私 立 東 海 大 學

資訊工程與科學研究所
碩 士 論 文

以聯想法則概念網路為基礎之
文章概念探索及相似性比對

指導教授：呂芳懌博士

研 究 生：顏 義 樺

中 華 民 國 九 十 二 年 七 月

Abstract

This paper proposes the representation of associative concepts to establish ACN (Associative Conceptual Network) with Fuzzy Petri Nets. In ACN, every place is considered as the single concept and certainty factor connected with transition is considered as the degree of the concept associated with others in thinking. This paper also presents the Maximum Contribution Ranking and Sigma Contribution Ranking based on the Transitive Inference and Combinational Inference of Concepts with ACN to find out the main concepts in articles.

Besides, this paper proposes Static Comparison based on that similar articles consists of the similar importance value of the same concepts and proposes Dynamic Comparison based on middle concepts to determine the degree articles meet users' requirements and similarity between articles via characteristics of associative concepts.

Keywords: Information Retrieval, Concept Rank, Document Similarity, Conceptual Network.

摘要

本論文旨在表示概念間聯想關係的方法，以模糊派屈網路建立聯想關係網路圖，其中將每個位置（Place**）視為概念，而轉置連結的確定因子大小（**Certainty Factor**）為概念間聯想的程度，並以此圖形為基礎利用遞移性推導（**Transitive Inference**）及概念合導（**Combinational Inference of Concepts**）提出最大貢獻排名和總和貢獻排名找出文章的重要概念，作為了解整篇文章的意義及大綱的依據。**

除此之外，本論文也將結合聯想關係，以兩文章的主要概念之重要性越相似，則文章也會越相似的特性提出靜態比較，和加上中間概念之相似程度提出動態比較，對照使用者提出的查詢，判斷該篇文章滿足使用者需求的程度，及比較文章間的相似程度。

關鍵詞：資訊檢索、概念排名、文件相似度、概念網路

誌謝

首先要感謝我的指導教授呂芳懌老師，總是悉心教導我許多資訊領域的專業知識，不厭其煩地回答我愚拙的問題，讓我在資訊領域中得以漸漸成長。也謝謝老師總是勉勵我以不同角度思考問題，接納並指導我突如其來的想法和創意，經常教導我做研究的方法，並時而提醒容易忽略的細節，才讓我從中體會到研究的樂趣，因此對研究有了濃厚的興趣。老師也經常逐字逐句地批改我的文章，願意花費大量時間教導我寫作的技巧，包括文章的結構安排和字句通暢編寫等方法，讓我在文章的寫作有了更多的長進。除此之外，他的做事踏實和堅持也是我未來不斷學習的方向。也感謝林宜隆老師、賴森堂老師、黃其泮老師和陳澤雄老師口試時的指導，給了這篇論文相當寶貴的意見，改進了論文中未盡周詳的缺失。

謝謝實驗室陪伴我在研究所日子打拼的這些夥伴，政瑋、俊維、嘉鴻、真真、清健、耀中、國煒、仁傑、津華、坤義和國榮，雖然研究生活有許多的壓力，但是因為他們的鼓勵並且分享經驗，讓我隨時備有活力為自己的目標衝刺。謝謝教育學程的陳世佳老師和趙長寧老師，也謝謝恒旭、昌諭、欣怡、嘉玲和鈺婷，因為你們，讓我修教育學程時有了特別精采的日子。

謝謝惠虹陪伴並勉勵我寫論文的日子，甚至為了陪我完成論文，晚上就在實驗室裡打地舖就寢，也體諒我沒有太多時間可以在你身旁。當然，謝謝草莓和球球兩隻可愛的貓帶給我的快樂。

謝謝東海靈糧堂的牧師、師母和文淵小組中的每個成員，文淵大哥，素如姐，文淑姐，耀星大哥，瑞琴姐，天祥大哥，淑雅姐，明輝大哥、頌馨姐，還有一位大姐是頌馨姐的鄰居，雖然忘記您的名字，卻是感謝您們經常為我禱告，讓我經常接受來自上帝的祝福，也謝謝您們總是照顧我這個外地來唸書的學生，讓我雖在外地卻也備感溫馨。

謝謝家人對我的支持，讓我可以全心全力衝刺自己的論文，經常關心我這個讓人擔心的小孩，雖然沒有經常在你們身邊，但還是想跟您們說我愛您們。

這本論文如果有那麼一點點貢獻，我願將這榮耀歸給　上帝，因為祂，我才能進入葡萄園裡，享受祂賜下喜樂而豐富的生活。

「甚願你賜福與我，擴張我的境界，常與我同在，保佑我不遭患難，不受艱苦。」

目錄

圖目錄

表目錄

第1章　緒論

1.1 文章檢索

資訊檢索技術係利用電腦自動地從大量的文件、聲音或影像等資料中，找出符合使用者需求的資訊，並將這些資訊展示給使用者。其目的是希望透過電腦自動化的特性，使我們可以較以往傳統人工尋找資料的方法更快速地取得自己需要的資料，除了可以大幅縮短搜尋的過程，減少人力和時間成本，也可以整合來自不同地方的資料，取得最新的訊息，避免人工建立資料的不一致，提高了資料的有效性及實用性。因此資訊檢索技術已經被大量地應用在網路上的搜尋引擎[1,2,11-14]、圖書館及各種知識管理系統等地方，成為目前一般人搜尋資料最重要的工具之一。

就檢索的資料類型而言，資訊檢索分為多媒體檢索和文字檢索兩種。前者包含了影像、聲音和動畫，而後者則是以文字為主的資料，例如，論文、小說或者網頁的本文內容等皆屬之。現今大部份文件仍以文字為主，多媒體為輔，來闡述內容，但由於多媒體的資料牽涉過多的變數，例如，下雨天由於視線不清，拍攝的照片也就容易模糊；而拍攝的角度不同，有時也很難從相片辨識被拍攝者的身分；或者錄音時，如果周圍環境的雜音過多，所錄製的聲音也不易辨認，諸如此類原因，往往使得電腦無法有效地辨識這些多媒體的資料，更遑論由電腦來解釋這些多媒體資料所代表的意義，例如，電腦即使可以在影片檢索出「機關槍」的資訊，但仍不知道這把機關槍在整個影片中代表的意義，自然也無法了解整部電影的情節。而由於文字資料較多媒體資料不易有資料模糊的情況，也較容易清楚資料欲表達的主題和想法，因此想要了解整個文件的意義主要還是從文字部分著手。

基於以上的原因，文字檢索成了現在資訊檢索中不可或缺的技術，其中以文

章檢索最為重要。文章檢索最終的目的是希望能夠了解整篇文章的意義，計算出文章內容符合檢索系統的使用者之需求程度，並將計算的結果送回使用者；而或者比較文章之間的相似程度，方便使用者僅需要利用少數文章就可以找出更多的相關文件，因此，文章檢索技術的良窳往往決定該檢索系統的成敗，也是許多學者研究的方向。

傳統的資訊檢索僅利用關鍵字的比對文章的內容來搜尋文章和計算文章間的相似程度，往往無法滿足使用者的需求，例如，使用者在資訊檢索系統輸入「籃球」，則和籃球比賽有關，和賣籃球的文件都被蒐集起來，兩者在意義上卻有相當大的差異，也由於有些關鍵字比較熱門，出現在許多文件上，因此有時蒐集回來的文件常是成千上百，令使用者不知如何挑選其中最重要且是他們最想要的。多年來，資訊檢索相關系統的業者也苦苦思考這樣的難題，但卻往往無法有效地解決。由此看來，資訊檢索系統必須依照查詢需求，有效地找出相關資料外，並能夠正確地排名這些資料，俾讓使用者能以最快的速度取得其中最重要者。

1.2 研究動機

如何正確地回應使用者的需求成為目前資訊檢索最重要的課題[3-4]，有些學者建議從文章的意義判斷是否符合使用者的需求，[15]認為對於同一關鍵字，不同的背景就有不同的意義和重要性，例如，就運動類別而言，「麥可喬丹」會比「貝多芬」來得重要，反之則否。利用事先建立每個關鍵字在各種背景的重要性，而後電腦再根據使用者輸入的查詢字詞和指定的背景，可以算出文章中每個關鍵字的重要性並加總之，就可以知道文章內容符合使用者需求的程度。[16]則建議利用字詞的同義詞、反義詞或屬性等語意關係建立文章語意的結構，電腦可以根據使用者查詢字詞中，與文章中事先建立的結構比較，找出相關的文章或判斷文章間的相似程度。

這些方法雖然已經改善了檢索的正確率，但仍然無法滿足使用者的查詢，其

中之一的原因是因為這些技術並沒有依照一般人的思考模式提出一個檢索的方法。人們通常會以自己的思考方式詮釋一篇文章的意義，有時候會根據文章中概念間的同義、反義或屬性等關係思索作者想要表達的含意，有時候則會利用其他難以定義出的關係來，例如，某個明星是產品的代言人，而廣告中又經常出現該明星與產品的訊息，則當我們在文章中看到該明星的姓名，自然也會加入該產品的概念推斷文章的意義。我們無法以一般的語意關係來表示「明星的姓名」與「產品的名稱」，自然也無法表示成電腦看得懂的資料格式，因此也無法利用它們之間的關係檢索文章。以上只是一個簡單的例子，其實生活中還有更多的語意關係不容易表示出來，我們認為與其定義這些複雜的語意關係，不如定義出描述人類思考模式的聯想關係，依據這些聯想關係，使設計出來的檢索方法更符合人類的思考方式，也可以增加檢索的正確率。

1.3 研究目的

本研究主要目的在於能有效表示與人類思考模式相似的聯想關係，並利用聯想關係找出文章中的概念，了解整篇文章的意義，並比較使用者提出的查詢，以判斷該篇文章是否符合使用者需求，除此之外，利用了解每一篇文章的意義，我們也可以比較文章之間的相似程度。因此，本研究將提出聯想關係的表示方法，並以此建構篩選文章中主要概念之流程，並利用選出每篇文章的主要概念及聯想關係的特性提出比較兩篇文章相似程度的方法。

1.4 研究步驟

為達到前面所述的研究目標，研究發展進行的步驟如圖 1.1 所示。本研究先以現有的文獻資料及現行系統瞭解各種文章檢索的方法，找出各種方法表示概念間關係的方式，探討其優缺點，並從中選出適合聯想關係的表示法。在建立聯想

關係的表示法後，也將討論聯想關係可以應用在資訊檢索的特性和可能遭遇的問題，利用這些特性設計一套可以找出文章主要概念和比較文章相似程度的方法，並針對這些問題提出因應的對策。根據這些設計的方法和因應的對策，進行系統的分析、設計與實作，並在取得實驗結果之後修正這些方法，而後再繼續實驗，如此不斷重複直到完善為止。

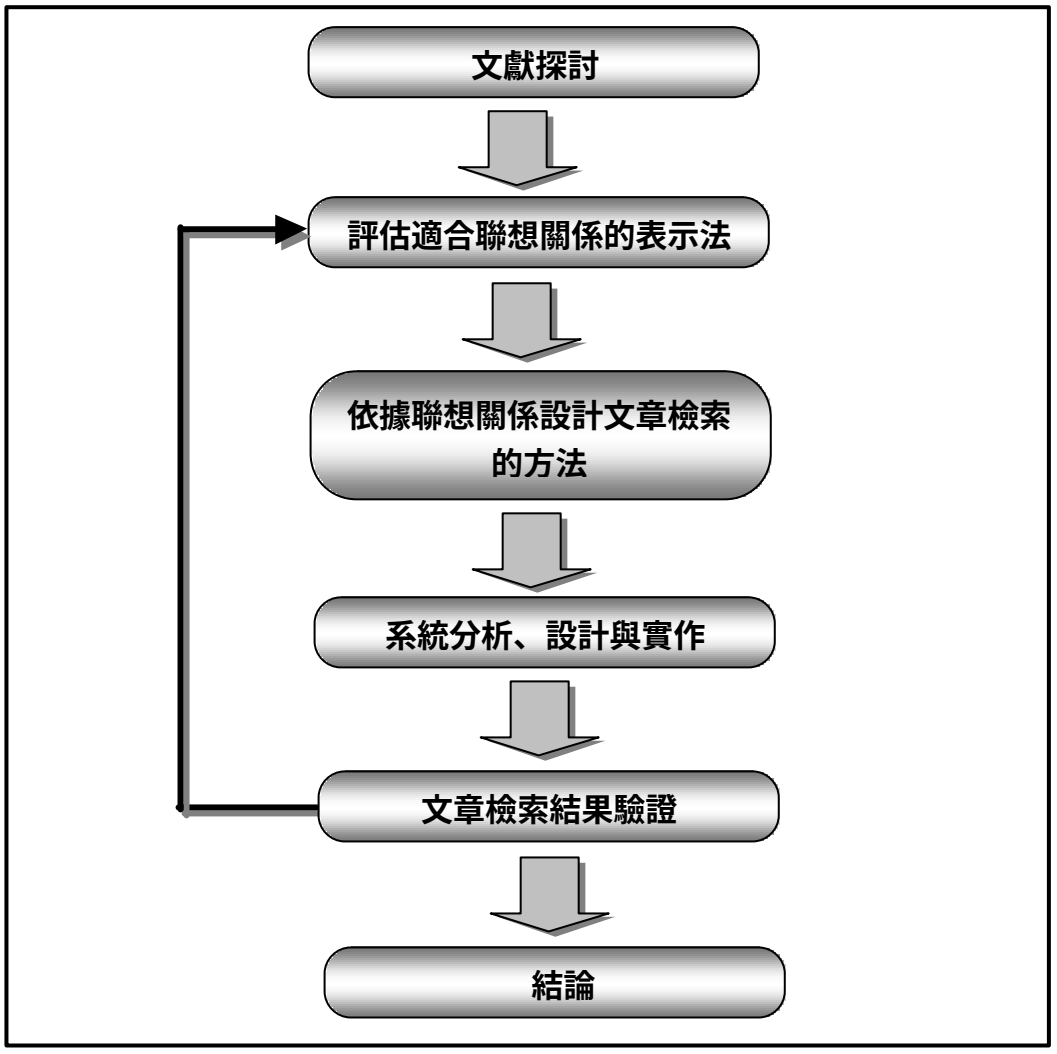


圖 1.1 本研究進行的方法及步驟[資料來源：本研究整理]

1.5 章節概論

本論文共分為七章，除第一章外其餘各章節概要如下所述：

第二章　文獻探討　介紹和本研究相關的資訊檢索技術，除了探討各種技術的優缺點，也將研究文獻中所提出的概念關係及其表示法。

第三章　聯想概念網路　本章將敘述概念間的聯想關係和聯想概念的意義，並利用聯想法則（Associative Rule）描述聯想關係。除此之外，本章也將介紹能表示聯想法則的方法，並以此建立可以包含眾多聯想關係的聯想概念網路。

第四章　概念排名　找出文章中的主要概念往往可以了解整篇文章所要傳達的意義，並可以之分類文章，本章將利用聯想概念網路提出：最大貢獻排名和總和貢獻排名兩個方法來找出文章中的主要概念。

第五章　文章相似比對　本章中將分別提出以最大貢獻排名和總和貢獻排名所找出之主要概念為基礎的文章相似程度比對的方法，此外，也將提出另一個利用概念間的聯想過程比較文章相似程度的方法。

第六章　實驗結果與分析　實作以上各章提出的方法，並分析及討論實驗的結果。

第七章　結論及未來研究方向　說明本論文的研究結論與後續研究發展的方向。

第2章　文獻探討

資訊檢索的發展迄今已經有四、五十年的歷史[17]，為幫助使用者從大量的非結構化或半結構化的資料當中，快速的取得使用者所需要的資料，人們利用電腦的協助儲存及搜尋大量資料。而這些電腦的協助技術已經被大量地應用在網路上的搜尋引擎、網路巡邏系統、圖書館及各種知識管理系統等地方，成為目前一般人搜尋資料最重要的工具之一。由於傳統的資料庫系統對於結構化的資料已經有相不錯的資訊處理能力，因此，目前資訊檢索的對象多半以結構化或非結構化的資料為主，例如，電子文件或網頁等。以往由於電腦硬體處理能力和成本的限制，有能力使用電腦的人並不多，電子化的文件往往是受到限制，例如，論文、書籍或著作等才會進行數位化儲存。然而，這類電子資料通常會包含某些領域的專有名詞，因此只要能夠從文章擷取這些字詞，就能有效地判斷文章的相似性。但近年來，由於科技的發達和網路的快速興起，一般人可以更方便地製作數位內容，因而產生了大量的電子資訊，如果僅以關鍵字而想從文章中檢索所需的資訊，往往會回應成千上萬的文章，使用者因而不知道如何從中選擇。因此為了有效地滿足使用者的需求，資訊檢索系統必須能夠正確地了解文章的意義才能讓使用者更精確地獲得所需要的文章。

2.1 傳統資訊檢索

傳統的資訊檢索技術包含關鍵字擷取、關鍵字索引、全文檢索、文件自動分類及文件自動摘要等[17]，主要目的在滿足專業的檢索單位像是圖書館的需求。其中可以包含布林模式（Boolean Model）、向量空間模式（Vector Space Model）和機率模式（Probabilistic Model），以下將逐一介紹。

2.1.1 布林模式

布林模式係根據集合理論和布林代數發展出來的簡易資訊檢索模式。當文章 D 中包含使用者的查詢 Q 時，則 D 和 Q 的相似性就為 1，若不包含則為 0；這種方法的優點是清楚，簡單；但缺點是如果使用者的查詢如果是一般人經常使用或不常使用的字詞，則系統會回應太多或太少的文件。因此，有些學者則提出擴充布林模型（Extended Boolean Model）建議加權每個關鍵字以提昇檢索的效果。

2.1.2 向量空間模式

向量空間模式是將資料庫中之文件與查詢文章的相似性量化，因此可以更正確地描述出兩文件間的相似性。

假設文件 d_u 及 d_v 所包含的所有關鍵詞集合為 $k=\{k_1, k_2, \Lambda\ , k_t\}$。則 d_u 及 d_v 相似度 $sim(\vec{d}_u, \vec{d}_v)$：

$$sim(d_u, d_v) = \frac{\vec{d}_u \bullet \vec{d}_v}{|\vec{d}_u| \times |\vec{d}_v|}$$

$$= \frac{\sum_{k=1}^{t} w_{k,u} \times w_{k,v}}{\sqrt{\sum_{k=1}^{t} w_{k,u}^2} \times \sqrt{\sum_{k=1}^{t} w_{k,v}^2}}$$

其中，

$$w_{k,u} = f_{k,u} \times \log \frac{N}{n_k}$$

而

$$f_{k,u} = \frac{\text{文件}u\text{內k出現次數}}{\text{所有文件中k出現次數的最大值}}$$

$\log \frac{N}{n_k}$ 則為文件逆頻率值（inverse document frequency），

N：文件集合中的文件總數量。

n_k ：文件集合中出現 k 的文件數量。

其中 $\vec{d}_u$ 及 $\vec{d}_v$ 分別代表文件針對 d_u 及 d_v 關鍵詞集合 K 的和向量，$w_{k,u}$ 與 $w_{k,v}$ 則分別代表關鍵詞 k 於 d_u 及 d_v 中出現的頻率，經過與出現最大頻率正規化後，乘上文件逆頻率值（此參數代表該關鍵詞於此類文件中的代表性）。由於向量空間模式易於了解，且正確率比現在許多的檢索方法高出很多，因此為現行資訊檢索系統最常採用的檢索技術。文章內經常會有隱含的概念，在了解文章上扮演重要的角色，而這些隱含的概念往往必須由其他的關鍵字才可以得到，但由於向量空間模式視任兩個關鍵字之間的關係彼此獨立，因此我們並無法利用向量空間模式找到文章內的隱含概念，也無法利用它判斷文章的相似程度，降低了檢索的正確性。

