南　華　大　學

資訊管理研究所碩士論文

領域本體論為基之網頁知識擷取機制設計

研 究 生：王亮超

指導教授：王昌斌　博士

中華民國　九十五　年 七 月

領域本體論為基之網頁知識擷取機制設計

Design a web documents extraction mechanism based on domain ontology

研 究 生：王亮超 Student : Liang-Chao Wang

指導教授：王昌斌 Advisor : Dr. Chin-Bin Wang .

南 華 大 學

資訊管理學系

碩 士 論 文

A Thesis

Submitted to Department of Information Management
College of Management
Nan-Hua University
in partial Fulfillment of the Requirements
for the Degree of
Master of Business Administrator
in
Information Management
June 2006
Chaiyi Taiwan, Republic of China.

中華民國九十五年七月

南　華　大　學

資訊管理學系

碩 士 學 位 論 文

領域本體論為基之網頁知識擷取機制設計

研究生：王惠超

經考試合格特此證明

口試委員：

指導教授：

系主任(所長)：資訊管理學系 系主任 吳光剛

口試日期：中華民國　95 年　6 月　30 日

口試合格證明

誌　　謝

碩班學習尾聲即將響起，此刻的心情洋溢著喜悅與不捨。回首這兩年來，亮超要感謝的人很多，首先最要感謝的是恩師王昌斌教授，在論文研究期間不辭辛勞地指導與提攜，以及時常地關心與鼓勵支持亮超，非常由衷地感謝，雖隻字片語，意更甚於言表，老師，真的很謝謝您！再者，十分感謝口試委員成大製工所陳裕民教授，以及所上鍾國貴教授，對於本論文悉心指正與精闢的建議，使得學生在研究寫作與思考邏輯上更加成長增進，論文能得以愈趨完善，非常地謝謝您們！此外，謝謝吳光閔所長、蔡德謙老師和陳宗義老師，幫助釐清思考困境與提供寶貴的意見。

這段期間感謝南華同儕們彼此的鼓勵與照應，尤其是嘉明、鎰聰在靜瑜、慶堂論文上的幫忙，以及實驗室成員育銘、昇衛、育弘在研究上的協助，與你們一同成長的相渡時光非常愉快。

最後要感謝我最親愛與敬愛的家人感謝你們的勉勵與支持，讓我能勇往前進，追求人生目標的完成，謝謝你們！

王亮超　謹識

于　南華大學資管所

九十五年　七月

領域本體論為基之網頁知識擷取機制設計

學生：王亮超　　指導教授：王昌斌 博士

南　華　大　學　資訊管理學系碩士班

摘　　要

文字探勘(Text mining)結合資料探勘、自然語言處理與資訊檢索技術，使大量不具結構的文字資訊能經由電腦自動的分析歸納，目前主要的應用有自動分類、自動摘要、文件檢索及知識管理。其中一個重要的主題「文件摘要(text summarization)」，經由電腦自動依照其文件內容，選取出重要的句子組合形成文件摘要，從而減少學習者閱讀時間並且增加效率。本研究著重於小學生數學學習障礙的問題領域，透過領域本體論(domain ontology)的應用，幫助我們從搜尋引擎所搜尋到的文章進行文件的擷取重要句子進一步組成摘要。由於領域本體論能夠描述特定知識領域內相關的概念與關係，利用此種特性，本研究提出一種以本體論為基的文件段落擷取方法，以期能有效的協助網路學習者在網路文件上獲取正確且快速的相關知識。本研究根據學習者的語意，以多個關鍵字詞來評斷文章段落中，哪些是學習者欲知的知識內容。而本實驗結果也發現，以多個關鍵字詞所擷取的文章段落比單一字詞擷取更能符合學習者的語

意。多個關鍵字詞所擷取出來的文章段落，能有效避免單一關鍵字詞本身擁有一字多義的情形發生。藉由領域本體論與自然語言處理的結合，把使用者所輸入的問題進行分析，找出文章中符合語意的段落回應給使用者，促進知識的分享與再利用。

關鍵詞：文字探勘、文件摘要、知識擷取、本體論

Design of Web knowledge extraction mechanism based on domain ontology

Student：Liang-Chao Wang Advisors：Dr. Chin-Bin Wang .

Department of Information Management
The M.B.A. Program
Nan-Hua University

ABSTRACT

Text mining combines data mining, information retrieval and NLP (Nature Language Processing) technologies. Text mining main application in auto-classification, auto-summarization and knowledge management, etc. Automatically text summarization is a problem which is using computer technologies to give an user quick and useful document's knowledge.Our research emphasis on elementary school students who have Mathmatic Learning Disabilities. In our research, Ontology can help us to define a word concept and relationship. We propose a multi-keyword method to judge a document's paragraph whether fit user's semantic. When a user input a query, we should according to user's semantic to find the suitable document's paragraph, and representation it to the users. It can be promoted knowledge reusability and sharability. And our experiments results finds that using multi-keywords extract document's paragraph can suitable user's query.

keywords：Text Mining, Text Summarization, Knowledge Extraction, Ontology

目　　錄

表 目 錄

圖 目 錄

第一章 緒論

近年來資訊技術蓬勃發展，由於網際網路盛行拉近了人與人之間的距離，所有的訊息都可以傳遞到世界上其它任何角落，也由於資訊數位化的因素，造成大量的資訊充斥在隨手可得的網路世界中，許多電子文件的服務也與日俱增，要如何尋找、收集資料，然後整理、探勘為有用的資訊便顯得一門重要的學問，要有效的管理這些資料也因此顯得格外重要。

