_{i=1}^{m} (1 - e^{-n_i * w_i}) * u_i}{\sum_{i=1}^{m} u_i} \qquad (3.7)$$

(3.7)式中 w_i 即是比對網頁 P_r 在類別 P_{r_i} 所計總的詞頻權重，而且此詞頻權重是標準化的結果（$0 < w_i < 1$）。所以當比對到某類語詞個數 3，且詞頻權重總計為 0.3 時，相似度值約 0.6 ($1 - e^{-3*0.4} \approx 0.6$);當比對到某類語詞個數 2，且詞頻權重總計為 0.8，則相似度值約為 0.8 ($1 - e^{-2*0.8} \approx 0.8$)。至於 w_i 要如何計得，我們引用(3.2)式，將計得的相似度值直接設為該分類的詞頻權重，因此表示為

$$w_i = \frac{\overrightarrow{W_{S_i}} \bullet \overrightarrow{W_{P_{r_i}}}}{\left|\overrightarrow{W_{S_i}}\right| \times \left|\overrightarrow{W_{P_{r_i}}}\right|} = \frac{\sum_{j=1}^{|S_i|} f_{j,S_i} \times f_{j,P_{r_i}}}{\sqrt{\sum_{j=1}^{|S_i|} f_{j,S_i}^{\ 2}} \times \sqrt{\sum_{j=1}^{|S_i|} f_{j,P_{r_i}}^{\ 2}}} \qquad (3.8)$$

(3.8)式 S_i 是第 i 個上義詞分類語詞庫，$\overrightarrow{W_{S_i}}$ 是該分類之語詞詞頻權重向量；P_{r_i} 指比對網頁中有出現 S_i 中語詞的集合，$\overrightarrow{W_{P_{r_i}}}$ 是 P_{r_i} 中語詞的詞頻權重向量

$\overrightarrow{W_{S_i}} = (f_{1,S_i}, f_{2,S_i}, ..., f_{i,S_i}, ..., f_{|S_i|,S_i})$

$\overrightarrow{W_{P_{r_i}}} = (f_{1,P_{r_i}}, f_{2,P_{r_i}}, ..., f_{i,P_{r_i}}, ..., f_{|S_i|,P_{r_i}})$

其中 $|S_i|$ 是指分類語詞庫 S_i 的語詞數目，f_{i,S_i} 指 S_i 中語詞 k_i 的詞頻權重，$f_{i,P_{r_i}}$ 即為 P_{r_i} 中語詞 k_i 詞頻權重。範例 3 將利用範例 2 的例子，說明本相似度計算。

範例 3：

語詞庫 S 及各分類的語詞集合如範例 2 所示為

S ={ {CD,音樂,燒錄}媒體技術,{精選集,熱門}廣告行銷,{購買,訂購,貨款}交易方式,{原版,每碗}數量單位 }

即 S 的分割 S_1、S_2、S_3 與 S_4 分別為「媒體技術」、「廣告行銷」、「交易方式」、「數量單位」，而個別的詞頻向量為

$\overrightarrow{F_{S_1}}$ =(24, 16, 5)　　$\overrightarrow{F_{S_2}}$ =(12, 3)

$\overrightarrow{F_{S_3}}$ =(10, 13, 7)　　$\overrightarrow{F_{S_4}}$ =(4, 1)

當 $freq_{\max, S}$ =24，詞頻權重向量為

$\overrightarrow{W_{S_1}} = \overrightarrow{W_{媒體技術}}$ =(1.0000, 0.6667, 0.2083)　　$\overrightarrow{W_{S_2}} = \overrightarrow{W_{廣告行銷}}$ = (0.5000, 0.1250)

$\overrightarrow{W_{S_3}} = \overrightarrow{W_{交易方式}}$ =(0.4167, 0.5417, 0.2917)　　$\overrightarrow{W_{S_4}} = \overrightarrow{W_{數量單位}}$ =(0.1667, 0.0417)

令兩個比對網頁 P_a、P_c，其各分類的語詞集合及詞頻向量為

P_a ={ {CD,音樂}媒體技術,{每碗}數量單位}

P_c ={ {CD,燒錄}媒體技術,{每碗}數量單位}

$\overrightarrow{F_{P_{a_1}}}$ =(2, 1, 0)　　$\overrightarrow{F_{P_{a_2}}}$ =(0, 0)　　$\overrightarrow{F_{P_{a_3}}}$ =(0, 0, 0)　　$\overrightarrow{F_{P_{a_4}}}$ =(0, 1)

$\overrightarrow{F_{P_{c_1}}}$ =(2, 0, 1)　　$\overrightarrow{F_{P_{c_2}}}$ =(0, 0)　　$\overrightarrow{F_{P_{c_3}}}$ =(0, 0, 0)　　$\overrightarrow{F_{P_{c_4}}}$ =(0, 1)

即 P_a 中語詞「CD」出現 2 次、「音樂」出現 1 次、「每碗」出現 1 次；P_c 中語詞「CD」出現 2 次、「燒錄」出現 1 次、「每碗」出現 1 次，則詞頻權重向量為

$\overrightarrow{W_{P_{a_1}}}$ =(2/24, 1/24, 0)=(0.0833, 0.0417, 0)　　$\overrightarrow{W_{P_{a_2}}}$ =(0, 0)

$\overrightarrow{W_{P_{a_3}}}$ =(0, 0, 0)　　$\overrightarrow{W_{P_{a_4}}}$ =(0, 1/24)=(0, 0.0417)

$\overrightarrow{W_{P_{c_1}}}$ =(2/24, 0, 1/24) =(0.0833, 0, 0.0417)　　$\overrightarrow{W_{P_{c_2}}}$ =(0, 0)

$\overrightarrow{W_{P_{c_3}}}$ =(0, 0, 0)　　$\overrightarrow{W_{P_{c_4}}}$ =(0, 1/24) =(0, 0.0417)

依(3.8)式，P_a 的各分類詞頻權重

$$w_1 = \frac{(1*0.0833)+(0.6667*0.0417)}{\sqrt{1.4878}\sqrt{0.0087}} \approx 0.9778 \qquad w_2 = 0$$

$$w_3 = 0 \qquad w_4 = \frac{(0.0417*0.0417)}{\sqrt{1.4878}\sqrt{0.0017}} \approx 0.0342$$

再由(3.7)式，並設 S_1 至 S_4 的類別重要程度均為 1，則 P_a 的相似度為

$$SIM\ (S, P_a) = \frac{\sum_{i=1}^{m}(1 - e^{-n_i * w_i}) * u_i}{\sum_{i=1}^{m} u_i} = \frac{\sum_{i=1}^{4}(1 - e^{-n_i * w_i}) * 1}{4}$$

$$= \frac{(1 - e^{-2*0.9778}) + 0 + 0 + (1 - e^{-1*0.0342})}{4} \approx 0.2230$$

依(3.8)式，P_c 的各分類詞頻權重為

$$w_1 = \frac{(1*0.0833)+(0.2083*0.0417)}{\sqrt{1.4878}\sqrt{0.0087}} \approx 0.8095 \qquad w_2 = 0$$

$$w_3 = 0 \qquad w_4 = \frac{(0.0417*0.0417)}{\sqrt{1.4878}\sqrt{0.0017}} \approx 0.0342$$