2.1.3 機率模式

機率模式最早由 S.E.Robert 所提出，他假設文章中關鍵字之權重為 0 或 1，而關鍵字之間是互相獨立的，其相似程度的計算方式為 $\sum d_i \times q_i$ ，其中 q_i 等於 $\frac{\log(PR_i \times (1-PNR_i))}{PNR_i(1-PR_i)}$ ，PR_i 為關鍵字 i 出現在相似性文章中之機率，而 PNR_i 為關鍵字 i 出現在非相關文章中之機率。採用此法最大的困難是不易決定 i 出現在相似性文章中或出現在非相關文章中之機率，且在評估相似性的方面，必須假設每個詞彙出現的機率是相等的，且之間沒有關聯性，才不會造成相似性重複計算的問題。

2.2 概念網路

以關鍵字為主的資訊檢索技術僅以出現在文章的關鍵字為基礎，並沒有考慮隱藏在文章內的概念，也無法真正了解文章中蘊含的意義，因此往往不能滿足使用者的需求。有鑑於此，有些學者就試圖藉由組合文章中的某些概念來理解文章

內的意義。概念就是人類的想法，可以是一個字或詞，或者是一個句子，甚或是一個文章的段落，整篇文章也可以代表某些概念。概念包含某種意義，因此，不同的字詞或字句都可能代表相同的概念。例如，「Wordnet」[18]中所有詞彙皆以「詞義概念」分群，具有相同意義的詞彙會被分在同一個集合中，稱為「Synset」，即同義詞集合（synonym set），例如，「fast」、「quick」、「alacritous」、「prompt」和「swift」就屬同一個 Synset，此外，也定義了反義詞（Antonymy）等概念間的關係，而由概念和概念間的關係所形成的網路就稱為概念網路。

認知心理學認為一篇文章中其實就是由許多重要的概念串聯而成，而文件中所描述的概念，其實是擁有概念的所有字詞組合的結果，其中將文件中的概念串聯起來，即可拼湊出文章所敘述的大部分意義，這種概念串聯稱為「詞義鏈」（Lexical Chain）。詞義鏈即是一篇文章中相同意義的字詞所組合成的集合，每個詞義鏈代表這篇文件所要描述的一個概念。

Wordnet 僅提供概念之間是否具有關聯，而為了能更正確描述概念之間的關係，因此[19]提出模糊化詞義概念網路（Fuzzy Term Concept Network，FTCN）表示詞彙之間的關係及其強度，如圖 2.1 所示，N_a 和 N_b 以有向邊連結，代表該關係的種類與關係強度，關係的種類可以由使用者自行訂定。

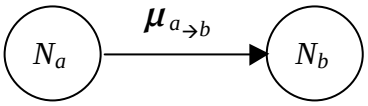


圖 2.1 從 N_a 連結到 N_b 的詞義概念網路

雖然我們可以使用模糊化詞義概念網路表示 *Wordnet* 的眾多關係，但人們閱讀文章的時候，並不一定會使用這些關係來找到文章中隱含的概念，例如，雖然「輪胎」是「車子」的部分關係，但人們看到「車子」聯想到「輪胎」的機會較小，聯想「廠牌」的機會較大。因此，如果能夠找出人們思考時經常使用的概念關係，才能減少不需要定義的關係，降低相似性計算的複雜度，卻可以提昇檢索的正確

率。

第3章　聯想概念網路

3.1 聯想概念、聯想關係和聯想法則

人們經常會利用概念間的聯想關係思考他們所面對的事物，例如，一位音樂家聽到少女的祈禱的樂曲通常會想到貝多芬，其中「少女的祈禱」和「貝多芬」皆為一個概念，因此對於音樂家而言，兩者間存在著聯想關係。事實上，聯想會影響人們言語表達的內容及接收概念的過程，與人談話、分析及推理、編輯或閱讀文章等皆如此。尤其是寫作文章，常為了詞句的優美，而利用隱喻的修辭法來撰寫，文章內不一定會直接出現作者想要表達的概念的陳述字句，而是利用與主題相關的概念迂迴地敘述。當讀者閱讀該文章時，則會根據自己的背景知識、生活的經驗及志趣等聯想、推論甚至猜測所有可能的概念，並從中歸納出自己認為最合適者作為主要的概念，以為該文章的意義。然而每個人聯想的方式以及對相關概念的深入程度會隨著背景知識、生活習性及風俗習慣等的不同而相異，因此對於同一個主題，兩個作家就能寫出不同的內容；對於同一篇文章，不同的讀者就有不同的詮釋。

個人思考事物的方法往往也會決定理解事物的程度，因而影響學習新知識的成效，其中思考時會仰賴各種概念間的關係，並利用舊知識循序漸近推導出新知識，爾後，再利用這些舊有及推導出來的知識繼續推導出更多更新的知識，如此不斷地重複亦不斷地建立新知識，而成就了個人的學問。由此可知，個人學習的首要工作就是必須清楚各種概念間的關係，才可以快速正確地推導出新的知識。據此，許多教學者就利用各種概念的關係來教導學生，其中最常採用的便是聯想關係。

聯想關係經常被用於兒童的學習，例如，聯想識字。聯想識字是教導兒童看到一個”刺激字”，而可以聯想到其他”反應字”的教學活動，讓兒童能夠藉由字詞

的聯想迅速記憶新的字詞。這種聯想是一種短暫的記憶，但經由不斷練習及回想新字詞，就可以變成長期記憶，而永久儲存在大腦。日後兒童在文章內看到某些字詞，自然地會從文章中找出與該字詞有聯想關係的其他字詞，決定該字詞在文章中的意義。

聯想識字包含「語意關係」、「音韻關係」、「句子結構關係」及「個別化的聯想」等方式。其中，

一、語意關係包含了

1. 「香蕉」和「水果」的繼承關係。
2. 「喜歡」和「喜好」的同義關係。
3. 「喜歡」和「討厭」的反義關係。
4. 「檸檬」和「酸」的屬性關係。
5. 「腿」和「身體」的部分關係等語意關係。

二、音韻關係則包含了

1. 「於」和「原」等同聲母關係。
2. 「青」、「經」和「晶」等同韻母關係。
3. 「中」、「鍾」和「終」則屬於同音關係。

三、「肚」-「餓」等詞形關係則屬於句子結構關係。

四、個別化的聯想則指個人對於某些字有特別的經驗、感受或主觀意識而聯想到的字，例如，看到「康乃馨」就很容易聯想到「母親節」。

智慧型系統最終的目標為正確地模擬人類思考方式及學習的行為能力，期望與人類有相同甚至更高的智力、智慧及知識。我們可以藉由觀察人類的學習過程，找出大腦儲存和處理知識的方式，並將之實作在系統中來達成目標，例如，類神經網路理論便是模擬人類大腦神經元儲存和傳遞知識的方式，在應用方面已發展出許多智慧型系統，並廣泛地應用在不同的領域。

聯想識字對於幫助兒童認識新字詞及學習使用新字詞的成效顯而易見，而其過程所仰賴的又是概念間的聯想關係。因此，若能有效表示聯想關係，模擬出聯

想識字的過程，並設法應用在資訊檢索系統，則必能有效提昇檢索效果。部分資訊檢索系統會事先定義兩概念間的關係[16]，例如，繼承關係，再利用這些關係提出檢索文章的演算法，而這些演算法往往只能針對某些特定關係，如果想要考慮利用其他關係檢索文章，就必須再發展不同的演算法。

由於聯想關係為定義兩概念在大腦中思考思維存在的關係，只要可以由概念 A 聯想到 B 或由 B 聯想到 A，A 和 B 就具聯想關係，而人為定義的關係往往也是人們為了有效表示兩概念之間的關係經由不斷思量而得到，表示兩概念間也具有聯想關係。因此聯想關係包含繼承、屬性、同義、反義和部份等人為定義的關係，並納入個人主觀意識等其他人為定義的關係。

此外，在資訊檢索系統中，聯想關係的實用性也較高於其他定義出來的關係，例如，部分關係，原因如下。如前述，概念間可以存在許多不同種類的關係，但我們在文章中看到某些概念後，通常所聯想的是最直接且相關的概念，例如是 X。也就是說個人會從眾多概念中找出最有關係者，過濾掉不相關 Y 者，被過濾掉的概念即使有繼承、屬性、同義、反義和部份等關係，也比較不會納入思考，除非文章內容後續部分無法從 X 繼續推論而得到合理的解釋或答案，人們才會再從 Y 找出次相關者。若仍無法得到滿意的答案，則又從 Y 中找出其他者，如此重複，直到滿意為止。例如，雖然「車子」和「輪胎」有部分的關係，但是我們看到「車子」時通常會想到「廠牌」，而不是「輪胎」，因此「車子」和「輪胎」的聯想關係比和「廠牌」者薄弱，因而「車子」的聯想以「廠牌」為優先。A 和 B 若存在聯想關係，則表示看到 A 時，確實可以比較容易聯想到 B。當然，聯想關係也有間接者，例如，A 可以聯想到 B，B 又可以聯想到 C，則 A 可以間接聯想到 C。結論是，若將之應用在資訊檢索系統上，其語意聯想必能有效表達文章之意義，而大幅提昇檢索的正確性。

其實在自然界中有許多事情都是根據規則在運作，聯想也是一種「當看到某種概念而產生其他想法」的規則，因此，可以利用 if…then…的條件敘述句來表示，稱為聯想法則（Associative Rule），例如：某個明星是產品的代言人，而廣

告中又經常出現該明星與產品的訊息，久而久之，我們會將兩者畫上等號。當我們在文章中看到該明星的姓名，也會很自然地將該產品的概念導入而繼續做後續文章內容意義的推斷，因此，可以將兩者表示成 if「明星姓名」then「產品名稱」。If 後的概念是觸發條件， then 之後的是結果概念。另一個例子是，當看到「素人自拍」就會有「色情」的想法，可以表示成 if「素人自拍」then「色情」，其意義是由「素人自拍」的概念在我們的大腦中觸發了「色情」的結果概念。而聯想有程度上的大小，例如：當看到「路由器」時會想到「網路」和「電腦」，很明顯的，聯想到「網路」的程度必大於聯想到「電腦」者，因為路由器與網路關係比較密切。聯想關係具有方向性，但不一定可逆，例如，古典樂迷聽到「少女的祈禱」會想到「貝多芬」，但聽到「貝多芬」不一定會想到「少女的祈禱」。但是有時候觸發概念或結果概念有可能會是兩個或兩個以上，例如，當聽到少女的祈禱，有人會想到貝多芬，有人則想到垃圾車，則可以表示成 if「少女的祈禱」then（「貝多芬」or「垃圾車」），而古代羅馬認為紅色代表高貴，以 if（「羅馬」and「紅色」）then「高貴」表示，則是兩個概念產生另一個概念的例子，色彩學家認為男性對於紅色的聯想為熱情、血、或罪人等，表示成 if（「男人」and「紅色」）then（「熱情」or「血」or「罪人」）則是更複雜的例子。

聯想法則描述聯想關係，唯有正確地表示聯想法則，才能將聯想關係應用於資訊檢索系統上。由於知識表示法（Knowledge Representation）可以扮演個人表達自己，也可以當作與他人溝通的方法，且可以轉化為計算的媒介，適合運用在資訊領域中。接下來我們將探討知識表示法的意義，及聯想法則的知識表示法。

3.2 知識表示法

認知心理學（Cognitive Psychology）的核心觀念，是以訊息處理理論（information processing model）為其理論架構，認知心理學家認為表示法是一種以概念代替實物的歷程，會影響個體對於知識的處理方法。而認知心理學家

Soloso[1]提出有關於人類語意組織的五種知識表徵方式，其中之一「網路模式」（network model）最具應用價值。又，聯想法則表示法可以表示出「if…then」的法則，能正確描述聯想關係。另外，模糊推論法則（Fuzzy Production Rule）可以用來進行模糊推論，使系統達成智慧思考的目標；派屈網路（Petri Net）是一種圖形和數學的模型工具，經常應用於製作自動化或動態的系統，而模糊派屈網路（*FPN*, Fuzzy Petri Net）則是以派屈網路為基礎，經常用於表示模糊推論的規則，亦可以用來表示法則知識(rule-based knowledge)並自動推論，以下論述之。

3.2.1 網路模式

該理論模式認為在人類的記憶系統中，知識是由儲存於記憶中的各個獨立單位所連結而成的網路，知識結構乃是由許多代表基本概念的節點，以及節點和節點之間的連結組成。網路中每個節點代表一個概念，所謂「概念」（concept），是個人知識體系中的基本單位，概念的本質中充滿著不確定性、模糊性，在概念的學習過程中，不僅是要配合先備知識的運用，以瞭解其基本的定義和屬性，更必須依靠對情境及相關線索的掌握、發展、調適、類化概念知識與程序性知識，方能獲得有效的概念知識的學習。因此，概念學習在教育上的含意即是，希望藉由概念具體意義化，幫助學生瞭解概念，以促進學生能以抽象的方式使用具體概念。我們能夠記憶每一個字的原因，乃是因為它與一個複雜的「關係網路」（network of relationships）連結在一起。也就是說，「香蕉」和「水果」都是獨立的概念，只是在關係網路中以繼承關係互相連結，以表示概念之間的關係。因此，以聯想關係為主的語意網路中，每個概念彼此獨立，且以聯想關係彼此連結。

3.2.2 聯想法則

聯想法則係人們聯想事物時所依循的規則，這些規則存在於人類的大腦中，

聯想關係取決於聯想者本身的生活經驗、背景、知識和個人興趣，往往是大量且複雜。 我們以 if…then…條件敘述句來表示聯想法則，其中 If 後的概念是觸發條件， then 之後的是結果概念，其中觸發條件和結果概念皆可以利用 and 和 or 組合之，而每個聯想法則皆有聯想程度λ，λ∈[0,1]。

3.2.3 模糊推論法則

模糊推論法則用以表示兩個命題（proposition）間的模糊關係。令 R 為模糊推論法則的集合，$R=\{R_1,R_2,\ldots R_n\}$，其中第 i 個模糊推論法則定義如下，$1 \leqq i \leqq n$：

$$R_i：IF\ d_j\ THEN\ d_k(CF=\mu_i)$$

其中，d_j 和 d_k 為包含模糊變數（fuzzy variable）的命題，每個命題皆對應到一個介於 0 到 1 之間的真值(truth value)。而 μ_i 為確定因子（certainty factor）的值，用以表示這個法則的可信賴強度，$\mu_i \in [0,1]$。

假設 λ 為一臨界值（threshold value），$0 \leqq \lambda \leqq 1$，如果 d_j 的真實程度（degree of truth）為 y_j，$0 \leqq y_j \leqq 1$，則：

第一種狀況：如果 $y_j \geqq \lambda$，則可以驅動（fire）R_i，因此，d_k 的真實程度為 $y_j * \mu_i$。

第二種狀況：如果 $y_j < \lambda$，則 R_i 不能被驅動，d_k 的真實程度為 0。

3.2.4 模糊派屈網路

FPN 為雙向圖（bipartite directed graph），由位置（place）和轉置（transition）兩種資料型態所組成，在圖形中分別以圓圈和條形表示。每個位置可能包含了一個零到壹之間的真值之記號（token），而轉置也連結到一個零到壹之間的確定因子。位置和轉置間以有向弧（directed arc）連結。

一般而言，一個模糊派屈網路可以被表示成 8 個組值（tuple）

$$FPN=\{P, T, D, I, O, f, \alpha, \beta\}$$

其中

$P=\{p_1, p_2 \ldots p_n\}$**為一位置的有限集合。**

$T=\{t_1, t_2 \ldots t_m\}$**為一轉置的有限集合。**

$D=\{d_1, d_2 \ldots d_n\}$**為一命題的有限集合。**

$P \cap T \cap D = \psi$**，**$|P|=|D|$**，**

$I: T \rightarrow P^{\infty}$**是一個輸入函數（**input function**），為轉置到輸入此轉置之位置的對應。**

$O: T \rightarrow P^{\infty}$**是一個輸出函數（**output function**），為轉置到由此轉置輸出之位置的對應。**

$f: T \rightarrow [0, 1]$**是一個連結函數（**association function**），為轉置對應到** 0 **與** 1 **之間的實數。**

$\alpha: P \rightarrow [0, 1]$**是一個連結函數，為位置對應到** 0 **與** 1 **之間的實數。**

$\beta: P \rightarrow D$ **是一個連結函數，為位置到命題的** 1 **對** 1 **對應。**

假設 A **為一有向弧的集合。如果** $p_j \in I(t_i)$**，則存在一個從位置** p_j **到轉置** t_i **的弧** a_{ji}**，**$a_{ji} \in$ A**。 如果** $p_k \in O(t_i)$**，則存在一個從位置** t_i **到轉置** p_k **的弧** a_{ik}**，**$a_{ik} \in$ A**。若** $f(t_i)=\mu_i$**，則** t_i **連結** μ_i**。而** $\beta(p_i)=d_i$**，**$d_i \in$ D**，則** p_i **連結** d_i**。模糊推論法則** R_i**：**$IF\ d_j\ THEN\ d_k (CF=\mu_i)$ **可以表示成圖** 3.1 **的結構**

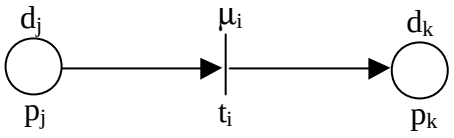


圖 3.1 **模糊派屈網路的範例**

在 *FPN* **中如果有些位置含有記號, 則稱為標記的模糊派屈網路(***marked fuzzy Petri Nets***)。我們以**

$\alpha(p_i)$

符號表示記號在位置 p_i**，**$\alpha(p_i)$ **為記號在位置** p_i **的數值，**$p_i \in P$**，**$\alpha(p_i) \in [0,1]$**。**

如果 $\alpha(p_i)=y_i$**，**$y_i \in [0,1]$**，且** $\beta(p_i)=d_i$**，表示位置** d_i **的真實程度為** y_i**。**

例如，如果定義：

FPN=(P, T, D, I, O, f, α, β),

P={p_1, p_2},

T={t_1},

D={it is raining, the ground is wet},

I (t_1)={p_1}, O (t_1)={p_2}, f (t_1)=0.90,

α (p_1)=0.90, α (p_2)=0,

β (p_1)=it is raining,

β (p_2)= the ground is wet

則可以表示成圖 3.2 **之圖形**

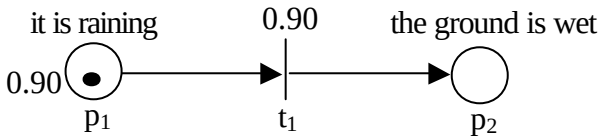


圖 3.2 **標記的模糊派屈網路**

則其驅動的狀況如圖 3.3 **所示：**

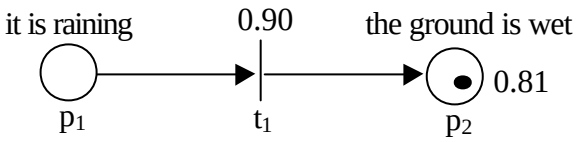


圖 3.3 **圖** 3.2 **中標記的模糊派屈網路驅動後之結果**

而模糊推論法則可以分成以下四種類型：

第一種：IF d_{j1} and d_{j2} and…and d_{jn} THEN d_k(CF= μ_i)**，表示方法如圖 所示，其推理可以表示成圖 之圖形，其中** d_{j1}**，**d_{j2}…d_{jn} **都成立，方可驅動此種推論法則，根據模糊推論的一般原則**[6]**，則** $Y_k=min_{(}y_{j1,}\ y_{j2,}..y_{jn)*}\ \mu_{i}$。