全球資訊網(World Wide Web, WWW)目前儼然成為網際網路資訊的重要來源。它所提供的資訊包羅萬象，資訊量增加的速度也越來越快。資訊公開普及化有正面的意義，但數量過多且來源分散的資訊，卻未必是好事。缺乏一致的管理，特定主題的相關網頁，散佈在各處，不知道有多少，也不知道如何去尋找相關資訊；此外，人類記憶與處理能力上亦有限制，不能無限量的瀏覽或儲存、分析資訊。因此急需一個搜尋引擎工具，以協助使用者瀏覽資訊，避免使用者迷失在網際網路的空間中。

在現今資訊爆炸的時代，每天都有新的資訊產生。為了從這些大量的資訊中，準確的獲取有用的資訊，文章的自動摘要處理變的越來越重要。通過閱讀文章摘要而不是全文能極大的加速資訊過濾速度，幫助人們了解概況或確定是否應該詳讀原文。這一技術是快速準確獲

取資訊的一個有用工具，在現代人們求於快速簡潔的獲取知識，它的市場需求相當廣泛。

第一節　研究背景

由於網際網路的興起，網路上的資料也越來越多，資料也越來越五花八門。對於使用者而言，常常在這浩瀚的網際網路上迷失方向。而搜尋引擎的出現，是網路使用者的一大福音。因為資訊的尋找問題在網頁數量越來越多時將會更嚴重，故很多以尋找資訊為目的資訊工具不斷開發。

搜尋引擎工具的發展，協助了使用者在網路上做資料的索引，它儘可能的找出使用者所要找尋的資料，大大節省了使用者在搜尋資料的時間。搜尋引擎工具提供的資訊服務如下：

一、關鍵字查詢服務(Keyword Query Service)：鍵入欲尋找的關鍵字，搜尋引擎比對索引或網頁內文，並回傳符合的網頁。

二、目錄服務(Dictionary Service)：建立主題架構，手動把網頁放入對應的主題中，使用者能瀏覽較高層的資訊結構，以找到所需的資訊。

由於網頁數量過多，用關鍵字查詢的回傳網頁的資訊量龐大，超過人類處理能力的上限，故需其他機制協助在大量網頁中，快速尋找所需資訊；此外若使用者與服務提供者在用語與表達方式上不相同，

亦不易找到真正所需的資訊。

因此網際網路與搜尋引擎的產生，造就了今天知識經濟時代的來臨。在一切講求效率的環境下，知識即是競爭力。企業最大的資產是員工以及員工腦中的知識，快速的獲取及善用知識成為企業的競爭優勢。企業利用數位學習科技協助員工職能教育訓練，提升員工附加價值以提升公司整體的競爭力。另外，網際網路的蓬勃發展及資訊科技的應用普及加速了數位學習的發展。坊間數位學習系統的數位教材是知識的獲取來源，知識的分享來自於網際網路。

第二節　研究動機與目的

網際網路的世界無遠弗屆，隱含著浩瀚的知識，經過整理後即可為一個龐大的資料庫。由於網際網路的資料量過於龐大且資料零亂，所以需要花費大量人力與時間進行整理。為了要減少使用者閱讀時間以及網路搜尋的時間，如何有效率並且快速的擷取文件內容是一個重要的議題。因此，在隱含的網頁知識中，如何根據學習者的語意並設計符合語意且有效率自動擷取網頁內容極為重要。

傳統教學方式受限於時間與空間，隨著科技及網際網路時代的來臨，數位學習為知識傳播與擴散的方法，與傳統的教學方式主要差異在於其結合資訊科技與網際網路以彌補傳統教學模式的不足與缺陷，而傳統紙本的教學內容在經過數位化的轉換程序後，將更易於編

修、彙整、互通及整合，大幅提升其再用性(Reusability)與可分享(Sharability)。由於數位學習平台可解決傳統教學所面臨的困難，而網路上又充滿著豐富的資料，如何以自動化的方式針對學習者想了解的問題，在網路上搜尋答案進而建構成一個有組織有系統的知識庫用於支援學習者進行學習的活動是一項大的挑戰。

本研究主要目的在於設計如何根據學習者的語意，找出文章裡面符合學習者語意的段落。依據程序設計「自動擷取符合語意段落演算法」，及「符合語意段落結果的儲存模式」，最後設計「符合語意段落儲存區」，使得知識能夠有效率的擷取、儲存與分享。由於本體論能夠用來表示描述與說明特定領域下的概念與知識，藉以分享與表達該知識領域下存在的事件與彼此間的關係，所以本研究運用本體論技術來推斷學習者語意的相似詞。

而在人工智慧與知識管理領域，本體論的應用逐漸增加，相關語言及開發軟體的支援，另外伴隨著第二代網路，也就是語意網路(Semantic Web)以及web services的興起，利用建構本體論來描述特定領域下的知識與服務等研究也漸漸熱絡[14]。

本研究希望利用本體論來協助語意辨識清楚的程度，因為一般詞庫並無法分辨相似詞。例如:學習障礙與學習阻礙。在一般人的認知中，這兩者意思是相近的，但是在電腦裡面卻辨識成兩個不一樣的詞