再由(3.7)式， P_c 相似度為

$$SIM\ (S, P_c) = \frac{\sum_{i=1}^{m} (1 - e^{-n_i * w_i}) * u_i}{\sum_{i=1}^{m} u_i} = \frac{\sum_{i=1}^{4} (1 - e^{-n_i * w_i}) * 1}{4}$$

$$= \frac{(1 - e^{-2*0.8095}) + 0 + 0 + (1 - e^{-1*0.0342})}{4} = 0.2089$$

因此原(3.6)式運算結果， $SIM\ (S, P_a)$ 等於 $SIM(S, P_c)$， 但經(3.7)修正的運算式， $SIM(S, P_a) > SIM(S, P_c)$， 也就是本式已能區別出 P_a 更相似於大補帖網頁。

3.4 誤差校正

根據 2.2 節的文獻研究顯示，當相關回饋運用於檢索系統上，將能夠有效提高檢索成效，因此本研究亦將藉本小節查詢擴展及 3.5 節詞頻權重重計的觀念，進行相關回饋研究，包括誤差校正及語詞庫的詞頻權重重計，以期待本系統效能更為提昇。

3.3 節的方法，雖能檢索出多數的相關網頁，但對於具歧義性的這一類網頁，仍會被系統檢索出來，也就是合法之唱片及軟體網頁，仍會計算得到極高的相似度值。因此，我們必需從查詢的結果中，挑出這些具歧義性的不相關網頁，並找出可區辨合法網頁與非法網頁間之特徵語詞，以同樣形成一個含有多個上義詞的誤差校正語詞庫，在大補帖例子上，經上義詞方式分類後，本誤差校正語詞庫，至少包括上義詞「版權聲明」，如表 3-3

表 3-3　包含有上義詞「版權聲明」的大補帖誤差校正語詞庫

版權聲明					
語詞	詞頻	詞頻權重	語詞	詞頻	詞頻權重
版權所有	11	11/11=1.0000	查詢專線	4	4/11=0.3636
轉載必究	8	8/11=0.7273	隱私權聲明	4	4/11=0.3636
台北市	8	8/11=0.7273	客服專線	4	4/11=0.3636
侵權	6	6/11=0.5455	原廠證明	3	3/11=0.2727
違法	6	6/11=0.5455	來電洽詢	2	2/11=0.1818
出廠證明	5	5/11=0.4545	傳真	2	2/11=0.1818
統一編號	5	5/11=0.4545	洽詢電話	1	1/11=0.0909

我們將誤差校正上義詞建為另一語詞庫 $\overline{S}$， 同時此 $\overline{S}$ 語詞庫與比對網頁 P_r 利用(3.7)

$$SIM(S, P_r) = SIM\ (S, P_r) - \varphi\ SIM\ (\overline{S}, \overline{P}_r)$$

$$= \frac{\sum_{i=1}^{m} SIM\ (S_i, P_{r_i}) * u_i}{\sum_{i=1}^{m} u_i} - \varphi\ \frac{\sum_{j=1}^{h} SIM\ (\overline{S}_j, \overline{P}_{r_j}) * v_j}{\sum_{j=1}^{h} v_j}$$

$$= \frac{\sum_{i=1}^{m} (1 - e^{-n_i * w_i}) * u_i}{\sum_{i=1}^{m} u_i} - \varphi\ \frac{\sum_{j=1}^{h} (1 - e^{-\overline{n}_j * \overline{w}_j}) * v_j}{\sum_{j=1}^{h} v_j} \quad (3.9)$$

式，將可定義一個為負值的相似度，是故，(3.7)式的相似度值修正為

(3.9)式中定義一個φ，$0<\varphi\le 1$，以表示誤差校正語詞庫的重要程度，另再定義一個v_j，$0<v_j\le 1$，以表示誤差校正語詞庫內，各類別的重要程度，而h代表$\overline{S}$誤差校正語詞庫中的類別個數，$\overline{S}_j$表示第j個分類集合，$1\le j\le h$，$\overline{P}_{r_j}$表示比對網頁P_r在誤差校正語詞庫的第j個分類集合。我們以範例 4 說明上述相似度的計算方法

範例 4：

假設語詞庫S、$\overline{S}$各分類的語詞集合及其詞頻向量分別為

S ={ {CD,音樂,燒錄}媒體技術,{精選集,熱門}廣告行銷,{購買,訂購,貨款}交易方式,{原版,每碗}數量單位 }

$\overline{S}$ ={ {版權所有,轉載必究,客服專線}版權聲明}

$\overrightarrow{F_{S_1}}$=(24, 16, 5)　$\overrightarrow{F_{S_2}}$=(12, 3)　$\overrightarrow{F_{S_3}}$=(10, 13, 7)　$\overrightarrow{F_{S_4}}$=(4, 1)　$\overrightarrow{F_{\overline{S}_1}}$=(11, 8, 4)

當 $freq_{\max,S}$=24，　$freq_{\max,\overline{S}}$=11 各分類的詞頻權重向量為

$\overrightarrow{W_{S_1}}=\overrightarrow{W_{媒體技術}}$=(1.0000, 0.6667, 0.2083)　　$\overrightarrow{W_{S_2}}=\overrightarrow{W_{廣告行銷}}$= (0.5000, 0.1250)

$\overrightarrow{W_{S_3}}=\overrightarrow{W_{交易方式}}$ =(0.4167, 0.5417, 0.2917)　　$\overrightarrow{W_{S_4}}=\overrightarrow{W_{數量單位}}$ =(0.1667, 0.0417)

$\overrightarrow{W_{\overline{S}_1}}=\overrightarrow{W_{版權聲明}}$=(1.0000, 0.7273, 0.3636)

令比對網頁P_d，其各分類的語詞集合及詞頻向量為

P_d ={ {CD,音樂}媒體技術,{每碗}數量單位}

$\overline{P}_d$ ={ {轉載必究}版權聲明}

$\overrightarrow{F_{d_1}}$=(2, 1, 0)　$\overrightarrow{F_{d_2}}$=(0, 0)　$\overrightarrow{F_{d_3}}$=(0, 0)　$\overrightarrow{F_{d_4}}$=(0, 1)　$\overrightarrow{F_{\overline{P}_{d1}}}$=(0, 1, 0)

則詞頻權重向量

$\overrightarrow{W_{P_{d_1}}}$=(2/24, 1/24, 0)=(0.0833, 0.0417, 0)　　$\overrightarrow{W_{P_{d_2}}}$=(0, 0)

$\overrightarrow{W_{P_{d_3}}}$=(0, 0, 0)　$\overrightarrow{W_{P_{d_4}}}$=(0, 1/24)=(0, 0.0417)　$\overrightarrow{W_{\overline{P}_{d_1}}}$=(0, 1/11, 0)=(0, 0.0909, 0)