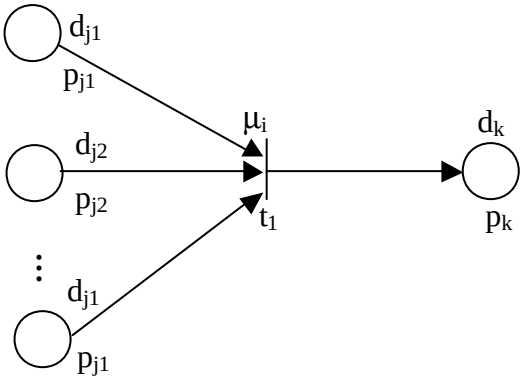
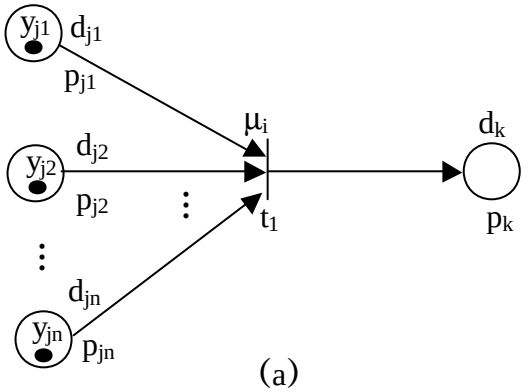
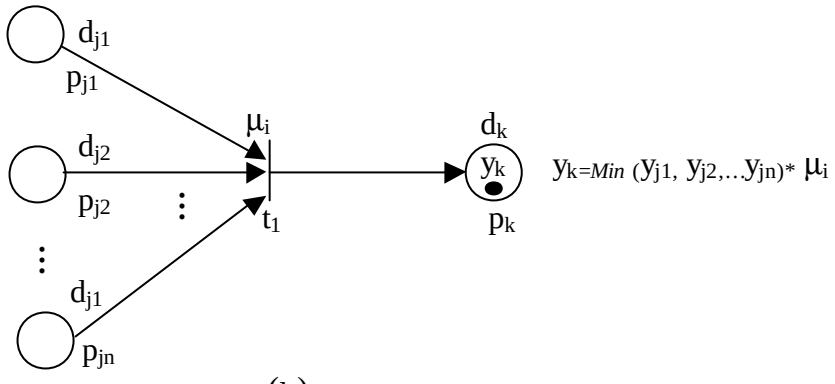


圖 3.4 **第一種類型的模糊推論法則表示方法**



(a)



(b)

圖 3.5 **第一種類型的模糊推論法則推理過程** (a)**驅動前** (b)**驅動後**

第二種：IF d_j THEN d_{k1} and d_{k2} and…and d_{kn}(CF= μ_i)，**表示方法如圖** 3.6 **所示，其推理可以表示成圖** 3.7 **之圖形，** $y_{ki}=y_j*m_i$， $i=1,2,3,\ldots,n$。

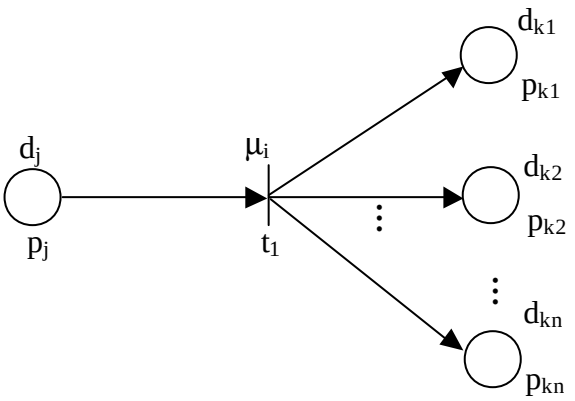


圖 3.6 第二種類型的模糊推論法則表示方法

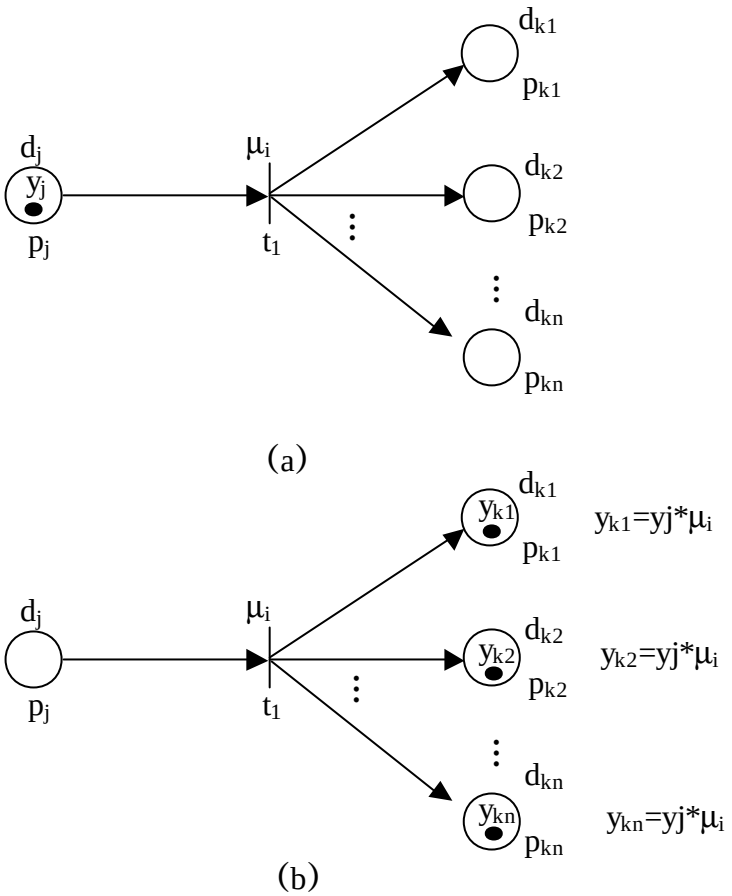


圖 3.7 第二種類型的模糊推論法則推理過程 (a)驅動前 (b)驅動後

第三種：IF d_{j1} or d_{j2} or…or d_{jn} THEN d_k(CF= μ_i)，**可以拆解成**

R_1：IF d_{j1} THEN d_k(CF= μ_i)

R_2：IF d_{j2} THEN d_k(CF= μ_i)

.
.
.

R_n：IF d_{jn} THEN d_k(CF= μ_i)

表示方法如圖 3.8 **所示，其推理可以表示成圖** 3.9 **之圖形，其中根據模糊推論的一般原則**[6]，$y_k=Max\ (y_{j1}*\mu_i, y_{j2}*\mu_{i,}...y_{jn}*\mu_i)$。

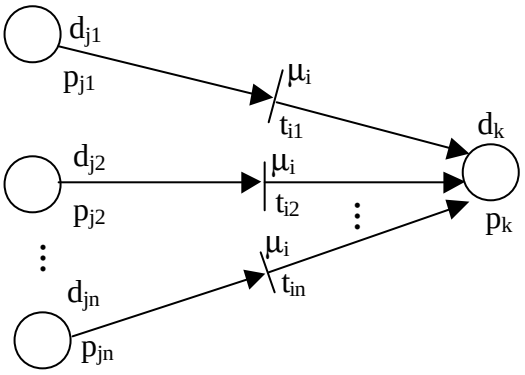
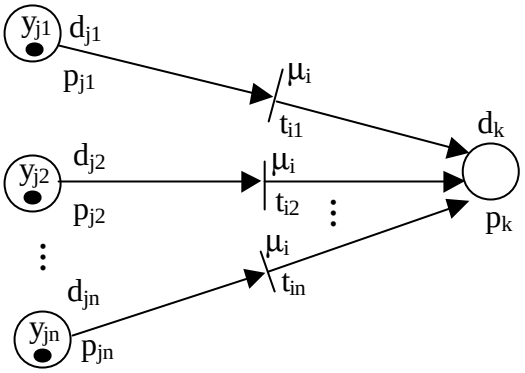


圖 3.8 **第三種類型的模糊推論法則表示方法**



(a)

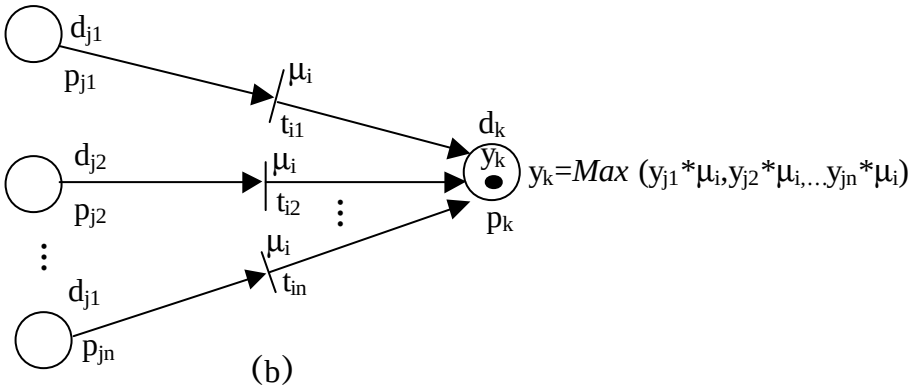


圖 3.9 **第三種類型的模糊推論法則推理過程** (a)**驅動前** (b)**驅動後**

第四種：IF d_j THEN d_{k1} or d_{k2} or…or d_{kn}(CF= μ_i)，**可以拆解成**

R_1：IF d_j THEN d_{k1}(CF= μ_i)

R_2：IF d_j THEN d_{k2}(CF= μ_i)

.
.
.

R_n：IF d_j THEN d_{kn}(CF= μ_i)

表示方法如圖 所示

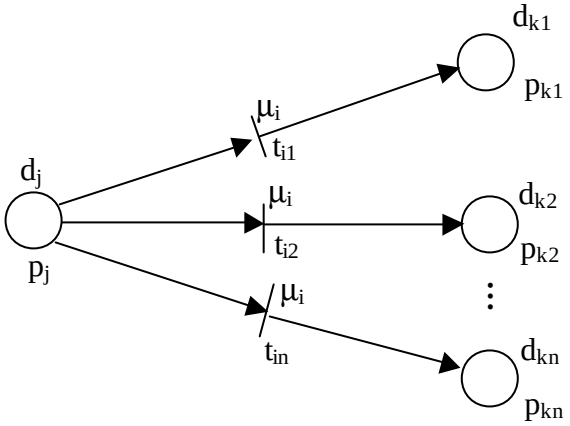


圖 3.10 **第四種類型的模糊推論法則表示方法**

由於第四種類型的模糊推論法則可以利用第一~第三種推論法則來表示及推理，而不另贅述。

3.3 聯想概念網路

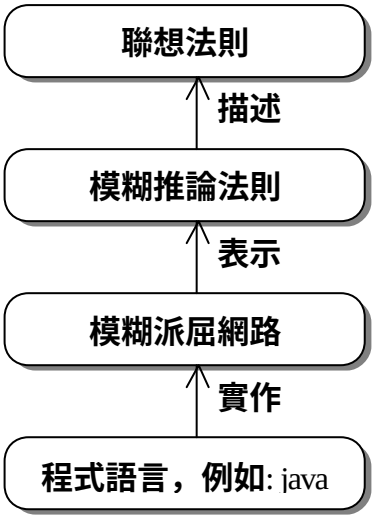


圖 3.11 聯想法則、模糊推論法則、模糊派屈網路和程式語言之間的關係圖

如圖 3.11 所示，我們可先以模糊推論法則（fuzzy production rule）描述聯想法則，再以模糊派屈網路表示模糊推論法則，並進行模糊推理，最後以程式語言，例如，Java，C++等實作模糊派區網路，期望建立自動化的推論機制，達成人工智慧的目標。而模糊推論法則以模糊派屈網路來表示，而由一群模糊派屈網路網路組成的網路圖形，我們稱之「聯想概念網路(Associative Conceptual Network, ACN)」，目的是表示人類大腦內的聯想法則，建立步驟如下：

3.3.1 定義聯想法則

建立 ACN 的第一個步驟，就是定義聯想法則以表示聯想關係。如前述，聯想法則包括觸發條件的概念、結果概念和可信賴強度。因此，定義聯想法則的首要工作就是確保概念間存在聯想關係。若兩概念之間不存在聯想關係，即使存在其他種類的關係，例如，屬性關係或部分關係，亦不能建立。例如，在文章看到「汽車」較少聯想到「輪胎」，「汽車」與「輪胎」的關係程度低，即使「輪胎」

為「汽車」之一部分，亦不建立「汽車」到「輪胎」間的聯想法則。其次就是定義每個法則的可信賴強度 CF，$CF \in [0,1]$。對於人類的思考而言，聯想關係是概念之間必要的連結。然而，為了有效模擬人類的思考模式，避免建立過多的聯想法則造成推論時過大的計算複雜度，應只紀錄"直接連結"，使用者可以事先訂定一個臨界值，CF 小於此臨界值的則不加入 ACN 中，間接連結則以遞移式規則推導之，亦不加入 ACN 中。

3.3.2 以模糊推論法則表示聯想法則

假設概念 A 和 B 有聯想關係，對於一篇文章 D 而言，若 A 在 D 中非常重要，具有相當的代表性，又 A 和 B 的聯想法則可信賴強度大時，則表示 B 也很重要。因此，A 和 B 不僅可以作為瞭解 D 的依據，也可以之檢索 D。對於聯想關係而言，B 的重要性除了來自 A 的重要性之外，也來自和 A 之間的聯想關係強度。

換言之，文章中若出現 A 時，可以藉由"A 本身"和"A 與 B 的聯想法則"來決定 B 的重要性。因此，若要以模糊推論法則 F 來表示聯想法則 R，必須定義 R 的觸發條件（IF 部分）中的每個概念在一篇文章 D 中的重要性，及 R 的結果概念（THEN 部分）受觸發條件影響之重要性。例如，有一聯想法則為 if「素人自拍」then「色情」，其可信賴強度為 0.8，對於文章甲則可以表示成模糊推論法則 R_1：

R_1：*IF*（「素人自拍」對甲之重要性）*THEN*（「色情」對甲之重要性）*(CF=0.8)*

其中，命題（「素人自拍」對甲之重要性）的真實程度即是「素人自拍」對甲的實際重要程度，命題（「色情」對甲之重要性）的真實程度即是「色情」對甲的實際重要程度。模糊推論法則中的每個命題的真實程度必須介於零和壹之間。但概念在文章中的重要性只用於比較各概念的重要程度，雖然也是數值，但

不一定介於零和壹之間。例如，概念 A 和 B 在文章 D 中的重要性為 8000 和 6000，但也可以寫成 0.8 和 0.6，都是代表概念 A 在 D 中的重要性大於 B，因此，必須重新定義模糊推論法則中的真實程度為大於零的值，且沒有上限，以適用於聯想法則。

3.3.3 以模糊派屈網路表示聯想法則

以模糊推論法則描述聯想法則後，接著將以模糊派屈網路表示模糊推論法則。

經過以上三個步驟後，建構出 ACN。每個使用者都可以根據自己定義的聯想關係建構自己的 ACN，由於每個人對於概念間的聯想關係不盡相同，因此建構出來的 ACN 也不會相同。再者，人類的知識相當多，想要以 ACN 完全表達出人類大腦中的聯想關係實屬難事，況且建構出所有的聯想關係對於資訊檢索也不見得有太大的用處。例如，某甲希望檢索出醫學的資料，此時只要醫學的 ACN 即可，和醫學無關的知識和關係則不應涉入。結論是 ACN 中所建構的知識應依專業領域分類，俾建立各種不同類別之 ACN。

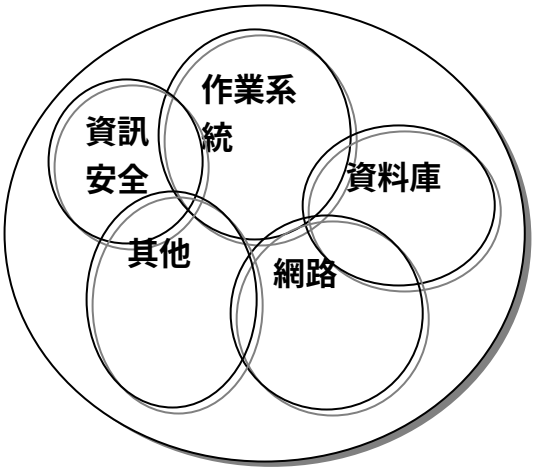


圖 3.12 計算機的 ACN 和子類別的 ACN

而每個類別的 ACN 中可以包含其子類別的子 ACN，如圖 3.12 所示為一例

子，其中計算機的 ACN 包含作業系統、資料庫、網路、資訊安全或其他子類別的 ACN，使用者可以根據需求取出或建立對應類別的 ACN。

這種類別和子類別之間的階層式關係，可以去除了許多在檢索中不是很必要的聯想關係（即聯想關係不夠強烈），而大幅降低計算的複雜度，ACN 也可以根據需要包含不同類別的 ACN。例如，我們要做一個網路安全的 ACN，則必須包含網路方面的 ACN 以及資訊安全的 ACN。但概念 A 和 B 在網路方面的聯想關係在資訊安全方面不見得相同，例如「思科網路公司」與「路由器」在網路方面的聯想關係會大於資訊安全方面，我們應降低這樣的情形，以避免過度的人為干涉（太主觀），降低了實用性。或者概念 A 和 B 雖然在網路和資訊安全方面的聯想關係相同，但這樣的關係又不一定適合網路安全領域，例如「阻絕服務攻擊（Denial Of Service）」與「駭客」在網路安全領域的聯想關係會分別大於在網路和資訊安全領域，這時 A 和 B 的聯想關係就必須重新定義。因此，如需建造一個跨越不同類別的 ACN，就必須先將各類別的 ACN 中的每個概念擺在同一個 ACN 中，並檢視和重新定義每個概念之間的聯想關係，如此才能確保利用跨不同類別的 ACN 進行檢索的正確性。

此外，在聯想概念網路中，我們修改模糊派屈網路的可驅動法則（Enabled Rule）和驅動法則（Firing Rule）如下，以圖 3.1 為例：

可驅動法則：若 $\exists p_j \in I(t_i), a(p_j) \geq l$ ， l 為使用者定義的臨界值，$0 \leq l \leq 1$，則稱 t_i 為可驅動的。

驅動法則：當 t_i 驅動時，會將記號從 t_i 輸入的位置移去，並在每個 t_i 輸出的位置置放一個記號，每個記號的數值為 $y_i * \mu_{io}$。

第4章　概念排名

4.1 主要概念和輔助概念

一篇文章是由許多概念所組成，有些只是用來修飾語句使之更加通順，對於了解文章意義並無太大助益；有些則是作者想要傳達給讀者的訊息。當讀者閱讀且領受了以後，就可以很清楚內容旨趣者，稱為主要概念。主要概念表露文章的意義，及語意核心，因此又可視為文章的摘要，也是讀者了解文章的重要依據。其中讀者會先決定概念 A 的重要性，利用此重要性與其他概念進行排名，並從中決定文章的主要概念。傳統資訊檢索系統以人工方式建立代表文章意義的關鍵字/詞，以為使用者查詢該文章是否符合所需時之用，這些關鍵字/詞，其實就是主要概念的一部份，通常必須將每篇文章都審閱完畢，並深入瞭解後才可以建立。否則，可能會建置不符合或無法表達文章意義的關鍵字/詞，當然檢索結果可能就無法符合使用者需求。結論是，以人工建立文章的主要概念既費時也不見得正確。如果可以電腦替代人工，而事先在系統內建立各個領域的專家知識，並在輸入文章至系統時自動擷取其主要概念，將可以提昇建立文章摘要系統的效率，及節省大量的人力和時間。因此如何從文章中取出主要概念就是文章檢索的首要工作了。