彙，所以本體論的建置可以清楚的分析學習者詢問的真正語意，辨別文章詞彙與詢問句是否同為相似詞，提高擷取的精準度。

第三節　問題分析

壹、網際網路自動摘要之知識擷取技術設計不易

一、網際網路未使用標準的詞彙或者一致的描述風格

科技日新月異，在非結構化的網路環境上語文也越多元化。時下一般口語用法大多因人而異，如何判斷並正確的擷取描述同一事物的詞彙並不太容易。

二、精確的分析網際網路的內容並選出重要的部份是非常困難的

如何根據使用者查詢問題的語意，真正擷取到文章相關的知識內容並且回應給使用者是非常困難。

貳、儲存區維護不易

新增的知識與儲存區的知識維護困難

隨著系統的使用與發展，網路上擷取的資料會演變成龐大的知識庫。如何在新增的知識與既有的知識內容儲存區兩者裡面找出相似並且重複的部分予以刪除並不是一件簡單的事情。

第四節　研究步驟

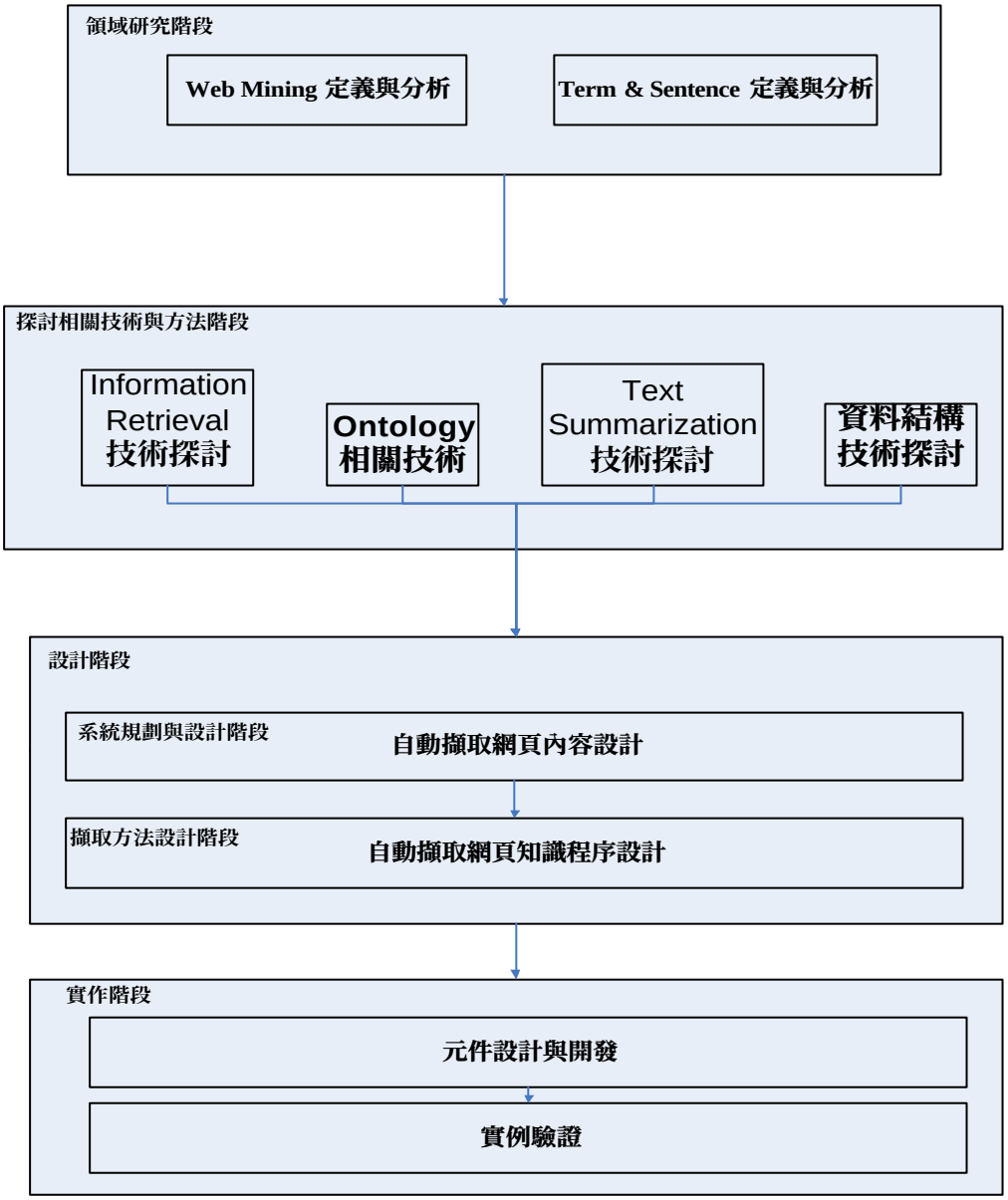


圖- 1 研究步驟

我們主要目的是發展出網路知識擷取機制，參照圖-1 研究步驟的階段依序為領域研究階段、探討相關技術與方法階段、設計階段和實作階段。

一、領域研究階段：清楚地了解網路探勘(Web Mining)**、詞彙**(Term)**和句子**(Sentence)**的定義與分析。**

二、探討相關技術與方法階段：探討在網路知識擷取機制方面目前的相關技術與方法。其中包含資訊檢索(Information Retrieval)技術探討、本體論(Ontology)相關技術、文件摘要(Text Summarization)技術探討和資料結構技術探討四大方面。

三、設計階段：在設計階段分為兩大部分，第一部分是系統規劃與設計階段，另一部分是擷取方法設計階段。在系統規劃與設計階段方面，我們根據學習者的語意並予以解析出關鍵字，最後到搜尋引擎尋找相關網頁，並把網頁去除掉標籤檔，保留文章內文部份。在擷取方法設計階段方面，根據學習者所解析出來的關鍵字與搜尋到的網頁內文進行正確率比較，把符合學習者語意部分的網頁文字擷取出來後並回覆給學習者，達到學習者自我學習的目的。