各分類的詞頻權重

$$w_1=\frac{(1*0.0833)+(0.6667*0.0417)}{\sqrt{2.3351}\sqrt{0.0087}}\approx 0.7833 \qquad w_2=0$$

$$w_3=0 \qquad w_4=\frac{(0.0417*0.0417)}{\sqrt{2.3351}\sqrt{0.0017}}\approx 0.0274$$

$$\overline{w}_1=\frac{(1,0.7273,0.3636)\bullet(0,0.0909,0)}{\sqrt{1^2+0.7273^2+0.3636^2}\sqrt{0.0909^2}}=\frac{(0.7273*0.0909)}{\sqrt{1.6611}\sqrt{0.0083}}\approx 0.5643$$

設誤差校正語詞庫的重要程度φ為 1，另S_1、S_2、S_3、S_4及$\overline{S}_1$的類別重要程度亦均為 1，因此比對網頁P_d的相似度為

$$SIM(S,P_d)=\frac{\sum_{i=1}^{m}(1-e^{-n_i*w_i})*u_i}{\sum_{i=1}^{m}u_i}-\varphi\frac{\sum_{j=1}^{h}(1-e^{-\overline{n}_j*\overline{w}_j})*v_j}{\sum_{j=1}^{h}v_j}$$

$$=\frac{(1-e^{-2*0.7831})+0+0+(1-e^{-1*0.0273})}{4}-\frac{(1-e^{-1*0.5643})}{1}\approx 0.2267$$

$$=\frac{(1-e^{|S_1\cap P_{d1}|*w_1})+(1-e^{|S_2\cap P_{d2}|*w_2})+(1-e^{|S_3\cap P_{d3}|*w_3})+(1-e^{|S_4\cap P_{d4}|})}{4}$$

$$-\frac{(1-e^{|\overline{S}_1\cap\overline{P}_{d1}|*\overline{w}_1})}{1}$$

3.5 詞頻權重重計

本節所使用相關回饋之詞頻權重重計技術，是將前次查詢所得到的相關網頁，從中找出符合於語詞庫的語詞，並改變語詞詞頻，如此作為下一次查詢比對時之依據。因此(3.9)式重寫如下

(3.10)式中，w_i'、$\overline{w}_j'$代表進行詞頻權重重計後，各分類所重新計總的詞頻權重，

$$SIM(S,P_r)=\frac{\sum_{i=1}^{m}(1-e^{-n_i*w_i'})*u_i}{\sum_{i=1}^{m}u_i}-\varphi\frac{\sum_{j=1}^{h}(1-e^{-\overline{n}_j*\overline{w}_j'})*v_j}{\sum_{j=1}^{h}v_j}\qquad(3.10)$$

此項權重值的大小，視各語詞詞頻與最大詞頻的變化情形而定，同時本節詞頻權重重計，可不斷重覆進行多次。範例 5 承範例 4 之語詞庫S及比對網頁P_d，說明上述方法。

範例 5：

若詞頻權重重計後語詞庫S異動為

$\overrightarrow{W_{S_1}}=\overrightarrow{W_{媒體技術}}$=(1.0000, 0.8333, 0.2083)　　$\overrightarrow{W_{S_2}}=\overrightarrow{W_{廣告行銷}}$ = (0.5000, 0.1250)

$\overrightarrow{W_{S_3}}=\overrightarrow{W_{交易方式}}$ =(0.4167, 0.5417, 0.2917)　　$\overrightarrow{W_{S_4}}=\overrightarrow{W_{數量單位}}$ =(0.1667, 0.0417)

其中假設僅有語詞「音樂」的詞頻，由 16 增為 20，所以詞頻權重重計後為 0.8333，則分類詞頻權重w_1'變為

$$w_1'=\frac{(1*0.0833)+(0.8333*0.0417)}{\sqrt{2.5851}\sqrt{0.0087}}\approx 0.8321$$

則相似度為

$$SIM(S,P_d)=\frac{\sum_{i=1}^{m}(1-e^{-n_i*w_i'})*u_i}{\sum_{i=1}^{m}u_i}-\varphi\frac{\sum_{j=1}^{h}(1-e^{-\overline{n}_j*\overline{w}_j'})*v_j}{\sum_{j=1}^{h}v_j}$$

$$=\frac{(1-e^{-2*0.8321})+0+0+(1-e^{-1*0.0273})}{4}-\frac{(1-e^{-1*0.5643})}{1}\approx -0.2218$$

4 系統績效評估

本研究主要的目的，是探討如何提高搜尋引擎之精確率及檢出率，來改善傳回太多不相關網頁的問題，因此我們建議第 3 節的相似度計算方式。以下將分為 4.1 實驗設計、4.2 實驗結果與分析，來驗證我們建議方式的可用性。

4.1 實驗設計

（1）測試集文件

2.2 節「網路上高精確率之犯罪資訊蒐尋系統」，已收集一個大小 6.8G，含有 53,240 個網頁的測試集文件，其來源係為各入口網站的分類目錄，從對應的網址下載網頁所形成。另因本研究以大補帖為例，所以必需蒐集一些大補帖的網頁樣本，作為實驗之素材，其來源係以警察實務機關破獲之大補帖案件 20 頁，以及直接在網路上蒐集到的 46 頁，共計 66 頁。

（2）實驗比較對象

「網路上高精確率之犯罪資訊蒐尋系統」之研究成果，已證明其效能較一般搜尋引擎為佳，因此本系統將與該計畫的成果作比較。

（3）績效評估測量值

採用文件相關程度的精確率與檢出率，作為績效評估測量值。並以 21 個標準檢出率（standard recall level）0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%，計算所對應精確率值後，再表示為一曲線圖(precision versus recall curve)，俾供效能比較。另外，我們利用 F 測量值求得最佳檢出率，並據此推論出一個相似度門檻值，以作為操縱本研究系統之門檻值數據。

（4）實驗步驟

前述測試集文件包括相關的大補帖網頁樣本 66 頁，及一般網頁 53,240 頁。而本研究首先從刑事警察局破獲的大補帖樣本網頁 20 個當中，擷取相關的語詞，統計詞頻並依不同之上義詞關係作分類，建成一個具階層架構的語詞庫。接著將隨機選出 23 頁的大補帖網頁樣本，投入一般網頁，並用詞頻權重相似度、分類指數相似度、分類指數權重相似度及誤差校正等方法，分別來搜尋並比較結果。最後，從上述 23 頁的大補帖樣本網頁，找出符合於語詞庫的語詞，並重計語詞詞頻，亦即本研究所稱之詞頻權重重計。同時，將另外 23 頁的大補帖樣本網頁，投入一般網頁，並用分類指數權重相似度及誤差校正來搜尋並比較結果。並於整個實驗結束後，我們將找出最好的相似度演算方法，

並藉由 F 測量值來求得最佳檢出率，且據此推論出一個相似度門檻值，表 4.1 表示整個實驗步驟。

表 4.1　實驗步驟

步驟	測試集	處理	結果
1	樣本網頁(20/66)	擷取相關語詞後 分類並統計詞頻	語詞庫
2	樣本網頁(23/66) 一般網頁(53,240)	1.詞頻權重相似度 2.分類指數相似度 3.分類指數權重相似度 4.誤差校正	描繪精確率、檢出率曲線圖
3	樣本網頁(23/66) 一般網頁(53,240)	1.詞頻權重重計	描繪精確率、檢出率曲線及趨勢線圖
4	*	F 測量值	最佳檢出率