再者，一篇隱喻式的文章寫作方式，是利用與主要概念相關的概念 R 間接地表達，側面式地引導讀者逐步地歸納出主要概念，R 稱為輔助概念。每個人的背景知識不同，寫作的方式也可能不一樣，因此雖然都是利用隱喻的方式表達相同的主題，輔助概念卻不見得相同。換句話說，兩篇文章的輔助概念雖然不同，但是主要概念有可能是相同的。

4.2 概念推導

藉由輔助概念來推導主要概念，事實上就是輔助概念和主要概念間存在著聯想關係。如果文章 D 出現某概念 A 的輔助概念越多，讀者很容易聯想到 A，越表示 A 在 D 的重要性。例如：文章甲和乙皆含有六十個概念，其中各自出現可以聯想到 A 的輔助概念分別為二十七個和十個，假設輔助概念在文章內的重要性以出現的次數為計算基礎，且每個聯想到 A 的程度皆相同，由於輔助概念之重要性係由概念出現次數及文章內概念的總數而來，非相對的數值，意即 A 在甲的重要性會大於乙，代表甲中較乙容易聯想到 A。因此，A 的輔助概念個數將會決定 A 在文章內的重要性。除此之外，輔助概念和主要概念的聯想程度也會決定主要概念在文章內的重要性，例如，甲中可以聯想到 A 的輔助概念雖然有二十七個，但是如果 A 的每個輔助概念可以聯想到 A 的程度皆為 0.2，而在乙中卻為 0.9，則 A 在甲的重要性（0.2×2=0.4）會小於在乙者（0.9×10=9）。

主要概念也有可能是其他概念的輔助概念。例如：A 可以聯想到 B，而 B 又可以聯想到 C，當 B 和 C 都是主要概念時，B 既是主要概念也是 C 的輔助概念。因此，A 在文章中的重要性不僅會影響 B 的重要性也會影響 C 的重要性。也就是說，只要某個概念在文章中具有重要性，將會影響其他有聯想關係的概念之重要性，再由這些概念影響其他概念，如此不斷地推進，我們稱為「遞移性推導」（Transitive Inference）；惟概念間聯想的程度定義在[0,1]，因此某概念可以影響其他概念重要性的數值會隨著聯想的層數逐漸減小，最後甚至趨近於 0。例如，A 可以聯想到 B，B 可以聯想到 C，其中聯想程度分別為 0.8 和 0.2，如果 A 的重要性為 0.9，則 A 給 B 帶來的重要性為 0.72(0.8×0.9)，帶給 C 的重要性卻只有 0.144(0.72×0.2)。這與人類的思考極為相似，當我們腦中有了某些概念，會聯想出其他概念，再由這些概念繼續引導出其他概念。但是如果聯想的層次過多，所聯想出來的概念就會與原始的概念漸行漸遠，因此通常也不會利用過多層次的概念來了解文章的意義。

假設概念 B 和 C 皆可以聯想到概念 A，當甲文章出現了 B 和 C，也必然隱含了 A。若文章中又出現可以聯想到 A 的概念 D，此時比只含有 B 和 C 時，更能確定 A 在文章中的重要性。意即，當一個概念可以被聯想的機會越多，越能確定它在文章中的重要性，我們稱為「概念合導」(Combinational Inference of Concepts)。因此，假使我們可以算出每個概念的重要性，並進行排序，而排名在最前面者，應該可以確定就是該文章的主要概念。至於概念的重要性該如何計算呢？以下敘述之。

假設一篇文章內的概念集合 $C=\{C_1,C_2,C_3,\ldots,C_n\}$ 中的每一個概念 C_i 都可以直接聯想到概念 A，則根據聯想法則的表示法，A 的重要性會受到 C 的影響，這種影響擬分成兩類來考量：第一類是以 C_i 中影響最大者，以其影響值為 A 的重要性，稱為「最大貢獻法」。如果取 C 的影響值總和，亦是一種累加的影響力，我們稱之「總和貢獻法」，其重要性可能會超過 1。若以最大貢獻法計算，圖 4.1 中 A 的重要性為 0.54（Max(0.6×0.9，0.8×0.5)），若以總和貢獻法則為 0.94（0.6×0.9+0.8×0.5），以下詳述之。

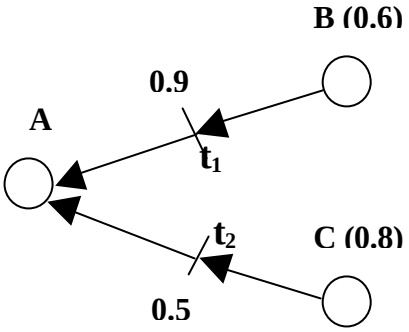


圖 4.1 聯想概念網路，其中 B 和 C 的重要性分別為 0.6 和 0.8

4.3 最大貢獻法

概念 A 在文章內的重要性決定了 A 是否為文章的主要概念，因此，對於聯想概念網路而言，決定 A 的重要性便成為找出文章內主要概念的重要課題。

人們瀏覽文章時常會藉由其中的許多概念直接或間接地聯想到 A，再由 A

聯想到 B，而如此持續聯想下去，則可將一篇文章的各主要概念串聯在一起，而成為作者想要借助文字透露的訊息。最大貢獻法就是選擇影響 A 最大的概念之影響值作為 A 的重要性。例如，B 和 C 都可以聯想到 A，當文章內先看到 B 時，我們會由 B 來決定 A 的重要性；而後如果後續在文章內看到 C，假設 C 與 A 的聯想關係大於 B 與 A，則會利用 C 取代 B 來決定 A 的重要性。反之，如果 C 與 A 的聯想關係小於 B 與 A，則我們還是會以 B 決定 C 的重要性，一個現實的例子，就是「石油」和「伊拉克」，雖然兩者都可以聯想到「美伊戰爭」，但是如果兩個字詞皆出現在文章內，還是會以「伊拉克」來決定「美伊戰爭」在文章內的重要性，因為「伊拉克」較「石油」容易聯想到該戰爭。而必須強調的是，概念本身的初始重要性也一定要考慮（通常由向量模式考慮 A 出現在文章的次數定義此值）。例如，「美伊戰爭」已經出現在文章內，其重要性為 0.9。此時，即使「石油」和「伊拉克」影響「美伊戰爭」的重要性分別為 0.7 和 0.8，仍然不會影響「美伊戰爭」為 0.9 的重要性。

定義 4.1：最大貢獻值

假設一篇文章 T 的所有概念集合 $G=\{G_1,G_2,G_3,\ldots,G_m\}$，$G_k \in G$，$k=1\ldots m$，其中有一概念集合 $C=\{C_1,C_2,C_3,\ldots,C_n\}$，$C \subseteq G$，$\forall Ci \in C$，$Ci$ 可以聯想到 G_k。假設 G_k 自身的重要性為 $R'(G_k)$，C 貢獻給 G_k 的重要性分別為 $\{R(C_1 \rightarrow G_k),R(C_2 \rightarrow G_k)\ldots R(C_n \rightarrow G_k)\}$，令 G_k 的重要性 $R(G_k)=Max(R'(G_k),R(C_1 \rightarrow G_k),R(C_2 \rightarrow G_k)\ldots R(C_n \rightarrow G_k))$。

定義 4.2：最大貢獻之主要概念

若 $R(G_k) \geq$ Y 則定義 G_k 為 T 的主要概念，稱為最大貢獻主要概念，其中 Y 為使用者定義之臨界值。這種以最大貢獻值為 G_k 之重要性加權，俾篩選 T 的主要概念的方法，稱為最大貢獻法。

利用最大貢獻法取出文章 T 的（最大貢獻）主要概念的步驟如下：

(1)輸入事先建立的 *ACN*。

(2)依照指定類別取出相關的子類別 *ACN*，組合成 *ACN'*，$ACN' \subseteq ACN$，假設 *ACN'*

之概念集合為 $C=\{C_1,C_2,\ldots C_p\}$。

(3)取出 T 之關鍵詞集合 $K=\{K_1,K_2,\ldots,K_L\}$。

(4)以向量模式計算 x(x∈K)在 T 中的重要性 $R'(x)$，每一 x 均在 C 中有一對應之 C_i

$$R'(x)=f_{x,T}*\frac{idfx}{Max_i idf_i}$$

其中，$f_{x,T}=\frac{freqx_{,T}}{\max_l freq_{l,T}}$，$freqx_{,T}$ 表示 x 出現在 T 的次數。$\max_l freq_{l,T}$ 表示 T 中所有的關鍵字 l 在 T 中出現的次數而取最大值。$idf_x=\log\frac{N}{n_x}$，N 為系統的文件總數，n_x 為所有出現 x 的文件總數。$Max_i idf_i$ 則表示 T 中所有關鍵字的 idf 取最大值。

(5)再以下公式計算每個概念 $u(u\hat{I}C)$在文章中的重要性 $R(u)$，其定義如下

$$R(u)=\begin{cases}(\tilde{\bigvee}_{t\in I(u)}FP(t))\ \textcircled{\vee}\ R'(u) & \text{如果 u 已存在重要性 } R'(u)\\ \tilde{\bigvee}_{t\in I(u)}FP(t)) & \text{其他}\end{cases}$$

$I(u)$是輸入至 u 的轉置集合（*transition set*），$FP(t)$則為轉置 t 傳送給 u 的重要性之值，這項傳送動作稱為驅動能量（*Firing Power,* FP）。若 t 傳送至 u 的驅動能量未到達臨界值 e，則令該驅動能量為 0，其所產生之重要性亦為 0。

其中符號$\tilde{Y}$的定義如下:

$\tilde{\bigvee}_{i=1}^{n}t_i=t_1\textcircled{\vee}t_2\textcircled{\vee}\ldots\textcircled{\vee}t_n$，其中 $t_1,t_2\ldots t_n$ 均為 [0,1] 之模糊值，$t_i\textcircled{\vee}t_j=\max(t_i,t_j)$。

此步驟完成後，$\forall u\hat{I}C$，$R(u)$即為 u 在 T 中的重要性，以 $ACN'(T)$表示，其意義是 T 之所有概念依 ACN'所獲得重要性集合。

(6)$\forall u\hat{I}C$，排序 $R(u)$得到 $C=\{C_1',C_2',\ldots,C_n'\}$，如果 $i<j$，則 $R(C_i')\geq R(C_j'),1\leq i,j\leq n$ 且 i≠j，假設 $R(C_k')\geq Y$ and $R(C_{k+1}')<Y$，則 $MC=\{C_1',C_2',\ldots,C_k'\}$稱為 T 的最大貢獻主要概念排名（簡稱 T 的最大貢獻排名），Y 為使用者定義之臨界值。

假設 ACN 包含 n 個概念，則本法中 ACN 的每個概念都要與其他概念進行比較並取最大值，對每個概念而言，找出重要性的時間複雜度為 O(n-1)，而要算出 ACN 所有概念的重要性，則需要花費 O(n(n-1))。

4.4 總和貢獻法

定義 4.3：總和貢獻值

假設一篇文章 T 的所有概念集合 $G=\{G_1,G_2,G_3,\ldots,G_m\}$，$G_k \in G$，$k=1\ldots m$，其中有一概念集合 $C=\{C_1,C_2,C_3,\ldots,C_n\}$，$C \subseteq G$，$\forall Ci \in C$，$Ci$ 可以聯想到 G_k。假設 G_k 自身的重要性為 $R'(G_k)$，C 貢獻給 G_k 的重要性分別為 $\{R(C_1 \rightarrow G_k),R(C_2 \rightarrow G_k)\ldots R(C_n \rightarrow G_k)\}$，令 G_k 的重要性 $R(G_k)=Sum(R'(G_k),R(C_1 \rightarrow G_k),R(C_2 \rightarrow G_k)\ldots R(C_n \rightarrow G_k))$。

定義 4.4：總和貢獻值之主要概念

若 $R(G_k) \geq$ Y 則定義 G_k 為 T 的主要概念，稱為總和貢獻主要概念，其中 Y 為使用者定義之臨界值。此種取總和貢獻值為 G_k 之重要性加權，俾篩選 T 的主要概念的方法稱為總和貢獻法。

總和貢獻法與最大貢獻法，在步驟(1)、(2)、(3)、(4)和(6)的做法完全相同，僅在第(5)步驟時的公式改成：

$$R(u)=\begin{cases}(\sum_{t\in I(u)} FP(t))+R'(u) & \text{如果 u 已存在重要性 } R'(u)\\ \sum_{t\in I(u)} FP(t) & \text{其他}\end{cases}$$

當然，如最大貢獻法，從步驟(6)亦可獲得總和貢獻主要概念排名（簡稱總和貢獻排名）。

由以上公式可以知道，每個概念要找出其重要性之時間複雜度為 O(n-1)，而要找出 ACN 每個概念的重要性之時間複雜度為 O(n(n-1))，由於為加法的緣故，每個概念都必須等待連接的概念之重要性漸漸穩定，而連接的概念又必須等待其

他概念重要性可以穩定，如此不斷等待，最多每個概念必須等待 n 個概念穩定，因此其時間複雜度為 $O(n(n-1)^2)$。

4.5 範例

假設使用者建立了一個關於美伊戰爭的聯想概念網路，如圖 4.2 所示。則有段簡短新聞如下：

在美國 20 日對伊拉克發動攻擊近 8 個小時後，法國總統席哈克在此間發表鄭重聲明，對美國繞過聯合國發動戰爭深表遺憾。他希望這場戰爭能夠儘快結束，並儘量減少人員傷亡。[7]

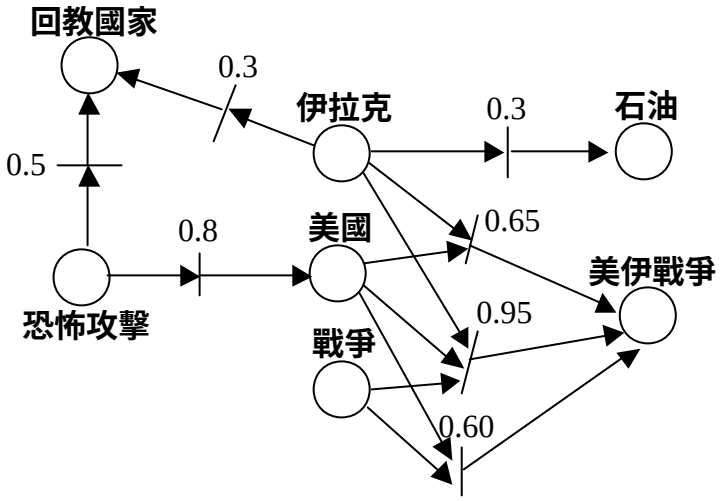


圖 4.2 美伊戰爭的聯想概念網路

以上文章中去除高頻字（stop word），高頻字為每篇文章都會經常出現的字，相對地就比較不重要，例如「小時」、「他」和「遺憾」等字，假設去除高頻字可以取出的關鍵字分別為：

美國（2），伊拉克，法國，席哈克，鄭重聲明，聯合國，戰爭（2），人員傷亡，共 8 個。

其中，括號內的數值表示該關鍵字出現在文章內的次數，沒有括號的關鍵字代表僅出現一次。因為本例僅考慮一篇文章，會造成 *idf* 的值為 0，因此 $R'(u)$將

只考慮 $fx_{,T}$，如圖 4.3 所示，伊拉克、美國和戰爭分別為 0.5（=1/2），1.0（=1/1）和 1.0（=1/1）。

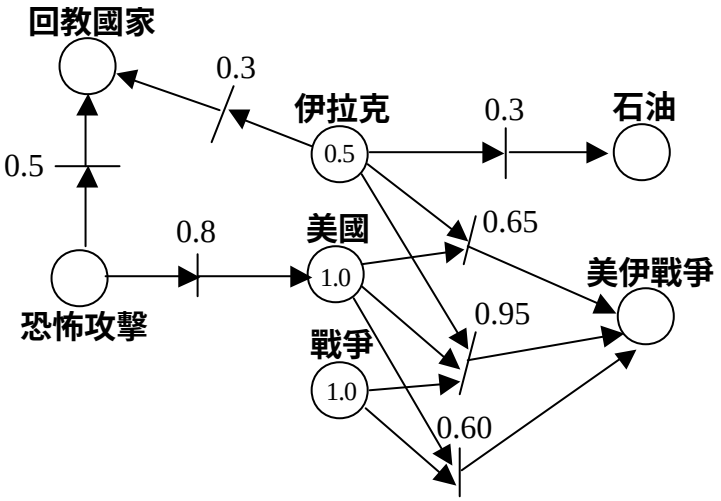


圖 4.3 聯想概念網路及其概念重要性之初始值

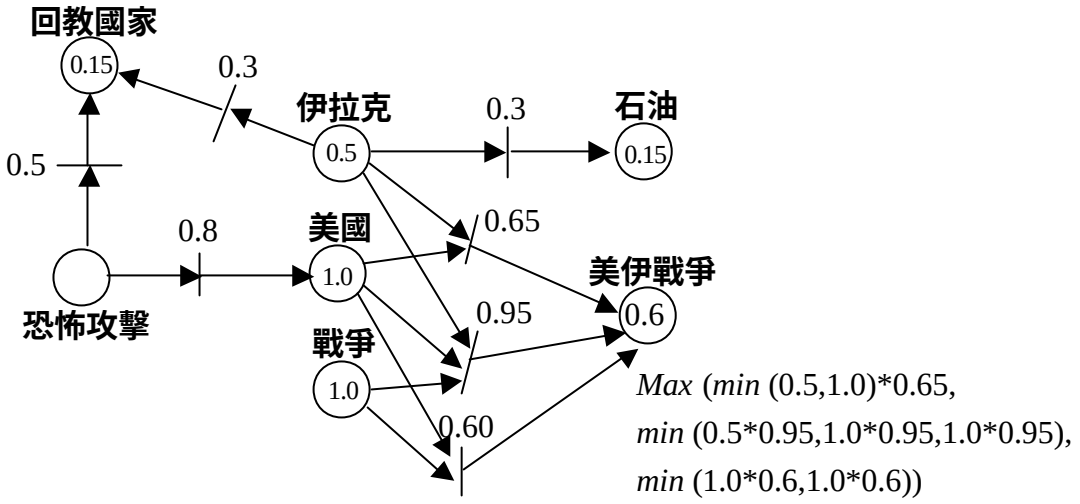


圖 4.4 經最大貢獻法計算後的聯想概念網路

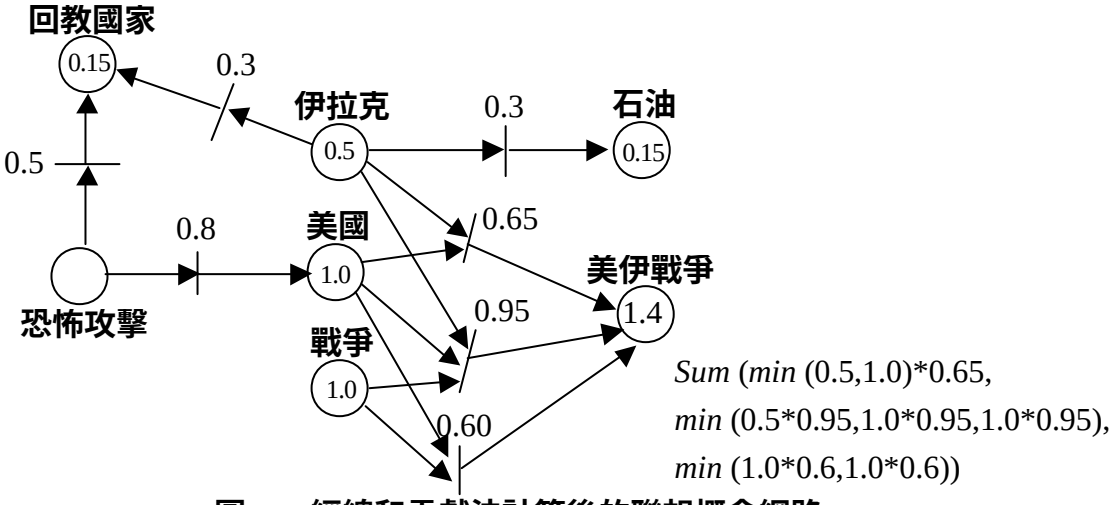


圖 4.5 經總和貢獻法計算後的聯想概念網路

如果利用最大貢獻法，可得到圖 4.4 的結果，若以 0.5 作為臨界值，則美國、戰爭、美伊戰爭和伊拉克都是本段文章的主要概念，其他關鍵詞對於了解文章意義並無太大助益，則不予理會。如果利用總和貢獻法，會得到圖 4.5 的結果，若以 0.5 作為臨界值，美國、戰爭、美伊戰爭和伊拉克仍是本段文章的主要概念，和最大貢獻法有相同的結果。其中，雖然文章未提到美伊戰爭，但兩種方法皆可推導出美伊戰爭，且是主要概念。