四、實作階段：在實作階段，設計元件與開發並且根據學習者語意搜尋實例，最後進行實例驗證。

第五節　研究預期貢獻

根據學習者詢問的問題，解析問題擷取關鍵字，並提出以多個關鍵詞彙來判斷文章段落是否符合學習者的語意。

第六節　論文架構

在第一章緒論裡，我們提到論文的研究背景、動機、目的、問題

分析、研究步驟和預期研究貢獻；在接下來的論文內容中，第二章將介紹相關的研究及本篇研究論文中所運用到的關鍵技術；第三章會詳細描述本篇研究論文的作法及整個系統的流程；第四章介紹系統的使用及說明並且介紹實驗測試資料、方法以及實驗後的結果；最後在第五章中會為本篇研究論文做一個整體性的結論及未來發展的描述。

第二章 文獻探討

第一節 本體論

壹、本體論定義

本體論，在英文的解釋為Ontology，最初使用在哲學領域中，表示「存在的、有意識的實體或主體」之意，後來延伸應用在人工智慧領域方面，表示用來描述與說明特定領域下的概念與知識，藉以分享與表達該知識領域下存在的事件與彼此間的關係；在[18]的描述中，其定義如下：

“An Ontology is a specification of a conceptualization”，Ontology 是某種概念上清楚的說明[18]。當我們使用Ontology 來描述特定領域下的知識，可把Ontology視為是概念(Concept、Object 或是Class)、屬性(Attribute、Property、Slot 或是Role)、實例(Instance)與關係(Relation)這些元素的組合，以下分別說明這些元素：

一、概念：Concept 就是以多個底層物件所組成的範圍，也就是由多個字彙(Vocabulary)所組成的集合，這個集合能夠作為一個概念性的描述，描述出主題的基本範圍，透過這個集合能讓系統了解到定義Concept 所代表的意思。

二、屬性：屬性可以當作是該物件的一個描述，描述該物件的特性或特徵，當我們使用Ontology 表達某個特定知識領域時，

Concept 就是其中的子集合，這些集合我們可以將其看成是物件，在物件間會有各種關係存在，而且每一物件本身也會有各種屬性存在。實際上，物件擁有屬性所建構出整個 Ontology 的資料架構，在應用上將提供更為多元及有用的訊息，我們不但可得知Concept 與其它的Concept 之間的關係，從單一的Concept 也可得知Concept 本身的重要性，另外，若是要將一個Ontology 作充分有效的利用，屬性對於訊息的多樣化是最有幫助的利器。

三、關係：若只使用物件與其屬性來清楚的描述特定知識領域內的概念與結構時候，在知識領域的表達上來說，還不足以提供物件之間相關的細節訊息，所以當建構出整個Ontology 的架構之後，除了清楚的描述出物件與物件屬性之外，還可以為這些物件定義其彼此間所有的關係。由於目前自動建構 Ontology 的技術尚未成熟，所以大都還是透過領域專家以人工方式提供一個有系統的知識領域架構，這個架構可以用來描述整個領域中的抽象結構與關係，提供相關應用系統上的使用與共享。

四、實例：實例可以用來更清楚的表達上層的Concept，因此實例與上層Concept 通常會存在著某種關係，並繼承某些上層

Concept 的屬性，當然，實例也可以擁有自己更細微的屬性來表示與其它實例的差異之處。這裡存在一個問題是，在建構 Ontology 的過程中，對於一個項目，要如何定義它是Concept 或是Instance 呢？根據[19]的建議：「Instance 是用來更清楚的表達Concept」，所以通常在Ontology 架構中最底層的部份來定義Instance。