（5）實驗參數

依據第 3 節相似度計算，我們設定下列各項變數值，作為實驗的參數來源。

變數	符號	值	說明
大補帖語詞庫	S	*	為表 3-2 之 198 個大補帖語詞
誤差校正語詞庫	$\overline{S}$	*	為表 3-3 之 14 個語詞
大補帖上義詞個數	m	4	大補帖上義詞包括「媒體技術」、「交易方式」、「廣告行銷」、「數量單位」計 4 個
誤差校正語詞庫重要程度	φ	1/4	誤差校正語詞庫的重要程度假設為 1/4=0.25
誤差校正上義詞個數	h	1	誤差校正上義詞僅有「版權聲明」
大補帖各分類重要程度	u_i	1	即各分類的重要程度均一樣
誤差校正各分類重要程度	v_i	1	即誤差校正語詞庫內之「版權聲明」類的重要程度

4.2 實驗結果與分析

依前一小節的實驗設計，我們實際進行第 3 節所說明的各種相似度比對計算，並從輸出的結果中，統計其精確率與檢出率，最後繪製成曲線圖，以比較其間的優劣。以下將對這五種相似度計算方法，從最差至最佳依序說明。

4.2.1 詞頻權重相似度實驗結果與分析

本相似度計算，係為國科會委託研究「網路上高精確率之犯罪資訊蒐尋系統」計畫 [陳志誠 1999] 的方法，我們將(3.2)式之詞頻權重相似度，透過本研究之實作系統，重新產生結果。表 4-2 係依相似度值高低列出大補帖網頁樣本排名的情形。

表 4-2　大補帖網頁樣本排名列表（依詞頻權重相似度）

排名	相似度	網頁名稱	實際網址
1	0.940214	gol.htm	http://a.f1.com.tw/11802/gol.htm
6	0.590400	booklaw.html	http://club.ntu.edu.tw/~sfa/booklaw.html
10	0.509098	msg00064.html	http://www.linux.org.tw/mail-archie/xcin/xcin.200007/msg00064.html
19	0.448452	m9040209.htm	http://www.chinatimes.com.tw/news/papers/online/tech/m9040209.htm

20	0.434992	index.htm	http://www.pros.com.tw/plaza/audio/video/index.htm
23	0.424307	info.htm	http://tinpan.fortunecity.com/suede/372/info.htm
38	0.364666	noodles.htm	http://members.nbci.com/cytunglo/noodles.htm
42	0.331677	x2mode.htm	http://users.vr9.com/ptleo/x2mode.htm
43	0.331338	page1.html	http://www.xxxdreamz.com/page1.html
52	0.319430	index.htm	http://members.tripodasia.com.tw/jazz999/index.htm
72	0.296436	new3-3.htm	http://www16.brinkster.com/vcdclub/new3-3.htm
77	0.294353	102.htm	http://go11.hypermart.net/wwwboard1/messages/102.htm
79	0.287919	newpage7.htm	http://www.dudu-sex.com/vcd/newpage7.htm
80	0.282673	ser.htm	http://www.ahpvcd.com/Vcdp/ser.htm
81	0.277379	c_reg.htm	http://www.sinica.edu.tw/~weric/c_reg.htm
82	0.274944	page05purchase.html	http://www.dj.net.tw/~gervin/cradle/page05purchase.html
92	0.254422	know.htm	http://www.xheaven.com/tw-vcd/know.htm
98	0.227164	msg00308.html	http://lists.debian.org/debian-user-9804/msg00308.html
99	0.226909	pass.html	http://www.taiwanese.com/tso/aokihp/pass.html
119	0.181612	send1.html	http://members.tripodasia.com.tw/coffee999/send1.html
122	0.175979	7766.htm	http://www.pros.com.tw/fishing/board/messages/7766.htm
140	0.153639	index.htm	http://www.belizehome.com/ilovecopy/index.htm
189	0.114848	order_.htm	http://210.208.222.150/~pchbox/taiwanxxx/order_.htm

從表 4-2 得知，當欲檢出所有 23 個樣本網頁，即檢出率為 100%時，需傳回其相似度排名之前 189 個網頁，也就是相對的精確率為 0.1217（=23/189）。表 4-3 為檢出率 0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%，及所對應的精確率值，而且平均精確率為 0.2359。另圖 4-1 的 Freq.表示其相關曲線圖。

表 4-3　詞頻權重相似度之檢出率及其精確率值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
需檢出樣本網頁數	*	2	3	4	5	6	7	9	10	11	12
實際傳回之網頁數	*	6	10	19	20	23	38	43	52	72	77
精確率	1.0000	0.3333	0.3000	0.2105	0.2500	0.2609	0.1842	0.2093	0.1923	0.1528	0.1558

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	平均
需檢出樣本網頁數	13	14	15	17	18	19	20	21	22	23	*
實際傳回之網頁數	79	80	81	92	98	99	119	122	140	189	*
精確率	0.1646	0.1750	0.1852	0.1848	0.1837	0.1919	0.1681	0.1721	0.1571	0.1217	0.2359

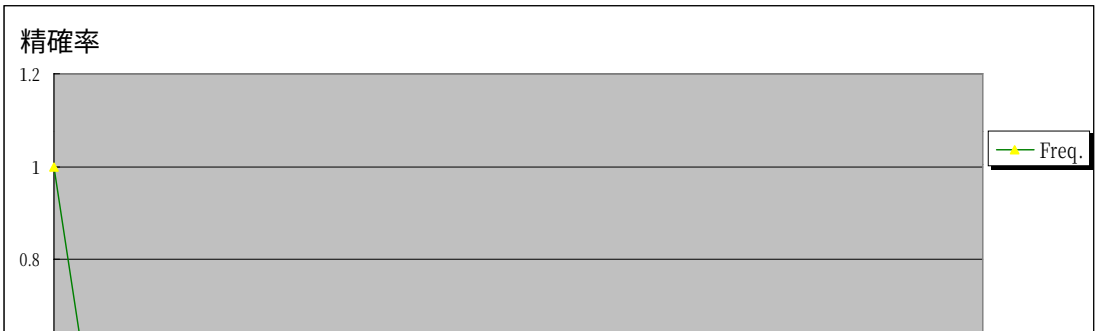


圖 4-1　詞頻權重相似度之檢出率、精確率曲線圖

4.2.2 分類指數相似度實驗結果與分析

為改善上一節精確率與檢出率，本研究建議將語詞庫依不同之上義詞關係作分類，並使用(3.6)式之分類指數相似度作計算，經由本研究之實作系統，產生表 4-4 之排名情形。