第5章　文章相似比對

5.1 以概念為基礎之比對

主要概念集合 CS 呈現了文章的大意，其重要性係由向量模式根據主要概念出現在文章的次數作為最大貢獻法或總和貢獻法的初始值算得，因此某主要概念 $C_i \in CS$ 在文章 D 的重要性也代表 C_i 足以表示文章意義的程度；若兩篇文章 D_1 和 D_2 的意義越相似，通常也應該會包含越多相同的主要概念 $CC=\{C_1,C_2,\ldots,C_k\}$，$C_i \in CC$，R(Ci)在 D_1 和 D_2 之值（重要程度）也會很相近。相反的，如果$\forall\, C_i \in CC$，$R(Ci)$在 D_1 和 D_2 之值很相近，則 D_1 和 D_2 之意義亦應該很相似。由於主要概念著重在文章的語意，不同於以往傳統資訊檢索系統僅是以兩文章內共同關鍵字及其出現次數作為文章相似程度的比對，所以比較的結果將更精確，更能滿足使用者檢索的需要。

假設一篇文章 T 包含了主要概念 $TC=\{T_1,T_2,\ldots T_n\}$，$\forall T_i \in TC$，令 T_i 的重要性為 $R(A_i^T)$。現在有兩文章 A 和 B，令 $sim(A,B)$ 為 B 對 A 的相似性，其中 A 中包含主要概念 $AC=\{A_1,A_2,\ldots A_n\}$，B 中的主要概念 $BC=\{B_1,B_2,\ldots B_m\}$，則 $sim(A,B)=x(AC \rightarrow BC)$。若 $A_i \in AC$ 且 $A_i \in BC$，$|R(A_i^A)- R(A_i^B)|=0$，則稱「B 完全滿足 A 的 Ai」，以 SAT(B,A)表示；若 $lim(|R(A_i^A)- R(A_i^B)|) \rightarrow 0$，則稱「B 部分滿足 A 的 Ai」，以 PSAT(B,A)表示，其意義是 B 中所有概念部分滿足 A 所有概念的程度。

定義：B 對 A 之靜態相似性

令 $C(B)=\{B_i|\ B_i$ 滿足 SAT(B,A)或 PSAT(B,A)$\}$，則 B 對 A 之靜態相似性 $sim(A,B)=x(AC \rightarrow C(B))$，其中 A 為標準者（又稱為被比較者），B 為比較者。

B 能滿足或部分滿足 A_i，$i=1,2,\ldots,n$ 的數量越多，則 B 越相似於 A。例如，A 的主要概念 $G=\{C_1,C_2,C_6,C_8\}$，B 的主要概念 $F=\{C_1,C_8\}$，假設 G 和 F 中每個元

素的重要性皆為 0.8，則若以 A 作為標準者，B 作為比較者，因為 F⊆G，則 B 僅滿足了 A 四項中的 C_1,C_8 兩項, 所以 $sim(A,B)$ 為 50%；但是 $sim(B,A)$ 卻有 100%，因為 A 已經完全滿足了 C_1,C_8 兩項。

接下來將針對最大貢獻法和總和貢獻法的特性提出判斷文章相似程度的方法，由於兩種方法都是以事先取得的主要概念（包括隱含在文章內的概念）來判斷，比較過程中不會再引入其他的概念，因此又稱為靜態比較法（Static Comparison）。

5.1.1 以最大貢獻為基礎

假設文章 T 的主要概念為 $C=\{C_1 \ldots C_m\}$，文章 S 為 $E=\{E_1 \ldots E_n\}$，而以 T 為標準，S 與之比較，則有以下步驟：

K	E_1 . E_j . E_n
C_1	$K_{1,1}$ . . . $K_{1,n}$
.	.
C_i	. $K_{i,j}$
.	.
C_m	$K_{m,1}$ . . . $K_{m,n}$

圖 5.1　T 和 S 的相似性矩陣

(1) 建立相似性矩陣 K，以 C 為縱座標，E 為橫座標，見圖 5.1。其中，$K_{i,j}$ 是 E_j 對 C_i 的聯想值，其意義為 ACN’中，可以從概念 E_j 以各種聯想路徑聯想到概念 C_i 的最佳（大）值。例如，圖 5.2 中的 $K_{A,D}=0.315$，如果 E_j 在 ACN’中無法到達 C_i，則 $K_{i,j}=nil$。

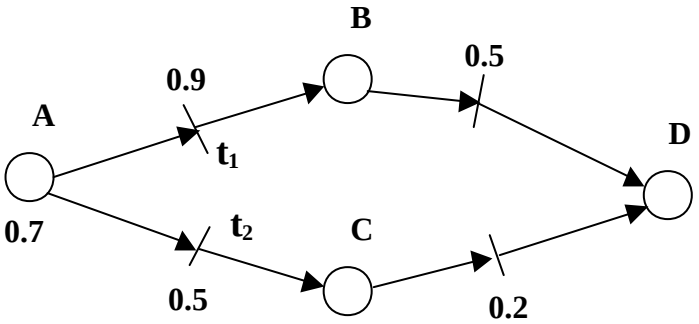


圖 5.2　**概念** A **與** D **之間之最佳聯想路徑為** ABD

K	E_1	E_2	E_3
C_1		. . .	
C_2(0.7)	0.3	0.8	0.9
C_3		. . .	

圖 5.3　S_2 **為** T_2 **的相似夥伴**

(2) **對** T **中的每個主要概念** C_i**，取出與其重要性最相近之** S **的概念** E_j**。意即**

$\forall C_i, \exists E_j \in S$**，** $abs(|C_i| - K_{i,j}) = \min(abs(|C_i| - K_{i,1}), abs(|C_i| - K_{i,2}), \mathrm{L}, abs(|C_i| - K_{i,n}))$**，**

我們稱 E_j **為** C_i **之相似夥伴（***Similarity Partner***），其中，**abs(X)**為取** X **之絕對值。例如，圖** 5.3 **中，與** C_2 **之重要性** 0.7 **最相近者為** E_2 (0.8)**，因此** E_2 **為** C_2 **的相似夥伴。**

(3) **假設** T **所對應的相似夥伴為** $E' = \{E_1', E_2', E_3', ..., E_m'\}$**，則** T **和** S **的相似程度的計算方法目前有許多種，例如，布林模式**(Boolean Model)**、向量模式和機率模式**(Probability Model)**等，其中向量模式：**

$$sim(T,S) = \cos ine(T,S) = \frac{\sum_{i=1}^{m} |C_i| \times |E_i'|}{\sqrt{\sum_{i=1}^{m} (|C_i|)^2} \times \sqrt{\sum_{i=1}^{m} (|E_i'|)^2}}$$

為最常用者，因此本論文採用之。找 T **的相似性夥伴時間複雜度為** $O(n^2)$**，而以向量模式計算** T **和** S **的相似性夥伴之重要性也為** $O(n^2)$**，因此，本法之時間複雜度為** $O(2n^2)$**。**

我們可以利用 T 中主要概念 P 之重要性 P_1 和 S 中主要概念 P 之重要性 P_2 差值判斷 T 和 S 的相似程度，當差值越小，則 T 和 S 越相似。即便如此，人類思考中並不會全部用最大貢獻法（或總和貢獻法）求得的重要性直接比較，讀者可能會因為個人喜好、認知及知識背景，從 T 中選定自己認定適當的數值 P_3 取代 P_1，並與 P_2 比較， P_3 通常是以 T 的其他主要概念聯想到 P 所導引出來的重要性，或是讀者主觀認定的數值。由於這些主觀的因素太過複雜也不易掌握，容易造成不同的人即使擁有相同的 ACN，在比較 T 和 S 時卻有不同的相似程度，為了觀念上的一致，本論文將以兩文章的最大相似程度為兩者之相似性。

根據靜態相似性的定義，如果可以從 T 中找到一概念可以聯想到 P 且其聯想值 P_4 最接近 P_2，並用 P_4 的 P_2 的差值計算文章的相似性，將可以得到 T 和 S 的最大相似性。因此，步驟(1)係利用 *ACN* 計算 T 中所有主要概念可以聯想到 E_j 的值，而步驟(2)再從這些值選出與 E_j 最相近者作為 E_j 的相似夥伴；相似夥伴的意義為 T 中可以聯想到 E_j 並與 E_j 的值最相近之概念。最後，步驟(3)則利用向量模式整合 T 的主要概念與其相似夥伴的差異程度作為 T 和 S 的相似性。

5.1.2 以總和貢獻為基礎

假設文章 T 以總和貢獻法算出的的主要概念為 $C=\{C_1,C_2,\ldots,C_m\}$，文章 S 為 $E=\{E_1,E_2,\ldots,E_n\}$，以總和貢獻所得的主要概念來計算相似度的方式如下（仍以 T 為標準，S 與之比較）：

其中，$|C_{i,s}|$ 為概念 C_i 在 S 的重要性，即 S 中與 C_i 相同概念的元素之重要性，

$$sim(T,S)=\frac{\sum_{i=1}^{m}|C_i|\times|C_{i,s}|}{\sqrt{\sum_{i=1}^{m}(|C_i|)^2}\times\sqrt{\sum_{i=1}^{m}(|C_{i,s}|)^2}}$$

若 $C\cap E=f$，則相似程度為 0。本法的相似性比對係利用向量模式，因此，其時間複雜度為 $O(n^2)$。

5.1.3 靜態比較的特性

靜態比較係以 ACN 中的概念為計算相似性之依據，其中 ACN 包含明顯概念及隱含概念，這些概念往往會表露出文章意義，因此靜態比較的基礎為文章意義，可以得到一項結論：檢索效果佳於以關鍵字為主的文章比對方法。因為傳統以關鍵字為主的文章比對方法是假設各關鍵詞都是獨立的個體，而以這些關鍵字（或加上其在文章中出現頻率）判斷兩篇文章的相似性。然而，這種方法並沒有考慮隱含在文章內的概念。由於每個人的寫作風格都不盡相同，可能會使用不同的字詞呈現相同的主題；此時可能會因作者的表達方式不同，將兩篇主要概念相同的文章誤判為不相似或不太相似。

若能將這些取出的關鍵字/詞再依據 ACN 中的聯想關係（概念間的連結）進行推論，導引出一些隱含的概念，必可得到比較客觀的重要性數據。因此，我們認為應該會比靜態比較的檢索效果好。

5.2 以聯想過程為基礎之比對

5.2.1 中間概念

閱讀文章時，通常我們會以文章內一部份或全部概念，思考其意義和它們之間的可能關係，才能確定某些概念在文章內是否為主要概念。而要決定 A 是否為主要概念，如前述，就必須借助於 A 的初始重要性、輔助概念 B 的重要性及 A 與 B 之間的聯想關係。輔助概念越重要，聯想關係越強，作用於 A 的重要性也越大。

而 ACN 中聯想關係的存在與否往往也是專家等級的區別，例如，B 和 C 都是 A 的輔助概念，但有些專家，例如甲，對於該領域某部分知識不熟悉，也不熟諳 A、B 和 C 之間的關係，就不會在 ACN 中建立這些概念之間的連結，也就

無法利用 B 和 C 決定 A 為主要概念。但是如果此時可以經由他人指導或講述某些概念，逐步地引導甲從 B 和 C 聯想至 A，建立 B、C 和 A 之間的關係，往後看到 B 和 C 的時候，就會直觀地聯想起 A。指導者可能會教導甲直接建立 B、C 和 A 的關係，也可能為了說明方便或者使甲容易瞭解，在 B、C 和 A 之間加入其它概念 D，讓讀者先從 B、C 聯想到 D，再從 D 聯想到 A，之後甲看到 B 或 C，就會藉由 D 聯想到 A。因此對於甲而言，因為 D 使得自己在該領域的專業可以更精進，因此 D 相當重要，也會在比較含有 B 和 C 的文章之相似性時，將 D 納入判斷的依據。我們稱這種在聯想過程中出現的概念為「中間概念」。

中間概念係根據聯想的特性，模擬人類思維的過程。當我們思考某些事物時，常會利用經驗或學習過的知識判斷概念之間的關係，這種經驗或知識往往決定了我們判斷事物的方向，甚而影響學習新知識的能力。一般資訊檢索僅僅比較出現在文章內的關鍵字，卻忽略了或無法表示這些相關的經驗或知識；相反的，若考慮太多與這些概念相關的特性，即使相關性薄弱者亦一併納入相似性的計算，反之有降低計算精確度之虞，例如，許多資訊檢索的研究討論到「車子」時，也會將「輪胎」納入，但因為許多人在文章中看到車子，通常很少會想到「輪胎」，如此只會提高計算複雜度，對相似性卻不見得有很大的助益。因此聯想關係建立 ACN 時，聯想關係強烈者，才加入 ACN 中。一方面，一篇文章中有些概念原本就不是很重要，導出兩次要概念之間的中間概念，實質意義不大，因此的中間概念將限制介於兩主要概念之間的聯想過程者。

這種以聯想過程動態地導出中間概念以輔助文章相似性比對的方法，相對於只以主要概念作為比較基礎的「靜態比較」，而稱為「動態比較」（Dynamic Comparison）。

5.2.2 動態比較

在兩主要概念 A 和 D 之間可能有若干條聯想路徑皆可由 A 聯想到 D，那麼

應該取哪一條路徑上的概念為中間概念？以下將敘述「最佳聯想路徑」和「最佳聯想概念集合」。

5.2.2.1. 最佳聯想路徑

在 ACN 中，由某一概念 A 和另一概念 D 之間若存在若干路徑，人們往往會選擇最容易聯想到 D 的方式來推導出 D，換言之，是以最大貢獻法讓 D 獲得最大重要性的路徑來聯想。這個路徑係利用最大貢獻法取得，稱為 A 到 D 的「最佳聯想路徑」（Optimal Association Path，OAP）。在圖 5.4 中，A 到 D 的 OAP，表示成 OAP(A,D)=A→B→D。令最佳聯想路徑元素集合（Optimal Association Path Element Set, OAPES）={ E|E 為 OAP 之所有節點}，則 OAPES(A,D)={A,B,D}。

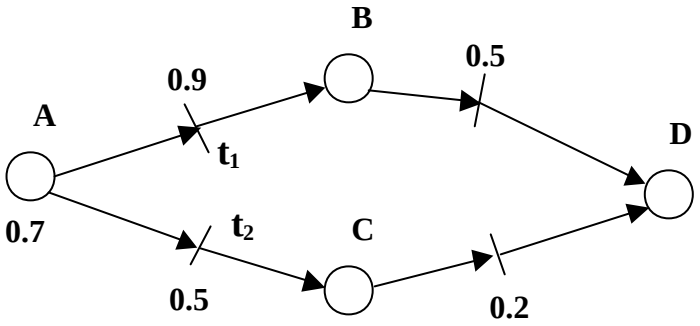


圖 5.4　A 與 D 之間之最佳聯想路徑為 ABD

5.2.2.2. 最佳聯想概念集合

主要概念和其中間概念都是思考文章意義時不可或缺的重要概念。令一篇文章 T 中之主要概念集合 CS={T_1,T_2,…T_n}，則令 T 的「最佳聯想概念集合」（Optimal Association Conceptual Set, OACS）$OACS(T)=CS\cup\{\underset{1\le i,\, j\le k}{U} OAPES(T_i, T_j)\}$。如圖 5.5 中，若文章 T 的 CS={A,D, E}，OACS(T)={A, B,D,E}。其中要加入 OACS 的概念

A 如果已經存在 OACS 中，則 A 將不加入 OACS，因此 OACS 中不會因為派屈網路的觸發序列（firing sequence）中存在無窮迴圈而收集重複的概念。

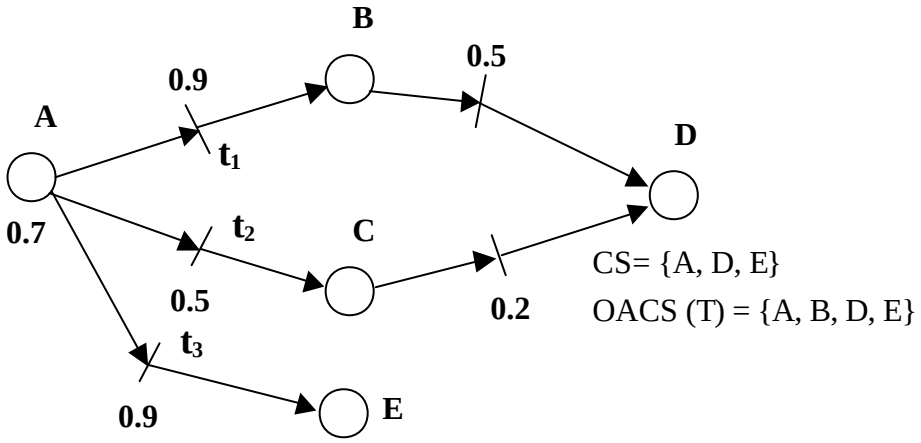


圖 5.5　文章 T 相符的 ACN 範例

動態比較的步驟如下：

1. 找出文章 T 和 TS 的主要概念，分別為$\{T_1 \ldots T_k\}$和$\{S_1 \ldots S_e\}$，這個部分可以依照資料的特性利用最大貢獻法或總和貢獻法得到。
2. OACS (T) = $\{T_1, T_2 \ldots T_k\}$，令 $X = \underset{1 \le i, j \le n}{Y} OAPES(T_i, T_j)$，令 $OACS = OACS(T) \cup X$，假設取得最後之 $OACS(T) = \{T_1 \ldots T_m\}$，$k \le m$。
3. 文章 S 亦以同樣方式處理，假設所得到的 $OACS(S) = \{S_1 \ldots S_n\}$，$e \le n$。
4. 以向量模式計算 T 和 S 之相似度：

$$sim(T,S) = \cos ine(T,S) = \frac{\sum_{i=1}^{m} |T_i| \times |T_{i,s}|}{\sqrt{\sum_{i=1}^{m} (|T_i|)^2} \times \sqrt{\sum_{i=1}^{m} (|T_{i,s}|)^2}}$$