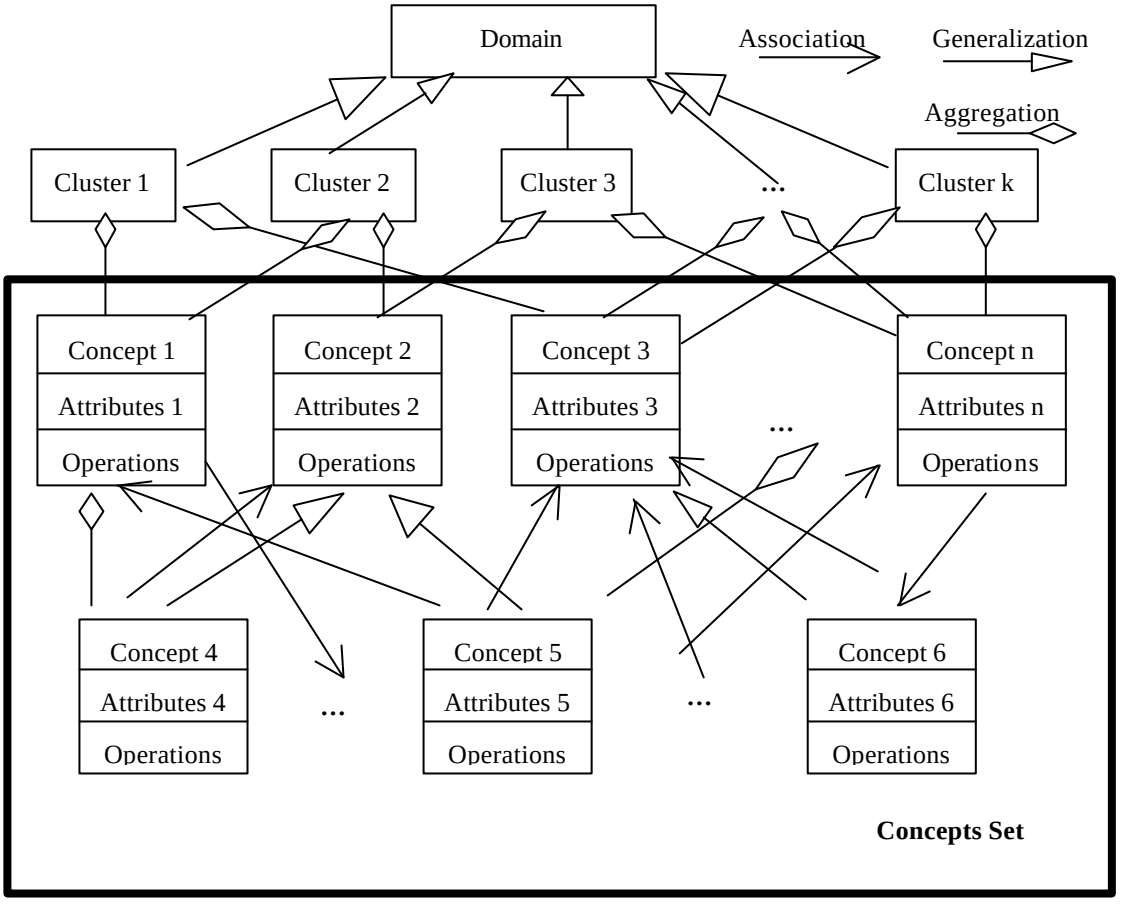


圖- 2 物件導向之本體論架構

資料來源：廖嘉欣 民 91 年

以下我們舉一例子說明 Concept class

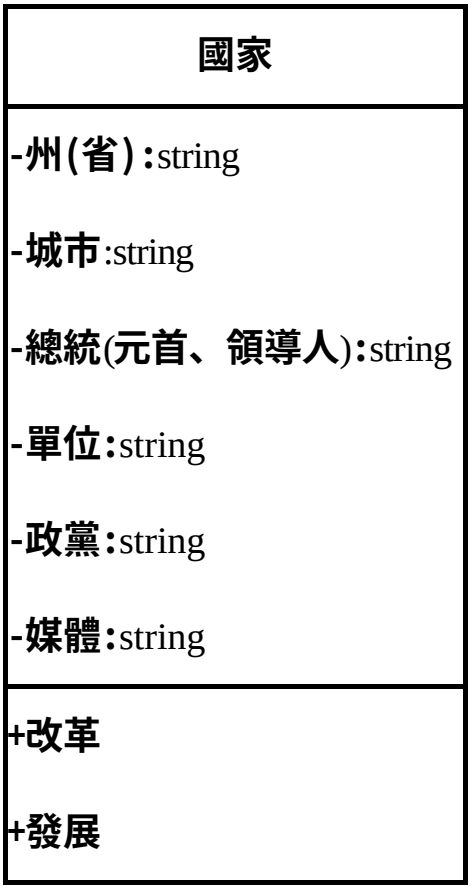


圖- 3 **本體論之中的概念**

資料來源：廖嘉欣 民 91 年

參見圖-3 **所示，**class **的名稱是「國家」，**attributes **包含「州(省)」、「城市」、「總統(元首、領導人) 」、「單位」、 「政黨」和「媒體」；**Operations **包含「改革」和「發展」。**

我們可以使用 Concept **的定義來創造領域本體論中個別的實例。舉例來說，我們創造一個別的實例「美國**(U.S.A) **」來描述一個「國家」型態。**

美國(美方,美利堅共和國)
-州(省):string=加州，德州 -城市:string=紐約，華盛頓，舊金山，芝加哥，費城 -總統:string=布希 -單位:string=白宮，五角大廈，聯邦調查局，國務院 -政黨:string=共和黨，民主黨 -媒體:string=CNN，華盛頓郵報
+改革 +發展

圖- 4 本體論之中描述特定型態的實例

資料來源：廖嘉欣 民 91 年

參見圖-4 所示，「美國」可稱為「美方」或者「美利堅共和國」。每一個屬性包含一些重要性來呈現 Concept 的特性。例如「政黨」包含「共和黨」和「民主黨」，「媒體」包含「CNN」和「華盛頓郵報」等等。