表 4-4　大補帖網頁樣本排名列表（依分類指數相似度）

排名	相似度	網頁名稱	排名	相似度	網頁名稱
1	0.994572	gol.htm	42	0.777648	index.htm
6	0.961275	index.htm	54	0.743117	x2mode.htm
7	0.948913	msg00064.html	55	0.739158	newpage7.htm
8	0.948521	c_reg.htm	66	0.728395	ser.htm
10	0.944562	index.htm	73	0.724878	info.htm
11	0.861738	page1.html	76	0.723471	pass.html
12	0.849303	know.htm	79	0.722406	msg00308.html
13	0.835784	booklaw.html	82	0.712660	102.htm
15	0.815946	new3-3.htm	88	0.677753	m9040209.htm
27	0.809797	noodles.htm	118	0.632121	7766.htm
29	0.782226	send1.html	161	0.566029	order_.htm
35	0.782196	page05purchase.html			

從表 4-4 得知，其檢出所有 23 個樣本網頁，需傳回其相似度排名之前 161 個網頁，也就是相對的精確率為 0.1429（=23/161）。表 4-5 為檢出率 0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%，及所對應的精確率值，而且平均精確率為 0.3940。另圖 4-2 的 Expo.表示相關曲線圖，其呈現分類指數相似度方法較詞頻權重相似度為佳。

表 4-5　分類指數相似度之檢出率及其精確率值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
需檢出樣本網頁數	*	2	3	4	5	6	7	9	10	11	12
實際傳回之網頁數	*	6	7	8	10	11	12	15	27	29	35
精確率	1.0000	0.3333	0.5000	0.5714	0.6250	0.5455	0.6364	0.6000	0.3704	0.3793	0.3429

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	平均
需檢出樣本網頁數	13	14	15	17	18	19	20	21	22	23	*
實際傳回之網頁數	42	54	55	73	76	79	82	88	118	161	*
精確率	0.3095	0.2593	0.2727	0.2329	0.2368	0.2468	0.2439	0.2386	0.1864	0.1429	0.3940

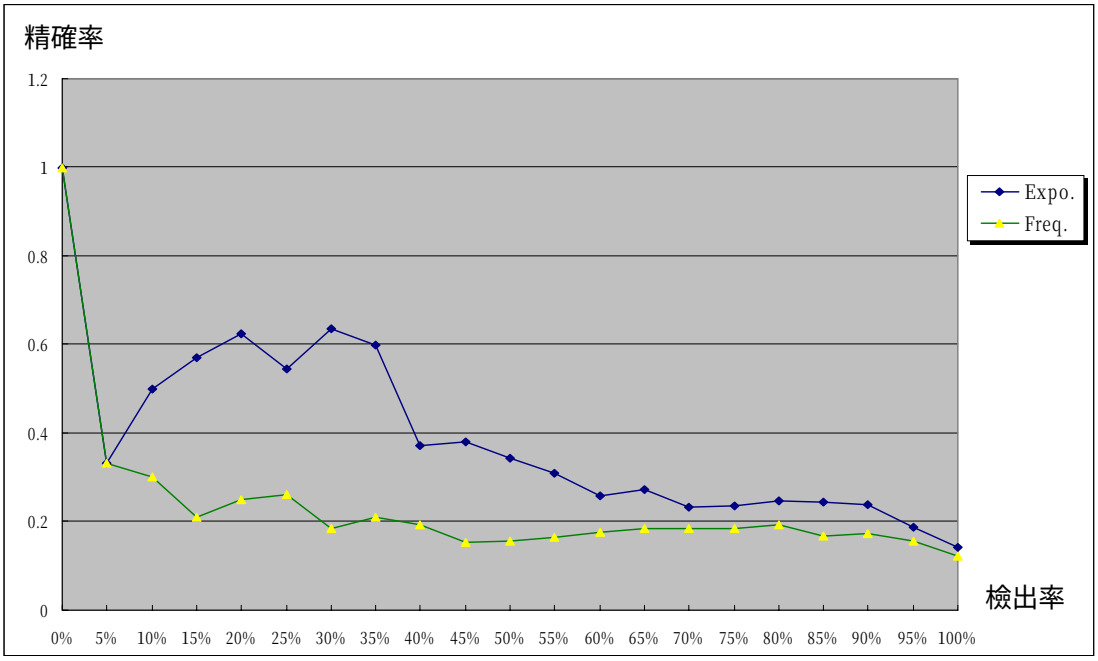


圖 4-2　分類指數相似度與詞頻權重相似度之檢出率、精確率曲線圖

4.2.3 分類指數權重相似度實驗結果與分析

分類指數相似度有一缺點，對於比對網頁僅計算不同語詞在各上義詞出現的次數，並未考量詞頻權重，因此無法區別出擁有較高語詞詞頻的網頁，應有較高之相似度值，所以(3.7)式加入語詞的詞頻權重觀念，即分類指數權重相似度，以改善這種現象，經過本研究之實作系統，產生表 4-6 之排名結果。

表 4-6　大補帖網頁樣本排名列表（依分類指數權重相似度）

排名	相似度	網頁名稱	排名	相似度	網頁名稱
2	0.983000	gol.htm	43	0.878554	7766.htm
4	0.982998	index.htm	46	0.857478	msg00308.html
5	0.982995	page1.html	51	0.835680	booklaw.html

6	0.982993	msg00064.html	54	0.803598	send1.html
10	0.982975	c_reg.htm	64	0.747080	102.htm
11	0.982904	index.htm	67	0.744623	ser.htm
13	0.982609	new3-3.htm	68	0.742165	newpage7.htm
14	0.982597	page05purchase.html	69	0.739708	m9040209.htm
16	0.979395	noodles.htm	72	0.737250	x2mode.htm
18	0.976746	know.htm	76	0.737248	info.htm
24	0.961367	pass.html	165	0.639100	order_.htm
30	0.941151	index.htm			

從表 4-6 得知, 其檢出所有 23 個樣本網頁, 需傳回其相似度排名之前 165 個網頁, 也就是相對的精確率為 0.1394 (=23/165)。表 4-7 為檢出率 0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%, 及所對應的精確率值, 而且平均精確率為 0.4482。另圖 4-3 的 Exp_W 表示相關曲線圖, 其呈現分類指數權重相似度的方法較其它方式為佳。

表 4-7　分類指數權重相似度之檢出率及其精確率值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
需檢出樣本網頁數	*	2	3	4	5	6	7	9	10	11	12
實際傳回之網頁數	*	6	7	8	10	11	12	15	27	29	35
精確率	1.0000	0.5000	0.6000	0.6667	0.7143	0.5455	0.6364	0.5625	0.5556	0.4583	0.4000

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	平均
需檢出樣本網頁數	13	14	15	17	18	19	20	21	22	23	*
實際傳回之網頁數	42	54	55	73	76	79	82	88	118	161	*
精確率	0.3023	0.3043	0.2941	0.3148	0.2687	0.2794	0.2899	0.2917	0.2895	0.1394	0.4482