其中，$|T_{i,s}|$為 T_i 在 S 中之重要性，若 $T_{i,s} \notin OACS(S)$中，則$|T_{i,s}|=0$。

由於我們可以事先定義出 ACN 中兩概念間的最佳聯想路徑，因此本比較法的時間複雜度只考慮向量模式即可，即 $O(n^2)$。

第6章　實驗結果與分析

6.1 實驗範圍及流程

本實驗以目前世界關切的 SARS（嚴重急性呼吸道症候群, Severe Acute Respiratory Syndrome）議題為主。我們將從各大電子新聞網站蒐集與 SARS 相關的新聞，並加入運動、電子科技和資訊科技等非關 SARS 的文章成為實驗資料。我們將利用這些資料逐一驗證第四章和第五章提出的檢索方法，以檢驗其效能及可行性。實驗的流程如圖 6.1 所示：

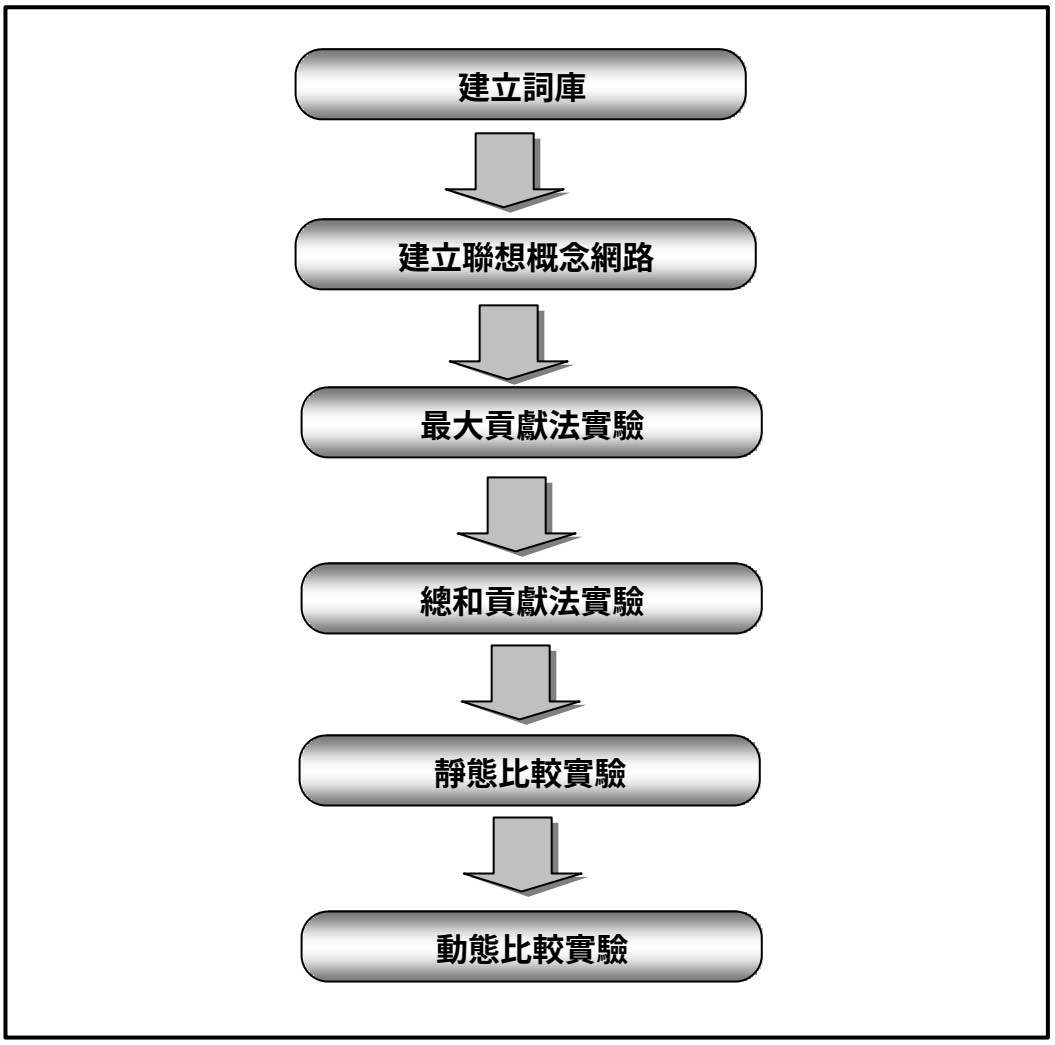


圖 6.1 本研究的實驗流程

6.2 建立詞庫

文章詞庫為文章檢索之基礎，其選取之準確度與適當與否，對於檢索效果有重大的影響。由於 SARS 領域為近幾個月來才發生的新議題，某些有效詞，例如，「嚴重急性呼吸道症候群」就不會出現在傳統詞庫。因此實驗將採用統計式斷詞法找出 SARS 的有效詞。統計式斷詞法係根據字詞出現在文章內的次數判斷該字詞是否為某個領域的有效詞，其中最常用的是 N-gram 斷詞技術，本論文採用之。

什麼是 N-gram？中文的文章是由一個個的句子組成，一個句子結束時都會加上標點符號，如分號、逗號、句號及驚嘆號等。而句子本身是由許多字組成，若不考慮它的語意，而將相鄰的任意兩個字組合起來成為一個「詞」，我們稱它為 Bigram。任意相鄰的三個字組成的「詞」就稱為 Trigram，N 個字的稱為 N-gram[8]。

如果用 SARS 相關的文章進行 N-gram 斷詞，則可以較一般文章擷取出相關 SARS 的有效詞。我們從東森新聞網站的「SARS 專區」[9]中選出 500 篇文章進行斷詞，文章範例如圖 6.2，每篇新聞內容大約是 150 字至 800 字之間不等。

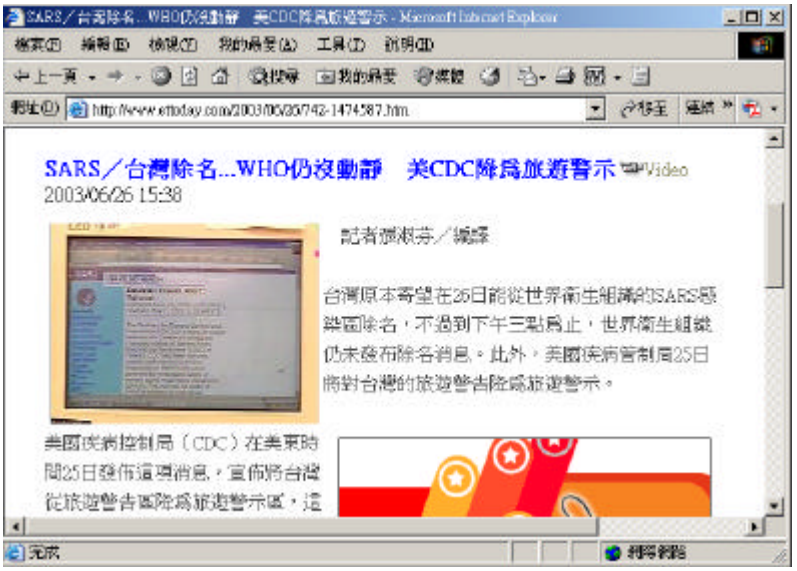


圖 6.2 東森新聞報的網頁範例

表 6.1 為對這 500 篇文章進行 bigram~10-gram 斷詞所得到 SARS 領域的部份有

效詞：

表 6.1 與 SARS 相關的部份有效詞

SARS	感染	衛生	WHO	烏爾巴尼	衛生局	長庚醫院
江秉誠	病毒	呼吸道	疫情	黏液病毒	CDC	加強消毒
通報	排除感染	侯彩鳳	防疫	溫家寶	感染病例	死亡病例
口罩	世界衛生組織	隔離	衛生官員	張上淳	勤太太	插管
章家敦	淘大花園	嚴重急性呼吸道症候群	肺炎	勤姓台商夫婦	支持性療法	隔離防護
防護面罩	可能病例	隔離通知書	檢疫	隔離治療	疾病管制局	冠狀病毒
檢體	發燒	肺間質病毒	病毒株	檢測	冠狀病毒基因組	PCR
疑似病例	防疫體系	和平醫院	弘武營區	心肺衰竭死亡	李龍騰	仁濟醫院
發燒	接觸史	抗氧化劑	淘大住戶	普勒托利亞	中國醫藥學院	江大雄
何大一	非典型肺炎	咳嗽	量耳溫	奈米機能口罩	流行病學	視訊追蹤管制系統

這些有效詞是近幾個月來 SARS 爆發後，大家耳熟能詳的字詞。如果要以人工去定義這些字詞，將會忽略或遺漏某些有效詞，例如，「烏爾巴尼」是發現 SARS 病毒的學者，因此「烏爾巴尼」在 SARS 中是相當重要的有效詞，卻很容易被一

般人忽略而未加到詞庫中。

此外，N-gram 中的 N 也是決定有效詞的重要因素。例如，N=9 將無法找出「嚴重急性呼吸道症候群」（共 10 個字）的有效詞，也將影響檢索正確性。

6.3 建立聯想概念網路

找到 SARS 的有效詞後，再利用這些有效詞之間的聯想關係，就可以建立 SARS 領域的 ACN。我們利用新聞中經常一起出現的有效詞很容易被人們聯想一起的特性，定義 SARS 領域由有效詞 A 聯想到 B 的程度 R_{SARS}，其意義是利用 A 出現的文章總數及 A 和 B 出現的文章總數決定 A 聯想到 B 的程度：

$$R_{SARS}(A \rightarrow B) = \frac{A\text{和}B\text{出現的文章總數}}{A\text{出現的文章總數}}$$

若 $R_{SARS} < \varepsilon$，則稱 A 無法聯想到 B，本實驗中定義 $\varepsilon = 0.3$。

本實驗利用 500 篇文章找出概念間的聯想關係，部份資料如表 6.2 所示。

表 6.2 本實驗中概念之間的聯想程度（部分）

觸發條件	結果概念	聯想程度	觸發條件	結果概念	聯想程度
SARS	感染	0.515	衛生	隔離	0.515
SARS	衛生	0.527	WHO	SARS	0.983
SARS	疫情	0.589	WHO	感染	0.633
SARS	防疫	0.34	WHO	衛生	0.883
SARS	隔離	0.446	WHO	疫情	0.7
感染	SARS	0.996	WHO	通報	0.3
感染	衛生	0.591	WHO	衛生署	0.517
感染	疫情	0.618	WHO	防疫	0.3
感染	通報	0.311	WHO	世界衛生組織	0.717
感染	防疫	0.335	WHO	傳染	0.35
感染	隔離	0.524	WHO	可能病例	0.3
感染	發燒	0.319	WHO	衛生署長	0.3
衛生	SARS	0.996	WHO	陳建仁	0.333

衛生	感染	0.577
衛生	衛生局	0.342
衛生	疫情	0.615
衛生	通報	0.338
衛生	衛生署	0.462

烏爾巴尼	SARS	1

由於出現「SARS**」的文章總數相當多，因此，可以從** SARS **聯想的概念數量亦不少，但部分聯想的程度不大。另外，和傳統詞庫中的一般概念相同，若干出現在文章的次數過於頻繁的字詞，例如，「你」、「我」或「他」，相對的其他有效詞就顯得不重要；至於「烏爾巴尼」出現的文章總數相當少，因此，當我們想到「烏爾巴尼」，也就容易聯想到與「烏爾巴尼」出現在相同文章的概念。**

同樣地，如果要人工建立這些聯想法則，將會耗費大量的時間，以本實驗大約 2457 **個概念而言，最多將考慮** $P_2^{2457} = 2457 \times 2456 = 6034392$ **個概念間的關係才能將** ACN **建置完成。自動化建立聯想法則雖然會犧牲聯想法則的些許正確性，卻可以大大提昇檢索的效率。加上這些聯想關係是由新聞得到的，相對於由個人主觀想法建置而成的** ACN **也會比較客觀。**

想要有效檢索 SARS **領域的文章就必須建置** SARS **領域的** ACN**。如果想同時檢索籃球領域的文章，就必須建置一個包含** SARS **領域和籃球領域的** ACN**。我們可以先找出大量與** SARS **或籃球相關的文章，並從中擷取有效詞，再重新計算這些有效詞的聯想關係，就可以得到** SARS **領域和籃球領域的** ACN**。也可以合併** SARS **領域和籃球領域的** ACN**，合併方法如表** 6.3**，設** A **和** B **為領域甲，乙的概念，**$R_{甲乙}(A \to B)$**為甲乙合併後** A **聯想到** B **的程度。其中** K **的定義如下：**

$$K = \frac{A\text{和}B\text{出現在甲的文章總數} + A\text{和}B\text{出現在乙的文章總數-A和B出現在甲和乙的文章總數}}{A\text{出現在甲的文章總數} + A\text{出現在乙的文章總數-A出現在甲和乙的文章總數}}$$

其中，K **考慮甲乙領域中有交集的文章，因此分子分母都會扣掉重複出現在甲和乙的部分。當甲乙領域其中一個領域由** A **聯想到** B **的程度大於** 0**，結合甲乙領域裡由** A **聯想到** B **的程度就為** K**。**

表 6.3　合併甲乙領域後 A 和 B 的聯想關係

	甲中 A,B 聯想程度= $R_{甲}(A \to B)>0$	甲中 A,B 聯想程度 $R_{甲}(A \to B)=0$
乙中 A,B 聯想程度 = $R_{乙}(A \to B)>0$	K	K
乙中 A,B 聯想程度 $R_{乙}(A \to B)=0$	K	$R_{甲乙}(A \to B)=0$

6.4 最大貢獻法檢索文章之實驗

本研究之 SARS ACN 係由「東森新聞報」網站的 500 篇文章所建置，為有效驗證本論文提出的檢索方法之可行性和效率，我們再由「WiseNews 慧科新聞網站」擷取出實驗之測試資料。「WiseNews 慧科新聞網站」目前收錄 300 多家由台灣，中國和香港等新聞媒體的新聞，除了新聞數量龐大，類別也相當廣泛。我們從中選出 SARS 相關新聞 400 篇，而電子、資訊科技、體育運動、影視戲劇四大類別的新聞資料 600 篇（合稱為非 SARS 文章），共 1000 篇新聞作為實驗對象。

最大貢獻排名的實驗開始先由檢索系統以亂數取出 1 篇文章 d_1，並利用最大貢獻法找出這些文章的主要概念，其中每個主要概念的重要性必須大於 0.3。如果「SARS」為該篇文章的主要概念，則稱該篇文章與 SARS 相關。以此法檢視文章，將可以得到每篇文章是否與 SARS 相關，答案為是與不是兩種，並檢查每篇之結果與事先由人工定義的文章歸類（SARS 與非 SARS 之類）相符，就知道檢索是否正確。可以算出召回率（Recall）和正確率（Precision）以判斷檢索的效能；兩者的定義如下：

$$召回率 = \frac{找到的SARS類別文章總數}{SARS類別文章在系統的總數}$$

$$精確率 = \frac{SARS類別文章被正確判定的數目 + 非SARS類別文章被正確判定的數目}{檢索過的文章總數}$$

之後再以亂數取出與前面不同的文章 d_2，並重新計算這兩篇文章的召回率和正確率；如此不斷地重複步驟直到取完 1000 篇文章為止，將可得到圖 6.3 的召回率和圖 6.4 的準確率。準確率最終可達 95.50%。

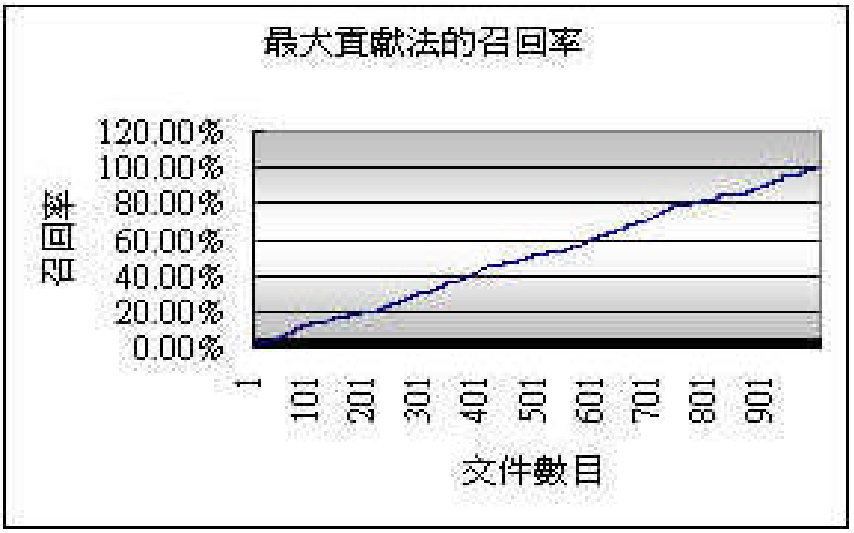


圖 6.3 利用最大貢獻法找出 SARS 文章的召回率

圖 6.4 利用最大貢獻法找出 SARS 文章的準確率

圖 6.3 的曲線越是線性，則代表 SARS 和非 SARS 相關文章的隨機分配越均勻。

在曲線任一點，代表處理到該對應之文件數量時之召回率，分母在每一點都是相同的，分子則會隨著處理結果而異，其最終值必為 100%。

6.5 總和貢獻法檢索文章之實驗

總和貢獻排名的實驗方法與最大貢獻法相同，其準確率可達 93.00%，實驗結果如下：

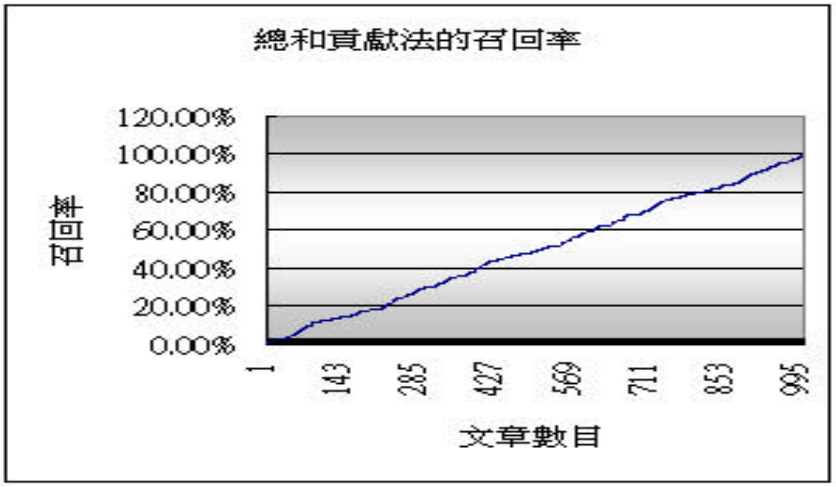


圖 6.5　利用總和貢獻法找出 SARS 文章的召回率

圖 6.6　利用總和貢獻法找出 SARS 文章的正確率

由於本實驗開始就有幾筆檢索錯誤的情形，因此圖 6.6 中的準確率有大幅下降的情形，不過由於檢索的資料量較大，這些誤差也會漸漸修正回來。

6.6 最大貢獻法之文章相似性比對實驗

本實驗以一篇與 SARS 相關的文章 A，與系統內的 1000 篇文章進行比對，我們以亂數來排出比對的順序，若文章的相似程度大於 0.5，則我們稱為相似的文章。其中 A 內容如下：