參見圖-5，我們另外再舉一例子說明領域本體論架構

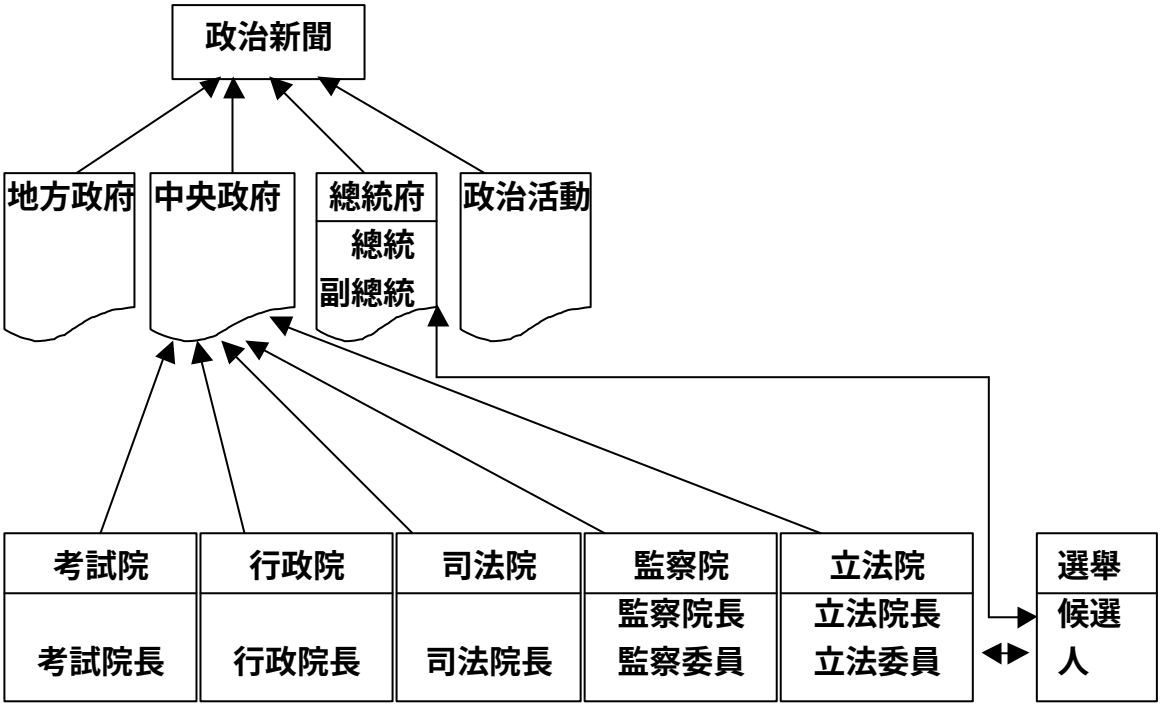


圖- 5 政治新聞本體論架構

資料來源：鐘明強 民 93 年

參見圖-5 表示的是一個政治新聞的本體論架構，「政治新聞」是這個本體論中最上層的概念(Concept)，在政治新聞中，考試院、行政院、司法院、監察院和立法院等訊息都是屬於政治新聞的一種，所以與政治新聞存在著一種「is a」的關係(relation)，都是屬於政治新聞下所衍伸出的概念(Concept)；而在總統府這個概念(Concept)中，範圍表示其屬性(Attribute)，在這裡我們利用屬性來描述其所屬之概念(Concept)下的某些特徵與特性，因此，在其它概念(Concept)與實例(Instance)中，也都會有各自的屬性來表達該 Concept 的結構與性質；在屬於政治新聞下的中央政府概念(Concept)中，分別由行政院、司法

院、考試院、監察院和立法院等五個中央行政機構所組成，因此這五院與中央政府存在著一種「part of 」的關係，而且各自有其屬性來表達各機構內的職務。如上述所示，當我們利用本體論，便可以簡單的來針對這些知識領域的分類來定義與描述出該特殊領域的專業術語、項目與彼此間的關係，再利用以 XML 為基礎，而且人與電腦皆可理解的本體論語言來描述上述的架構圖，進而分享、再利用或擴張相同資料。

貳、本體論的應用

通常建置本體論的原因，可能有下面這些因素[19]：

一、讓人類或電腦軟體代理人之間彼此分享資訊的知識

二、讓領域知識可重覆被利用

三、讓領域知識可清楚的被描述

四、在現有的知識中，分割出新的領域知識

五、分析領域知識

因此，使用本體論來定義某個領域中的一些基本概念及它們之間的關連，其主要的用意是為了讓電腦更容易閱讀這些知識，讓某特定領域的知識與資訊能夠以人與電腦皆可理解的架構來描述與呈現，並清楚的說明抽象資料概念間的關係，使得這樣的資訊能夠配合電腦系統以支援知識的分享及重複利用。因此在本體論內所定義的元素皆是

可以充份代表此特定領域內的相關知識，目前在建構本體論除了是經由該領域下的領域專家(domain expert)來制定；也有越來越多相關研究提出自動化或半自動化建構本體論的方式與機制 [9][11][13]。

另外，本體論在許多資訊應用上已經成為一個常被使用的技術，例如使用在代理人系統[4]、知識管理系統、電子商務等系統，藉以產生以自然語言為基的整合智慧型資訊系統，以及在網路上提供以語意為基礎的服務[17]，並已經漸漸成為在知識管理上做語意處理的一種關鍵技術，或是在資料交換時能夠精確的對應一些意義相同的詞彙與符號，倚靠由概念性模組所構成的本體論來精確的定義不同符號或詞彙的意義。

另外，在下一代網路，也就是語意性網路(Semantic Web)中，也使用本體論做為其中一項極重要的技術，希望利用本體論架構來表達其語意性，協助軟體代理人(agent)在語意網路上搜尋最適合的資訊[19]。

第二節　自然語言處理技術

壹、斷詞處理

由於中英文體系的不同，造成中、英文在字詞的表現方面，有明顯的不同，因而造成兩者在斷詞處理上的方式迥異。而在資訊檢索（Information Retrieval）領域中，字詞處理技術希望藉由分析文件，

取出能代表文件的關鍵字詞（Keyword）或特徵值（Feature）。

由於中文字詞表現並沒有像英文字詞有明顯的分隔符號，因此造成了中文文件在辨詞上的困難。此外，中文語言學中一個有意義的詞，通常由數個字連接而成，造成了電腦在判斷詞彙上的困難。因此，以下針對各種斷詞法作簡單介紹[12]：

一、詞庫斷詞法：本法主要是藉由事先建置好的詞庫，對於所要分析的文件，逐字逐行的比對，並判斷出其所包含之詞彙。本法優點在運算快速，缺點是人工詞庫永遠比不上快速的社會變動，因此常常無法有效分辨出新的詞彙，例如：囧rz、南迴鐵路搞軌案等等。