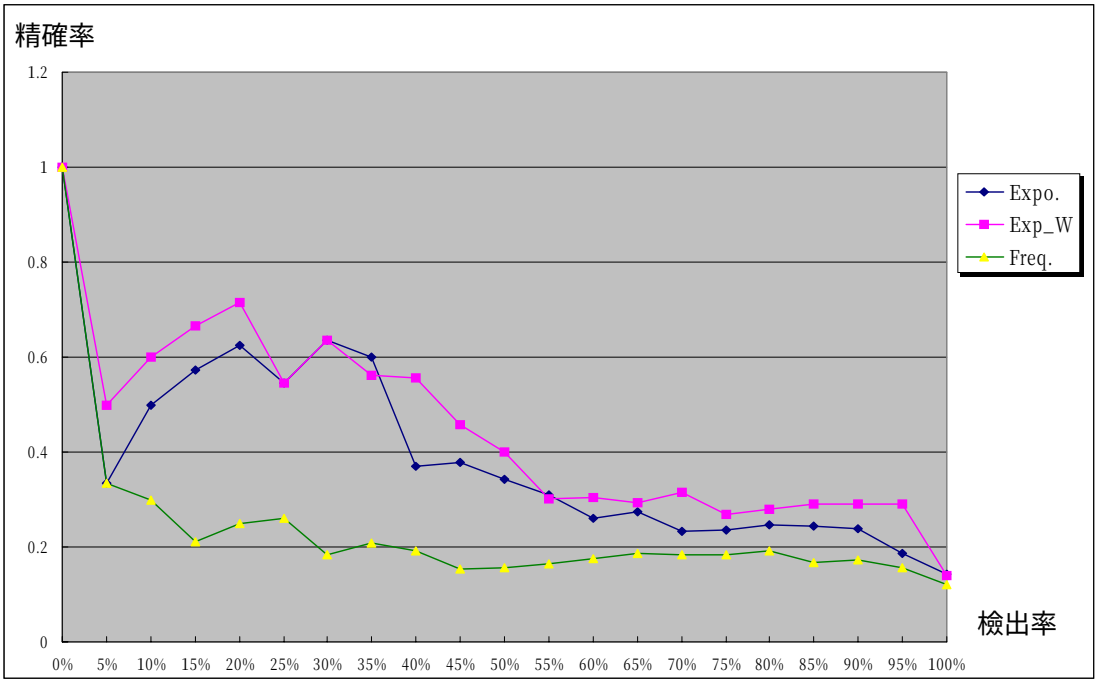


圖 4-3　分類指數權重相似度與前二個方式之檢出率、精確率曲線圖

圖 4-3 中，發現檢出率從 20% 至 5% 的區間，其精確率反而明顯下降，分析當中原因，主要是對於具歧義性的這一類網頁，即不屬於大補帖的合法正版唱片、軟體網頁，仍會計算得到極高的相似度值。

4.2.4 誤差校正實驗結果與分析

由於具歧義性的網頁，即本研究之合法正版唱片、軟體網頁，仍會計得極高的相似度值，因此建議加入誤差校正相似度計算，以有效改善這種情形。依 3.4 節說明，我們從查詢結果中，挑出歧義性網頁，找出足以區別的特徵語詞，並形成一個誤差校正語詞庫，在大補帖例子上，我們分類出「版權聲明」的誤差校正語詞庫，並以(3.8)式計算相似度，產生表 4-8 排名結果。

表 4-8　大補帖網頁樣本排名表列（依誤差校正相似度）

排名	前次排名	網頁名稱	排名	前次排名	網頁名稱
1	6	msg00064.html	22	51	booklaw.html
2	4	index.htm	25	24	pass.html
3	5	page1.html	31	46	msg00308.html
6	10	c_reg.htm	36	54	send1.html
7	2	gol.htm	39	14	page05purchase.html
8	13	new3-3.htm	42	11	index.htm
9	16	noodles.htm	44	64	102.htm
14	18	know.htm	47	67	ser.htm
18	30	index.htm	48	68	newpage7.htm
50	69	m9040209.htm	119	165	order_.htm
59	72	x2mode.htm	144	43	7766.htm

65	76	info.htm			

從表 4-8 得知，其檢出所有 23 個樣本網頁，需傳回其相似度排名之前 144 個網頁，也就是相對的精確率為 0.1597（=23/144）。表 4-9 為檢出率 0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%，及所對應的精確率值，而且平均精確率為 0.5293，另圖 4-4 的 Err_C 表示相關曲線圖，其呈現誤差校正的方法較其它方式為佳。

表 4-9　使用誤差校正計算相似度之檢出率及其精確率值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
需檢出樣本網頁數	*	2	3	4	5	6	7	9	10	11	12
實際傳回之網頁數	*	2	3	6	7	8	9	18	22	25	31
精確率	1.0000	1.0000	1.0000	0.6667	0.7143	0.7500	0.7778	0.5000	0.5556	0.4400	0.4800

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	平均
需檢出樣本網頁數	13	14	15	17	18	19	20	21	22	23	*
實際傳回之網頁數	36	39	42	47	48	50	59	65	119	144	*
精確率	0.3611	0.3889	0.3571	0.3617	0.3750	0.3800	0.3390	0.3231	0.1849	0.1597	0.5293

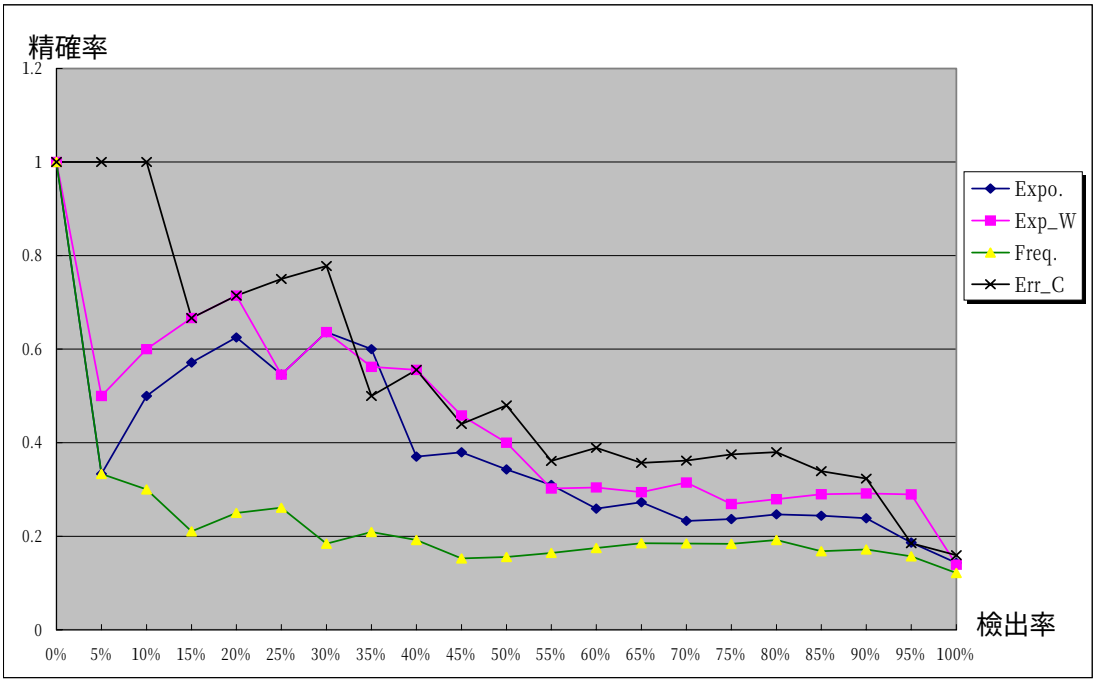


圖 4-4　使用誤差校正計算相似度之檢出率、精確率曲線圖

觀察表 4-8，其誤差校正及分類指數權重相似度的排名情形，有 18 個樣本網頁的排名顯著升高，代表具歧義性的一般網頁排名下降了。不過仍有 5 個樣本網頁的排名卻下降，這表示其內容含有「版權聲明」的語詞，因此，除應持續維護誤差校正的語詞庫外，