SARS「嚴重急性呼吸道症候群」持續蔓延，衛生署已宣佈列為第四類法定傳染病，對此國民黨嘉義縣黨部昨日發佈「健康概念」文宣，提醒鄉親關心自己與家人的健康。相關資料可上嘉義縣黨部網站：http://www.kmtcy.org.tw/或請參考行政院衛生署疾病管制局網站 http://www.cdc.gov.t 或致電 0800024582 免付費電話諮詢。近日來一直引起民眾關注的 SARS「嚴重急性呼吸道症候群」，嘉義縣縣黨部提醒大家要注意自己的健康，唯有先做好保健預防工作才能將危險降到最低。相關的預防措施如：1.勤洗手 2.保持環境衛生及空氣流通 3.避免到人群聚集或空氣不流通的地方 4.避免不必要的探病 5.均衡飲食 6.適量休息及運動縣黨部主委翁重鈞並指示各鄉鎮市黨部利用機會廣為宣導 SARS 預防之道，並配合衛生、教育單位關心地方民眾，特別是小朋友的健康保健。並提醒鄉親，在此期間民眾非必要盡量勿前往中國大陸、香港及越南等地捆 C，如需出國儘量勿到人口密集通風不良之場所，如果從該地或鄰近地區回國後，出現異常發燒或呼吸道症狀，應即就醫，並告知醫護人員到過的地區，作為參考。

本篇文章宣導 SARS 的預防之道，因此包含許多 SARS 領域的有效字，但是文章中尚未提及 SARS 以外的事情，我們認為該文章僅隸屬於 SARS 類別，因此與本篇相似的文章也必然是 SARS 類別的文章。我們以這樣的特性檢視 1000 篇文章中，哪一些是隸屬於 SARS 類別的文章，並統計之，再一一檢視這些文章

是否歸屬於 SARS，就可以得到相似性比對的召回率和準確率，圖 6.7 和圖 6.8 分別為最大貢獻法取得主要概念之召回率和準確率。其中準確率為 93.09%。

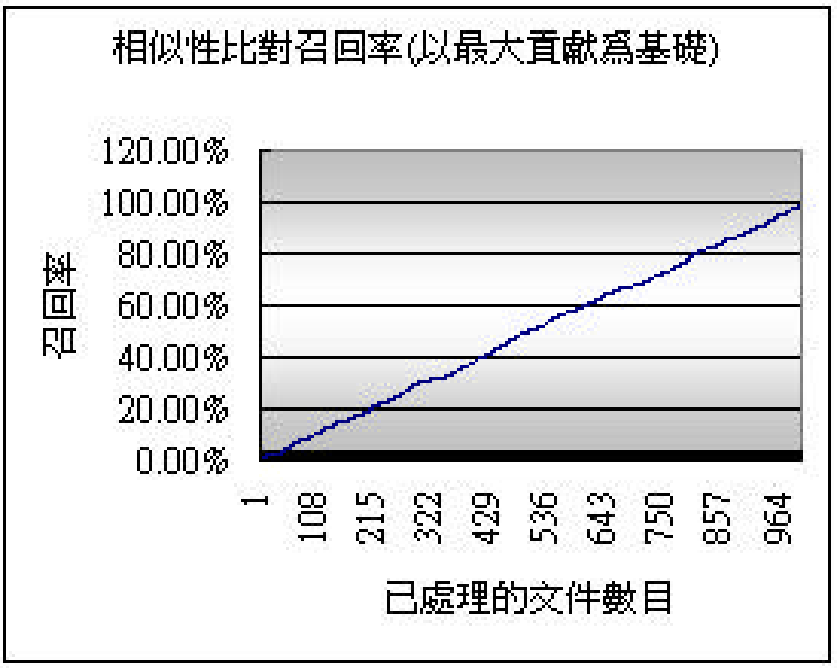


圖 6.7 以最大貢獻為基礎的相似性比對召回率

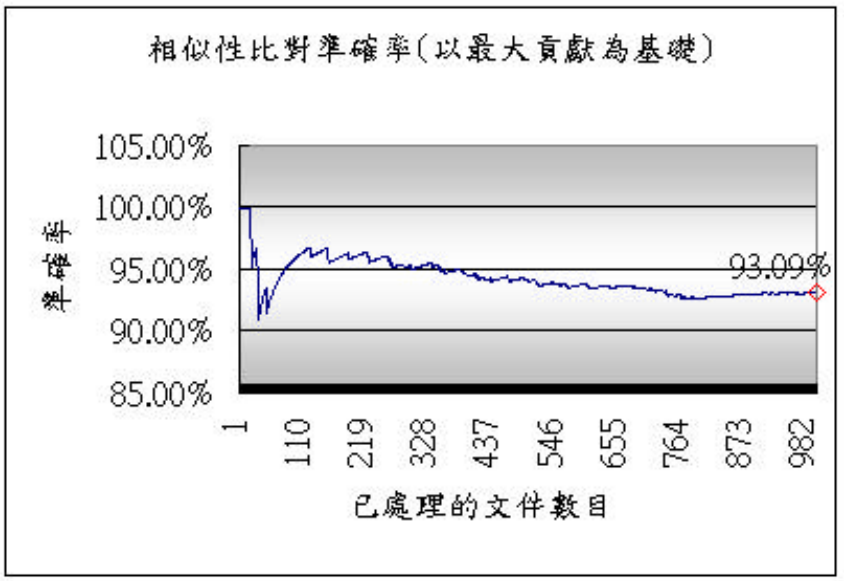


圖 6.8 以最大貢獻為基礎的相似性比對準確率

文章的相似性比對是比較兩篇文章 A 和 B 的相似度，A 和 B 可以由使用者任意選擇。然而，若兩篇均非 SARS 文章，可能來自於不同類別（運動、電子科

技和資訊科技），若不相同對結果毫無意義，因此，每一組別中必須至少有一篇為 SARS 類。共隨機抽出 1000 組均不重複的組別以供驗證。兩篇均屬於 SARS 類或兩者均不屬於 SARS 類為"相似"，一篇是一篇不是則為"不相似"。圖 6.9 為以最大貢獻為基礎之相似性比對的實驗，最終之準確度為 84.4%。

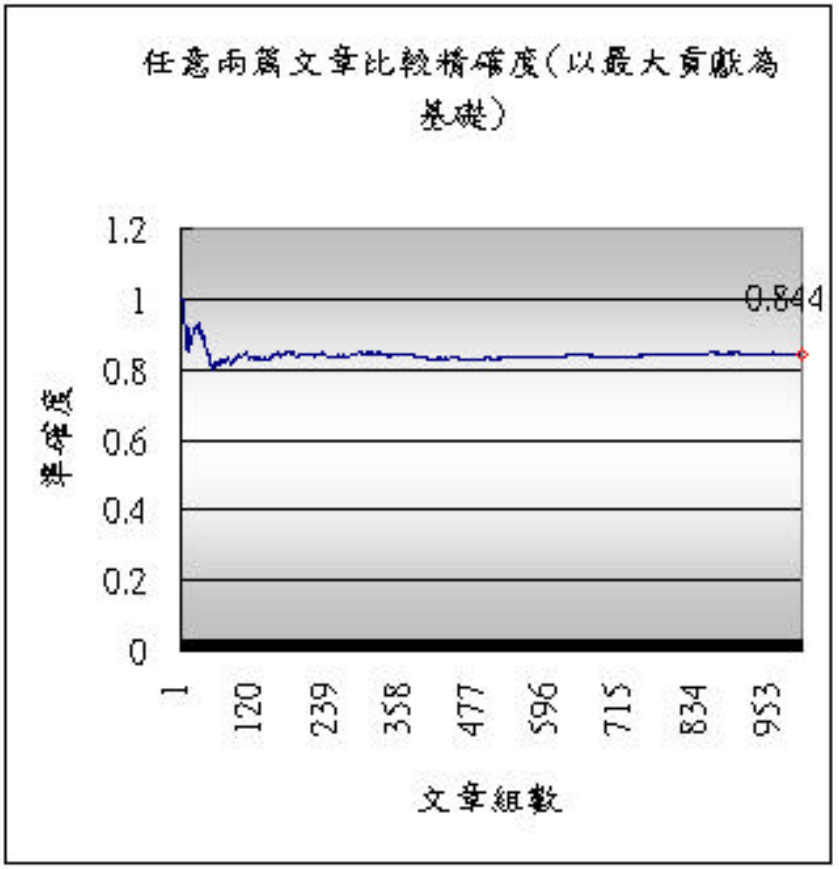


圖 6.9 以最大貢獻為基礎之任意兩篇文章的相似性比對準確率

我們可以發現利用一文章找出其他文章的精確度（93.09%）與任意兩篇文章的相似性比對得到的精確度（84.4%）有些微差距。這是因為前者選擇的是經過挑選的 SARS 領域文章，可以很清楚地與其他文章界定為相似或不相似。但後者是隨機挑出，在 SARS 領域裡的重要性就不一定高，可能會參雜其他領域的知識，與某些文章就不容易界定相似或不相似，準確率自然也會受影響。這也得到一項結論，如果要利用文章 A 查詢相同領域的其他文章，A 最好與這個領域非常相關。

6.7 總和貢獻法之文章的相似性比對實驗

以總和貢獻為基礎的相似性比對與以最大貢獻為基礎的相似性比對之實驗環境相同，也是以第六節之同一篇文章為對象，其 1000 篇文章進行相似性比對，其召回率與準確率見圖 6.10 和圖 6.11，準確率為 93.00%。

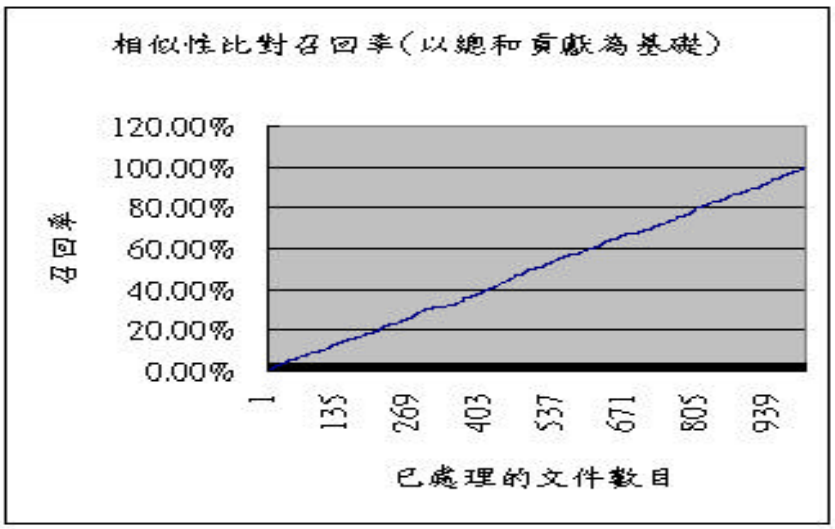


圖 6.10　以總和貢獻為基礎的相似性比對召回率

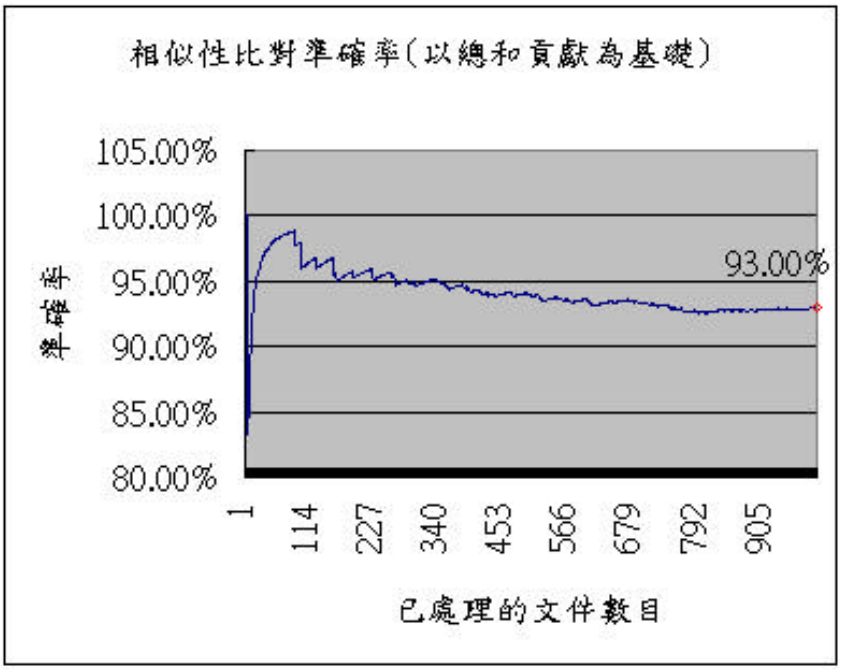


圖 6.11　以總和貢獻為基礎的相似性比對準確率

圖 6.12 所示係任意兩篇文章為一組，共 1000 組，而以總和貢獻為基礎之相

似性比對準確率，平均準確率為 93.70%。。

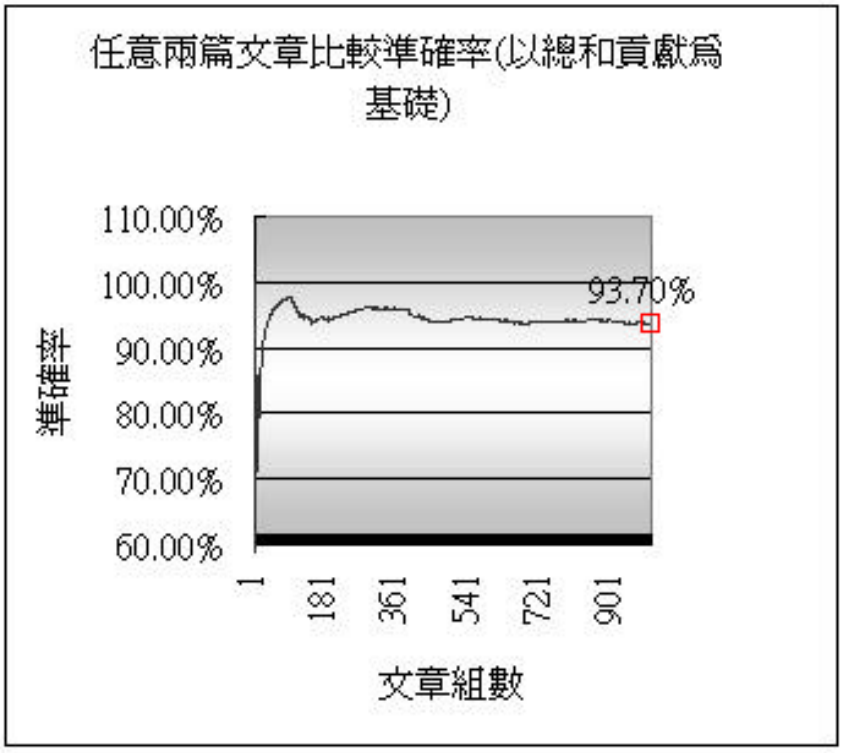


圖 6.12 **以總和貢獻為基礎之任意兩篇文章的相似性比對準確率**

6.8 動態比較之實驗

動態比較與靜態比較的實驗環境亦相同，而得到圖 6.13 **的召回率、圖** 6.14 **的準確率（**93.90%**）和圖** 6.15 **任意兩篇文章的相似性比對準確率（**85.10%**）。**

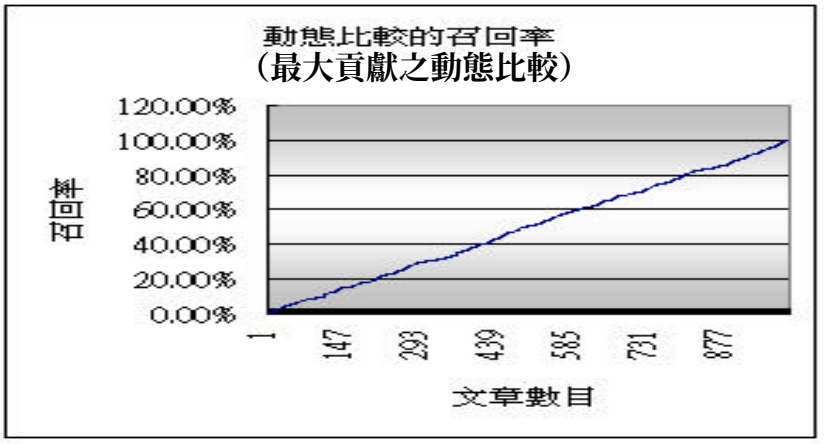


圖 6.13 **動態比較的召回率**

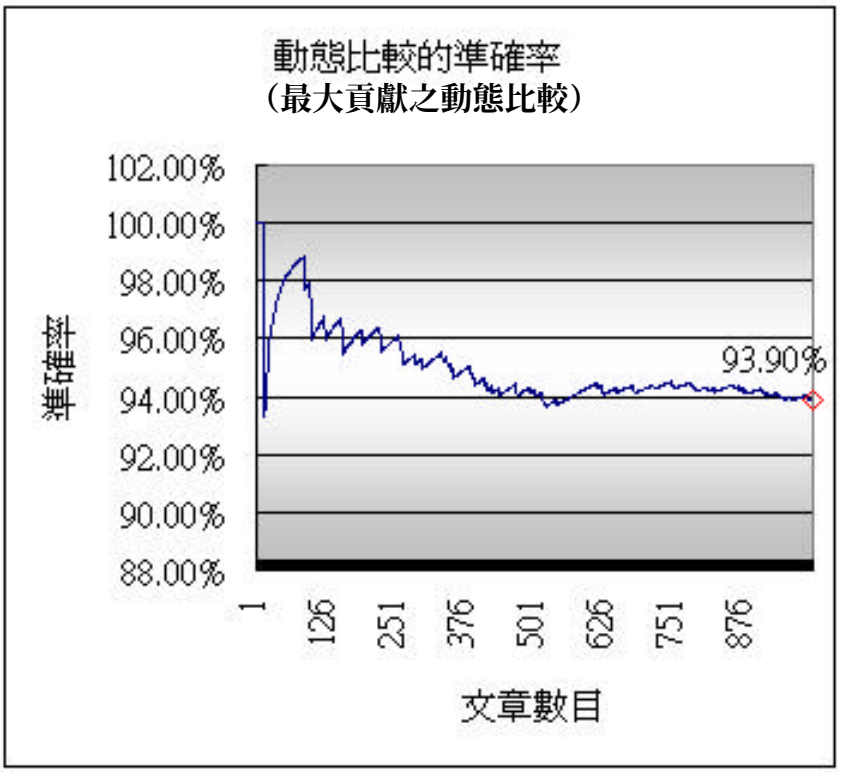


圖 6.14 動態比較的準確率

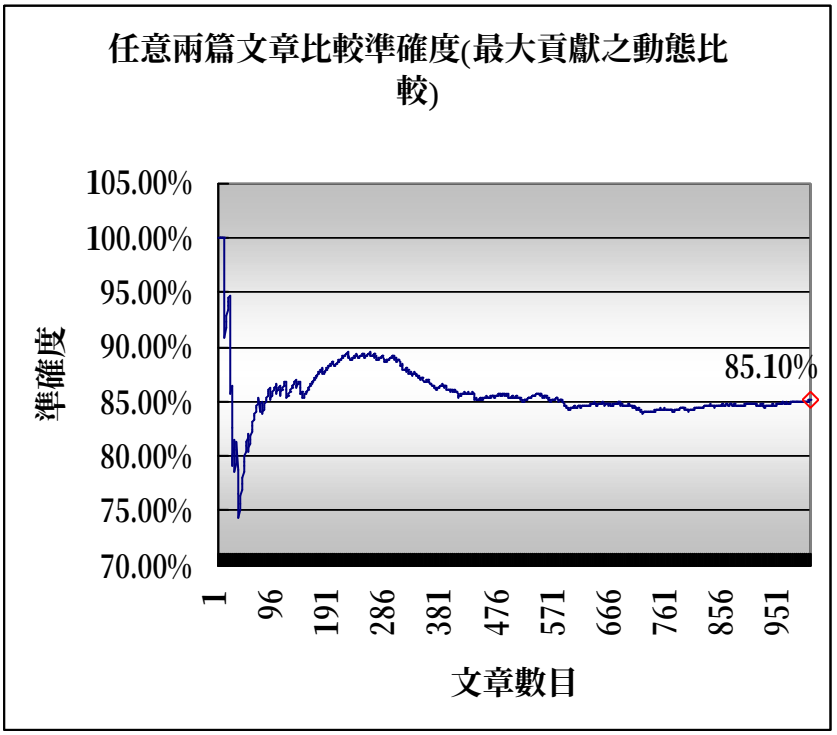


圖 6.15 動態比較的任意兩篇文章比較準確率

6.9 與向量空間法的比較

我們將最大貢獻法與總和貢獻法與向量空間法比較，結果如圖 6.16 所示。而利用最大貢獻為基礎、總和貢獻為基礎、最大貢獻法之動態比較和向量空間法的相似性比對法的比較圖則如圖 6.17 和圖 6.18 所示。