二、N-Gram 選詞法：針對大量的語料庫，對於文件內容進行分割，所分割的字詞出現次數若高於門檻值，則認定可能為一個獨立詞彙。依照所取出的字數長短，分為2-Gram (Bi-Gram)、3-Gram（Tri-Gram），依此類推至N-Gram。優點是可不受詞庫限制，判斷文件所包含的詞彙；例如:台北市市長;依據2-Gram取出:台北、北市、市市、市長;依據3-Gram取出台北市、北市市、市市長。而根據[1]研究顯示，中文名詞以2-Gram(雙連詞)佔所有名詞中的80%以上，又名詞與動詞佔所有詞性的絕大多數，所以研究中以動詞與名詞

為主要詞庫。而四連詞以上者大部分為成語居多。

缺點是斷出詞彙易呈現與語料庫相依(Corpus dependent)的特性。

例如：電腦科學:依據2-Gram取出:電腦、腦科、科學。但是腦科一般泛指醫院部門，所以容易出現與語料庫相依的缺點。

三、混合斷詞法：混合式斷詞主要是結合詞庫式斷詞及N-Gram 選詞法優點，首先利用詞庫式斷詞判斷文件詞彙，再利用N-Gram選詞法判斷相關新詞彙。然而本法仍須藉由維護詞庫及蒐集語料庫來維持品質。由於本法兼具詞庫式斷詞及N-Gram選詞法優點，故本研究採用此法作為斷詞處理。

貳、資訊檢索

傳統的資訊檢索大致上可分為三種，字典方式(Dictionary Approach)、文法、法則方式(Linguistic Approach)和統計方式(Statistical Approach)。

一、字典方式：利用事先訂定的詞彙字典來對文章進行資訊擷取。

二、文法、法則方式：建立完整的文法知識庫來擷取文章資訊。

三、統計方式：統計方式利用文字的數字資訊來擷取文章資訊，例如詞彙所出現的頻率、詞彙與詞彙出現的關係程度等

第三節　網路文件摘要

壹、 詞彙權重計算

TF(Term Frequency)概念與IDF(Inverse Document Frequency)由[21]所提出來的概念。提出該架構的理由，是因為每個文件中的詞佔整篇文件的重要性，其實是不太相同的。可能「筆記電腦」在資訊類和科技業文章出現次數較多，但是在體育類或者是植物類文章出現的次數較少，更重要的是就算出現體育類或植物類的文章出現，各篇的重要性也不太相同。因此TF和IDF的概念就由此產生了，這兩者的結合，就可以權衡詞的顯著值(重要性)。其概念包含如下：

一、TF(Term Frequency)：計算關鍵詞在某文件的出現頻率

$$tf_{ij} = \frac{n_j}{n_{all}} \qquad \text{(公式1)}$$

關鍵詞 j 代表在文件 i 出現的頻率，其中

n_j:i 文件 i 出現的次數

n_{all}:表示文件 i 所有具意義的總詞頻

例如：「筆記電腦」在單一文章中出現5次，但是文章裡面的總詞頻為50個，所以「筆記電腦」的TF值為0.1。

二、Inverse Document Frequency(IDF)：計算一篇文章關鍵詞彙權重值，更進一步推估所有文章關鍵詞彙權重值。

$$\mathrm{IDF_j} = \log_2 \frac{N}{df_j} \qquad \text{(公式2)}$$

單字 j 代表在所有文件裡內的權重值

N:代表所有文件的總數

df_j :代表單字 j 有出現過的文章總數

例如:例如文章總篇數 100 篇,其中出現電腦的篇數為 20 篇,所以 IDF 值為 $IDF_{\text{筆記電腦}} = \log_2 {}^{100}\!/_{20} = 2.32$

當這兩者相乘(TF×IDF)之後，即代表修正過後的關鍵詞 T_j 在文件 D_i 的加權(weight)。如下式所述:

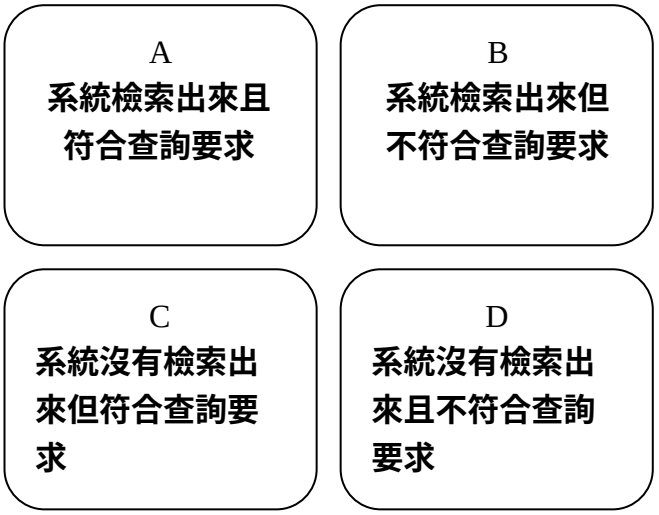
$$w_{ij} = \frac{n_j}{n_{all}} \times \log_2 \frac{N}{df_j} \qquad \text{(公式3)}$$

所以上述例子：「筆記電腦」頻率(TF=0.1)和權重(IDF=2.32)，所以「筆記電腦」在所有文章裡面的權重值為 0.232。

這樣作法的意義是說，通常衡量重要性，皆是以該詞彙在單一文件內出現的頻率作為決定性因子(TF 的涵義)，但若該詞彙同時出現在多篇文字內，相對而言該詞出現比出現在少數文件內較不具價值(IDF 的涵義)，所以利用 TF 和 IDF 相乘的結果就可算出詞的重要的性。在利用 TF-IDF 組成的摘要上，首先要加總出現在同一句的關鍵詞 TF-IDF 值；其次，依各句的 TF-IDF 值大小遞減排序。