還可調整誤差校正語詞庫的重要程度值φ，以達到最佳化。

4.2.5 詞頻權重重計實驗結果與分析

自前四個實驗所用的第一組 23 個大補帖樣本網頁中，找出跟語詞庫相符的語詞，並重計語詞詞頻。之後，把另外第二組 23 個大補帖樣本網頁投入測試集，再使用(3.9)式計算相似度，表 4-10 為實驗後之排名結果。

表 4-10　大補帖網頁樣本排名列表（依詞頻權重重計相似度）

排名	前次排名	網頁名稱	實際網址
1	1	new3-26.htm	http://www16.brinkster.com/vcdclub/new3-26.htm
2	2	main.htm	http://www.geocities.com/fdsy6324/main.htm
3	3	music.htm	http://school-2.8m.com/MUSIC/music.htm
4	6	rule.htm	http://home.kimo.com.tw/viagratest01/rule.htm
7	7	readmel.htm	http://www.fortunecity.com/skyscraper/overdrive/224/readme.htm
8	8	call.html	http://school-2.8m.com/call.htm
9	9	new.htm	http://www.fortunecity.com/skyscraper/overdrive/224/new.htm
14	14	new3-3.htm	http://www16.brinkster.com/vcdclub/new3-3.htm
15	18	msg00633.html	http://freebsd.sinica.edu.tw/~majordom/freebsd-taiwan-questions/freebsd-taiwan-questions.199911/msg00633.html
16	22	index.html	http://school-1.8m.com/index.html
17	25	msg00634.html	http://freebsd.sinica.edu.tw/~majordom/freebsd-taiwan-questions/freebsd-taiwan-questions.199911/msg00634.html
19	31	money.htm	http://gooddbt.8m.com/money.htm
32	36	1.html	http://star88.virtualave.net/1.html
33	39	order.htm	http://members.tripod.com/~jdbt1/order.htm
36	42	index.htm	http://gooddbt.8m.com/index.htm
37	44	right.htm	http://www.fortunecity.com/skyscraper/java/537/right.html
39	47	main1.htm	http://star88.virtualave.net/main.htm
40	48	readme1.htm	http://members.tripod.com/~AV1998/readme.htm
49	50	main2.htm	http://buy.faithweb.com/main.htm
60	59	info.htm	http://chiba.cool.ne.jp/jdbt/info.htm
61	65	know.htm	http://www.geocities.com/fdsy6324/know.htm
87	119	index.htm	http://www.angelfire.com/va/mianmian/index.htm
90	144	1.htm	http://members.tripod.com/~AV1998/1.htm

從表 4-10 得知，其檢出所有 23 個樣本網頁，需傳回其相似度排名之前 90 個網頁，也就是相對的精確率為 0.2556（=23/90）。表 4-11 為檢出率 0%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、100%，及所對應的精確率值，而且平均精確率為 0.5890。另圖 4-5 的 Fee_B 表示相關曲線圖，呈現語詞詞頻重計為最佳之方法。

表 4-11　　語詞詞頻重計相似度之檢出率及其精確率值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
需檢出樣本網頁數	*	2	3	4	5	6	7	9	10	11	12
實際傳回之網頁數	*	2	3	4	7	8	9	15	16	17	19

精確率	1.0000	1.0000	1.0000	1.0000	0.7143	0.7500	0.7778	0.6000	0.6250	0.6471	0.6316

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	平均
需檢出樣本網頁數	13	14	15	17	18	19	20	21	22	23	*
實際傳回之網頁數	32	33	36	39	49	40	60	61	87	90	*
精確率	0.4063	0.4242	0.4167	0.4359	0.3673	0.3878	0.3333	0.3443	0.2529	0.2556	0.5890

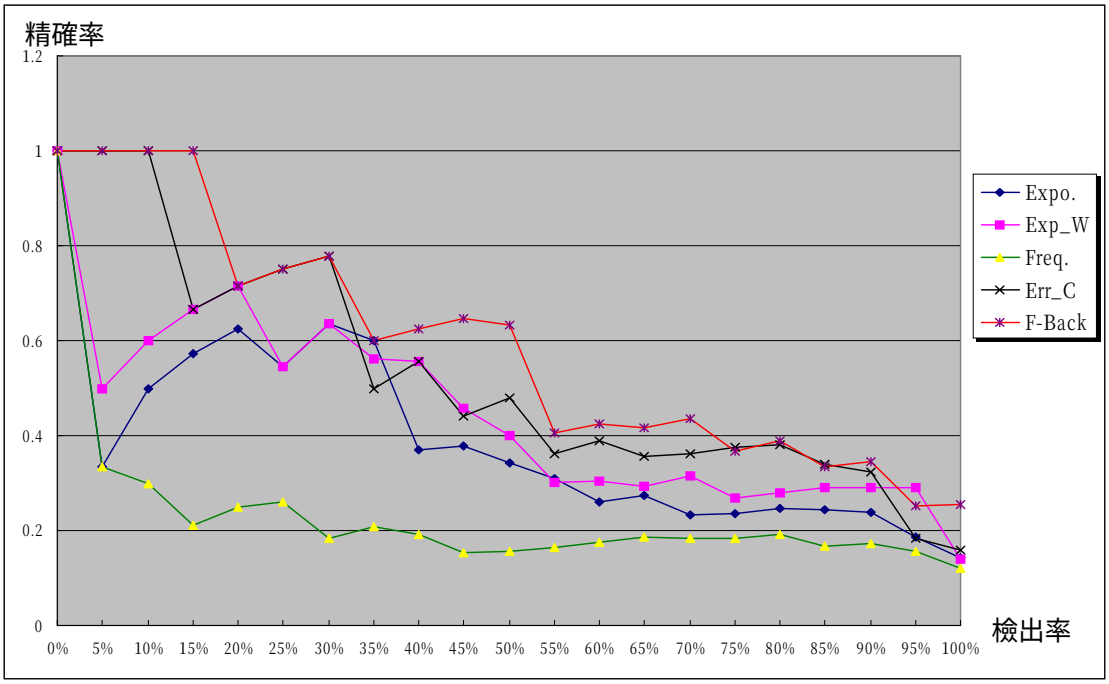


圖 4-5　語詞詞頻重計相似度之檢出率、精確率曲線圖

雖然語詞詞頻重計為最佳之相似度演算方式，但是我們必須求得最佳的檢出率，並推論出一個相似度門檻值，以作為操縱系統之參考數據。其推論的方式是先使用 2.3.3 節所說明之 F 測量值，找到最佳的檢出率，再依此最佳檢出率，計算所屬相關樣本網頁之語詞詞頻重計相似度值，最後就可將其相似度最小的定義為系統之門檻值。表 4-12 為語詞詞頻重計相似度之 F 測量值，表中顯示最佳檢出率為 50%，經查表 4-11，含有最小相似度值之樣本網頁，位於實際傳回的第 19 個網頁，亦即是表 4-10 的 http://gooddbt.8m.com/money.htm，透過語詞詞頻重計相似度計算，求得實際門檻值應為 0.5512，也就是系統比對及計算一個網頁的相似度時，一旦超過此門檻值，系統將直接認定為大補帖網頁。