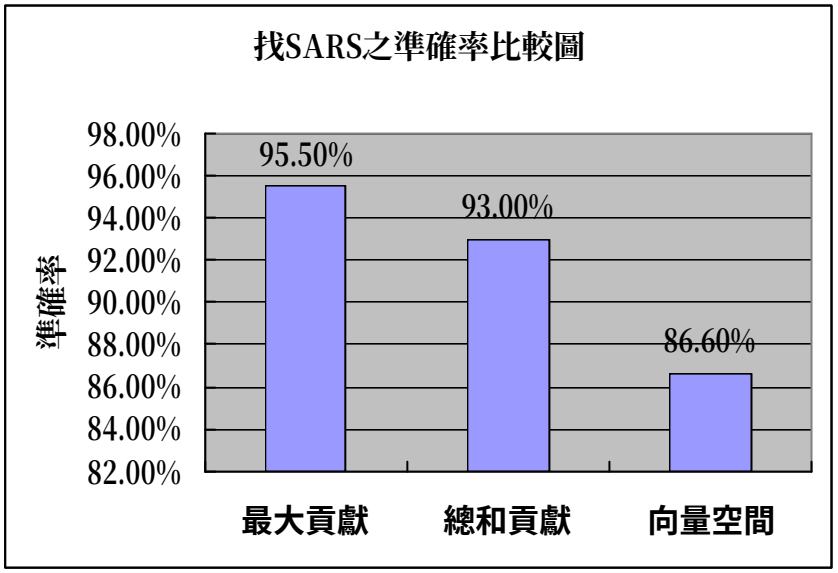


圖 6.16 以最大貢獻法、總和貢獻法和向量空間法檢索文章的準確率比較圖

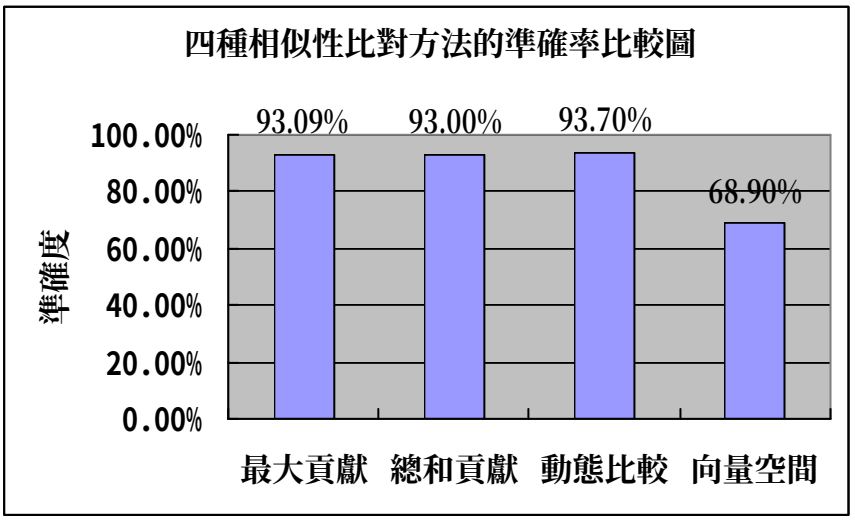


圖 6.17 四種方法相似性比對的準確率比較圖

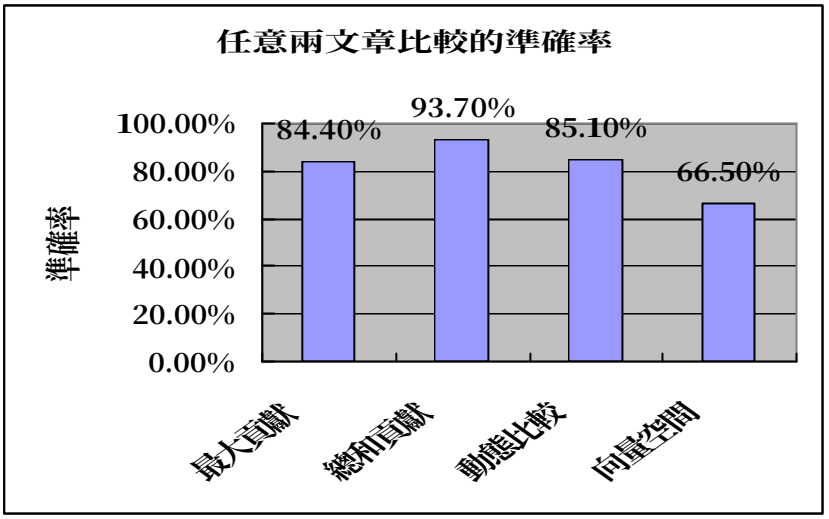


圖 6.18 四種方法任意兩篇文章比對的準確率比較圖

6.10 以聖經中的「創世紀」領域作實驗

為了驗證本論文所提出的方法也可以適用在其他領域，我們選擇了聖經的「創世紀」作為檢索的主題。聖經中的文章經常以隱喻的手法表達，因此如果本論文提出的方法可以有效地提昇檢索效率，就能夠證明聯想關係確實可以找出文章中的隱含概念，並以之檢索文章。

本次實驗以聖經的「創世紀」中的章節建構「創世紀」ACN，實驗中檢索文章和比對文章相似性的方法與 SARS 領域的實驗相同，不同的是以 100 組文章為檢索對象，其中 20 篇為與創世紀相關的文章，其餘則不相關。圖 6.19 為最大貢獻法、總和貢獻法和向量空間法檢索文章的準確率比較圖。而利用最大貢獻為基礎、總和貢獻為基礎、最大貢獻法之動態比較和向量空間法的相似性比對法的比較圖則如圖 6.20 和圖 6.21 所示。

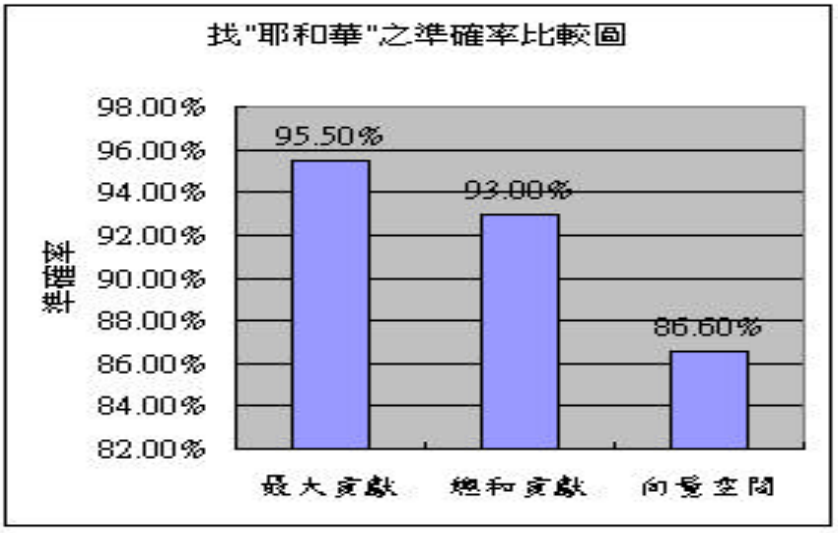


圖 6.19 三種方法檢索「創世紀」文章的準確率比較圖

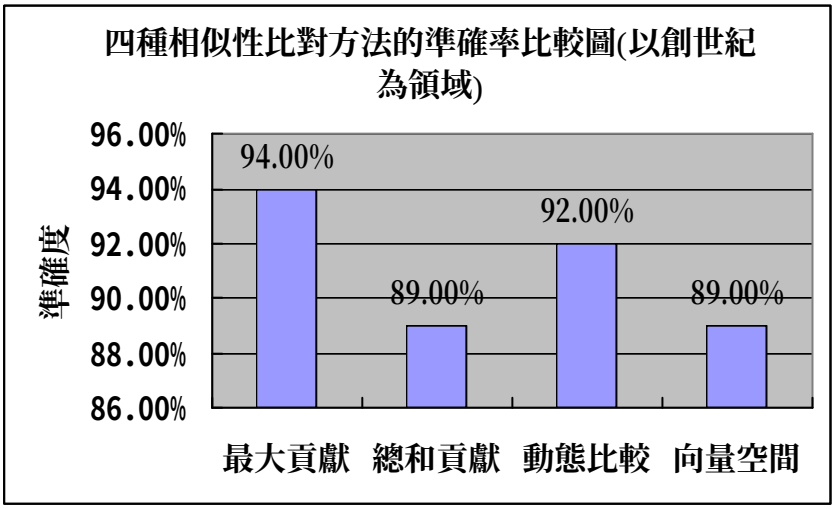


圖 6.20 四種方法相似性比對的準確率比較圖

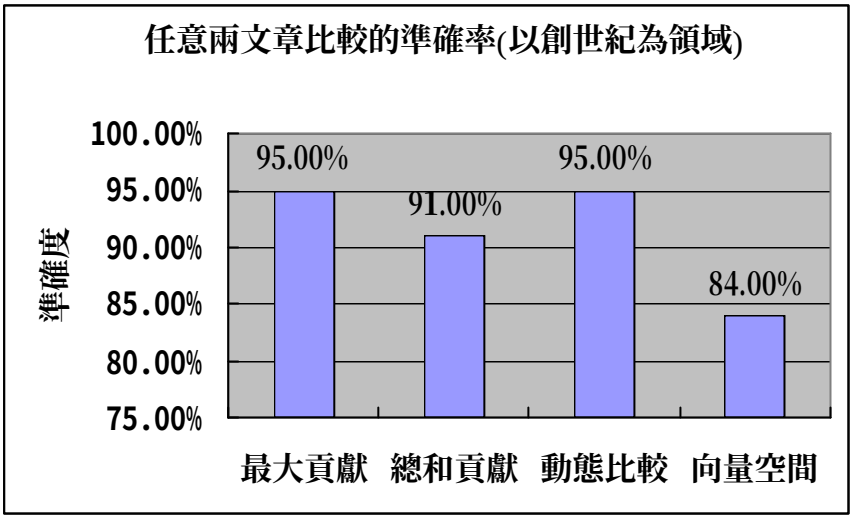


圖 6.21 四種方法任意兩篇文章比對的準確率比較圖(以創世紀為領域)

實驗證明如果以「創世紀」為領域，本論文提出的方法之準確率都佳於向量空間法。

6.11實驗結論

最大貢獻法和總和貢獻法之準確率高於向量空間法，證明最大貢獻法和總和貢獻法確實能大幅提昇向量空間法檢索文章的效率。

至於文章相似性比對的部分，以最大貢獻、總和貢獻和動態比較為基礎之相似性比對法也高於向量空間法之相似性比對法，也證明了以聯想為基礎的比對方法可以大幅提昇檢索效率。其中，我們可以發現總和貢獻法或以總和貢獻法為基礎之文章相似性比對法之準確率都可以高達 93.00%以上，但其時間複雜度也很高。而最大貢獻法為基礎之靜態比較和以最大貢獻法為基礎之動態比較準確率幾乎相同，這是因為我們已經將 ACN 中的每個概念之間的聯想關係最大化，因此任何概念間的聯想就不須經過中間概念，所以最大貢獻法為基礎之靜態比較和以最大貢獻法為基礎之動態的準確率會很相近。而總和貢獻法求得之重要性係由可以聯想到該概念的其他概念重要性加總而來，其中也包含了中間概念的重要性，因此沒有以總和貢獻法為基礎之動態比較。

實驗的結果符合預期，當使用者輸入與 SARS 越相關的文章，則從文章找出關於 SARS 的主要概念，並與其他文章從 SARS 的角度比較相似性的準確率也會大幅提昇。可以證實聯想關係確實可以有效地模擬人類的思考方式，進而利用聯想的特性，改善了資訊檢索的效率。

第7章　結論與未來研究方向

本論文以模糊派屈網路表示概念間的聯想關係，以此建構聯想概念網路，並根據聯想概念網路建置完整程度和文章的特性，例如，文章所屬類別，分別找出文章內的主要概念，包括隱含的概念，改善傳統資訊檢索只從文章的關鍵字找出文章特徵的缺點。此外，我們也利用聯想概念的特性，依兩文章的主要概念之重要性越相似則文章也會越相似的特性提出靜態比較，和以中間概念的相似程度提出動態比較，俾判斷文章的相似程度，模擬了人類的思考方式，經實驗結果，以聯想關係為基礎來找出文章的主要概念和比對文章相似性的準確率都遠大於向量空間法，確實有效地提昇檢索的準確性。

另外，本論文的方法對於概念 A 和 B 之間關係的強弱並不考慮它們在文章中的距離。其實人們對於發生時間越近的事件記憶會越清晰，也容易將發生時間相近的事件聯想一起。因此一般人寫作時，為了使讀者更容易理解所傳達的內容，相關的概念編寫在文章的位置通常越相近，也就是說，兩概念如果越相近，可能越相關。其中，文章述及的事件往往會結合這些相關的概念。例如，”John loves Mary.’就是描述 John 和 Mary 存有 love 關係的事件。也因此，如果可以找出文章包含的事件，就可以正確地表示概念間的關係，將有效提升檢索的效率。以文章包含的事件為主之文章比對法也是未來研究的目標。據此，我們就可以利用事件來找出兩概念間的關係，並利用事件間可能的聯想關係來組合事件，建構出文章完整的事件結構圖，並以基礎事件此比較文章之間的相似程度。我們可以將事件拆解成更細微的事件，也可以組合許多事件成為更大的事件。從人類的思考而言，事件越大，更能全面性地得到事件中概念間的關係，也可以正確地描繪出事件的時間序列，其中，概念與概念之間的聯想程度也會隨著時間而有所不同，所以事件與事件間的聯想程度也會隨著時間而有所不同，因此未來我們也將

研究組合事件的表示法和相似性，以增進檢索正確率。

此外，由董強和董振東所提出的知網是將概念和概念之間以種種不同的關係進行連結，例如，同義關係、反義關係、組合關係等結合，形成了網狀的知識系統，便於進一步自然語言處理，每個概念之間可以再細分成其他概念，直到不易於再分割意義的最小單位義原(Sememe)為止。因此，未來我們也擬利用知網蓋善 ACN，以提昇檢索的效率。

參考文獻

1. 呂芳懌, 顏義樺, 施政瑋（2002），"以網頁特性為基礎之重要性排名,"第六屆 2002 年資訊管理學術暨警政資訊實務研討會，pp489-495。
2. 呂芳懌, 顏義樺, 施政瑋（2001），"探查疑似不法資訊網站之分散式網路巡邏系統,"第三屆 2001 年網際空間：資訊、法律與社會，pp97-108。
3. 呂芳懌, 顏義樺,（2003），"以聯想法則概念網路為基礎之文章概念探索及相似性比對,"第十四屆資管年會（ICIM' 2003 Conference，Taiwan）。
4. 呂芳懌, 顏義樺,（2003），"以事件聯想為基礎之文章相似性比對,"2003 年資訊技術應用與發展研討會。
5. 林曉芳（民 90）。知識表徵與概念學習之研究--以路徑蒐尋網路分析為評量工具。教育與心理研究，24 期，pp229-262。
6. 楊英魁、孫宗瀛、鄭魁香、林建德、蔣旭堂，模糊控制理論與技術，全華圖書股份有限公司，民國 85 年。
7. 東森新聞報 "美伊／戰爭開打 8 小時　德法總理譴責美國不合法行徑", http://www.ettoday.com/2003/03/20/334-1428392.htm, 3 月 20 日, 2003.
8. 黃文鴻,呂芳懌,"文字資料庫在犯罪偵查之應用---竊盜犯案手法之比對,"第十屆國際資訊管理學術研討會論文集,台北,1999 年 6 月。
9. 東森新聞報，"SARS 風暴"，http://www.ettoday.com，2003 年 5 月。
10. WiseNews 慧科新聞網站，"WiseNews"，http://twpwisenews.wisers.net，2003 年 5 月。
11. J. Cho, H. Garcia-Molina, L. Page "Efficient Crawling Through URL Ordering", http://www-db.standford.edu/~cho/crawler-paper, July 2001.
12. J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5): 604-632, November 1999.

13. S. Brin, L. Page. “The anatomy of a large-scale hypertextual Web search engine”. In *Proceedings of the Seventh International World Wide Web Conference*, pages 107-117,April 1998.

14. L. Page, S. Brin, R. Motwani, T. Winograd. “The PageRank Citation Ranking: Bringing Order to the Web”, http://www-db.stanford.edu/~backrub/pageranksub.ps, July 2001.

15. M. Murat, K. Kanzaki, K. Uchimoto, Q. Ma and H. Isahar, “Meaning sort – three examples: dictionary construction, tagged corpus construction and information presentation system,” in *Computational Linguistics and Intelligent Text Processing*, Springer, 2001.

16. G. Miller, R. Beckwith, C. Fellbaun, D. Gross and K. Miller, “Introduction to WordNet: An online lexical database” in *International Journal of lexicography*, vol. 3, no.4 235-312, 1990.

17. B.Y. Ricardo and R.N. Berthier, “Modern Information retrieval” Addison-Wesley, 1999.

18. Christiane Fellbaum Wordnet, MIT Press, 1999.

19. Shyi-Ming Chen; Jeng-Yih Wang; Document retrieval using knowledge-based fuzzy information retrieval techniques, Systems, Man and Cybernetics, IEEE Transactions on , Volume: 25 Issue: 5 , May 1995.

20. J. F. Sowa (2000), “Knowledge representation: Logical, Philosophical, and Computational Foundations”, BROOKS/COLE, ISBN 0 534-94965-7.