表 4-12　語詞詞頻重計相似度之 F 測量值

檢出率	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
精確率	1.0000	1.0000	1.0000	1.0000	0.7143	0.7500	0.7778	0.6000	0.6250	0.6471	0.6316
F 測量值	*	0.0952	0.1818	0.2609	0.3125	0.3750	0.4330	0.4421	0.4878	0.5308	0.5581
最佳											ˇ

檢出率											

檢出率	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
精確率	0.4063	0.4242	0.4167	0.4359	0.3673	0.3878	0.3333	0.3443	0.2529	0.2556
F 測量值	0.4673	0.4970	0.5078	0.5373	0.4932	0.5223	0.4789	0.4980	0.3994	0.4071
最佳檢出率										

5 結論與未來發展方向

本研究參考相關領域之知識，以逐一推導不同的相似度計算，並用大補帖網頁為例，來評估這些相似度對系統效能的影響。根據實驗結果得知，我們建議使用「分類指數權重相似度」並配合「誤差校正」，將使搜尋引擎找到更正確的結果，當然進行「相關回饋之詞頻重計」，系統效能就更為理想了。本文未來仍有許多值得繼續研究及改善的地方，茲提出下列數端供有興趣者參考：

（1）建立各種專屬的語詞庫

本研究雖然僅建立了大補帖的上義詞語詞庫，包括媒體技術、廣告行銷、交易方式、數量單位及版權聲明等，但我們應能依此方法，陸續建立其它更多的上義詞語詞庫（包括誤差校正語詞庫）。未來，當欲進行其它的標的搜尋，即可先分析其上義特徵，再重複組合這些上義詞語詞庫，以供作另一的語詞庫來源。

（2）提高上義詞階層數

本系統僅利用二層的架構來建立上義詞語詞庫，較缺乏彈性，因此若能研究多階層的上義詞語詞庫，則可解決這問題並擴大應用，例如相似度計算結果相近的一群網頁，透過更多階層的上義詞比對，以加大相似度計算結果的差距，而能有效作出區分。

（3）增加相關回饋研究

我們雖利用相關回饋技術，進行誤差校正及詞頻權重重計，但因語詞庫會隨時空遷移，而不斷變化與翻新，因此，必須將新的相關網頁，斷詞及分類新詞，俾使語詞庫保持時間敏感性及代表性。

對於資訊檢索理論應用上，本文提供了科際整合的實例，同時本研究成果除提供檢警機關搜集網際網路上犯罪情報，解決使用人工過濾及判別內容的問題外，而且能延伸應用領域，提供個人及企業搜尋更為正確之資訊。

參考文獻

[中文部分]

[王朝煌 1998] 王朝煌，資料檢索技術及其警察文件管理應用之探討，*警學叢刊*，28卷5期，1998。

[何三本 1995] 何三本、王玲玲，*現代語義學*，1995。

[邱承迪 1998] 邱承迪，*網際網路上可疑不法資訊之自動化蒐集系統*，國立中央警察大學資訊管理研究所碩士論文，1998。

[陳志誠 1999] 陳志誠，*網路上高精確率之犯罪資訊蒐尋系統*，國科會，NSC 89-2420-H-015-002-QA。

[曾元顯 1997] 曾元顯，關鍵詞自動擷取技術與相關詞回饋，*中國圖書館學會會報*，第59期，1997。

[顏志平 2001] 顏志平、徐熊健，專用搜尋引擎之設計、建置與績效評估--以網路上犯罪情報自動搜集系統為例，*第五屆資訊管理學術暨警政資訊實務研討會*，中央警察大學，2001。

[英文部分]

[Chen 2000] Chen, P. S., "An Automatic System for Collecting Crime Information on the Internet", *Journal of Information Law and Technology*, 2000, Issue 3.

[Chen 2002] Chen, P. S. and Chu, P., "Foundamentals of Facet Analysis Method", summit to *Journal of Decision Science*, 2002.

[Farreres et al. 1998] Farreres, X., Rigau, G. and Rodriguez, H., "Using WordNet for building WordNets", *Proceedings of the COLING-ACL '98 Workshop: Usage of WordNet in Natural Language Processing Systems*, 1998, pp. 65-72.

[Frankes 1992] Frankes, W. B. and Ricardo, B. Y., *Information retrieval: data structures & algorithms*, Prentice-Hall, Inc., 1992.

[Frants et al. 1997] Frants, V. I., Shapiro J. and Voiskunskii, V. G., *Automated information retrieval theory and method*, Academic Press, California, USA, 1997.

[Jones 1997] Jones, K. S. and Willett, P., *Readings in information retrieval*, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1997, pp.167-190.

[Lin et al. 1998] Lin, S. H., Shih, C. S., Chen, M. C. and Ho, J. M., "Extracting classification knowledge of Internet documents with mining term associations: a semantic approach", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp.241–249.

[Martin 1992] Martin, J. and Odell, J. J., *Object-oriented analysis and design*, Prentice Hall, 1992.

[Miller 1990] Miller, G., "Five papers on WordNet", *Special Issue of International Journal of Lexicography* 3.

[Lesk 1995] Lesk, M., “The seven ages of information retrieval”, *conference for the 50th anniversary of as we may think*, MIT, Cambridge, Mass, 12-14 October 1995.

[Ricardo 1999] Ricardo, B. Y. and Berthier, R. N., *Modern information retrieval*, Addison-Wesley, ACM Press, New York, 1999.

[Rijsbergen 1979] C.J. van Rijsbergen, *Information retrieval*, Butterworths, 1979.

[Rocchio 1971] Rocchio, J. J., “Relevance feedback in information retrieval”, *The SMART Retrieval System: Experiments in Automatic Document Processing* (Edited by G. Salton), pp. 313-323, Prentice Hall, 1971.

[Salton 1973] Salton, G. and Yang, C. S., “On the specification of term values in automatic indexing”, *Journal of Documentation*, Vol. 29, No. 4, 1973.

[Salton 1975] Salton, G. and A. Wang, “ A vector space model for automatic indexing”, *CACM*, No. 11, Vol. 18, 1975.

[Salton 1983] Salton, G. and McGill, M. J., *Introduction to modern information retrieval*, McGraw-Hill, New York, 1983.

[Shaw et al. 1997] Shaw, W. M., Burgin, R. and Howell, P., “Performance standards and evaluations in IR test collections: Cluster-based retrieval models”, *Information Processing & Management*, 33(1):1-14, 1997.