貳、正確率評估

對資訊檢索系統而言，正確率也是一個很重要的評估要點，正確率的評估有兩種[8]，一種是回現率(recall ratio)，另一種是準確率(precision ratio)。這兩種評估法參見圖-6所示



A+B+C+D=資料庫所有文件

圖- 6 資訊檢索的正確率評估

資料來源：邱立豐　民 91 年

一、回現率(recall ratio)：系統所擷取到的文件數目佔查詢問題相關的比例，高的回現率代表資料庫中與查詢字的有關文件大部分已被檢索出來，其公式如下：

$$recall = \frac{\left|\{\text{relevant}\} \cap \{\text{retrieved }\}\right|}{\left|\{\text{relevant}\}\right|} = \frac{A}{A+D}$$ **公式(**4**)**

二、精準率(precision ratio)**：原則上，精準度作法與回現率差不多，差異點在於以系統檢索的文件總數作為評估基準；高的精準率代表檢索出來的文件大部分與查詢字有關，其公式如下：**

$$precision = \frac{\left|\{\text{relevant}\} \cap \{\text{retrieved}\}\right|}{\left|\{\text{retrieved}\}\right|} = \frac{A}{A+B}$$ **公式(**5**)**

另外，透過 F-measure **的作法融合回現率與精準率兩種評估準則。其公式如下**

$$F-measure = \frac{2P \times R}{P+R}$$ **公式(**6**)**

P **表示精準率，***R* **表示回現率**

以下我們舉出一例子說明，請參照圖 7

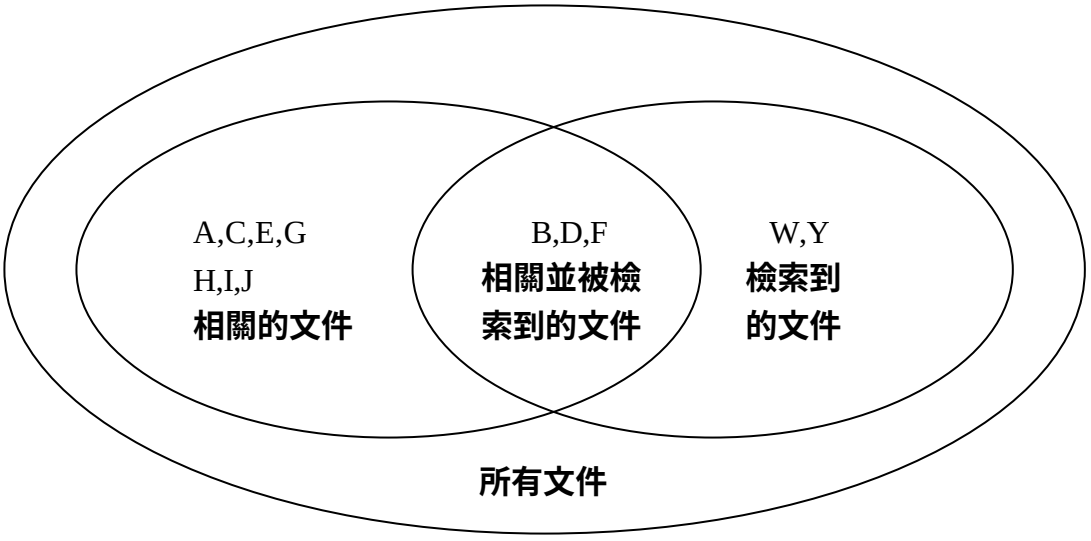


圖- 7 **文件檢索**

{ relevant } ={A,B,C,D,E,F,G,H,I,J}=10

{ retrieved } ={B,D,F,W,Y}=5

{relevant∩retrieved}={B,D,F}=3

招回率(Recall ratio)=3/5=60%

精準率(Precision ratio)=3/10=30%

$$F-measure=\frac{2P\times R}{P+R}=0.4$$

參、文章自動摘要

自動摘要的目地在於將重要句子組成一篇「言簡意賅」的文章，輔助使用者能短時間內得知文章的涵義，固摘要不僅僅能節省使用者時間，更能幫助使用者快速判讀文件內容，進一步搜尋其感興趣的內容。為了將重要句子組合成摘要一篇「摘要」，需要評估句子的重要性。[3]**提出了下列四種方法，作為句子重要性的評估準則**

一、關鍵詞法：關鍵詞法認為字詞的重要性與在文章內出現的詞頻及出現的文章篇數據有相關性；其中，字詞的權重與詞頻成正比，而與文章出現次數成反比。這也是一般人最常使用的方式。

二、位置法：一篇文章最重要的部份大多位於文章的首句與末句。

三、線索法：線索法假設文句中皆具有一些正面、反面的辭彙，能用來評估句子的重要與否，正面字：如 great , significant **等**;

負面字如 hardly 被用以判斷句子的重要性。

四、標題法：一篇文章的標題往往選取與主題相關的字詞所組合而成；因此，出現在標題的字詞要給予較高的權重值。至於標題詞彙的權重值的多寡，大多與前面三種方法互相參照，然後在主觀給定權重。

對於自動摘要的形成步驟，[2]分成六個步驟。參照圖-8 所示：擷取網路超文件、超文件分析作業、關鍵字萃取、句子權重計算、重要句子選取和自動產生摘要六個步驟；不過大致上，自動摘要可分為斷詞、句子權重計算、重要句子選取並鏈結三個主要部份。同時，在計算句子的權重方面，主要又可分為 TF-IDF 及相似度兩種計算方式，而依照權重計算方式的不同，句子鏈結的方式也有所不同。

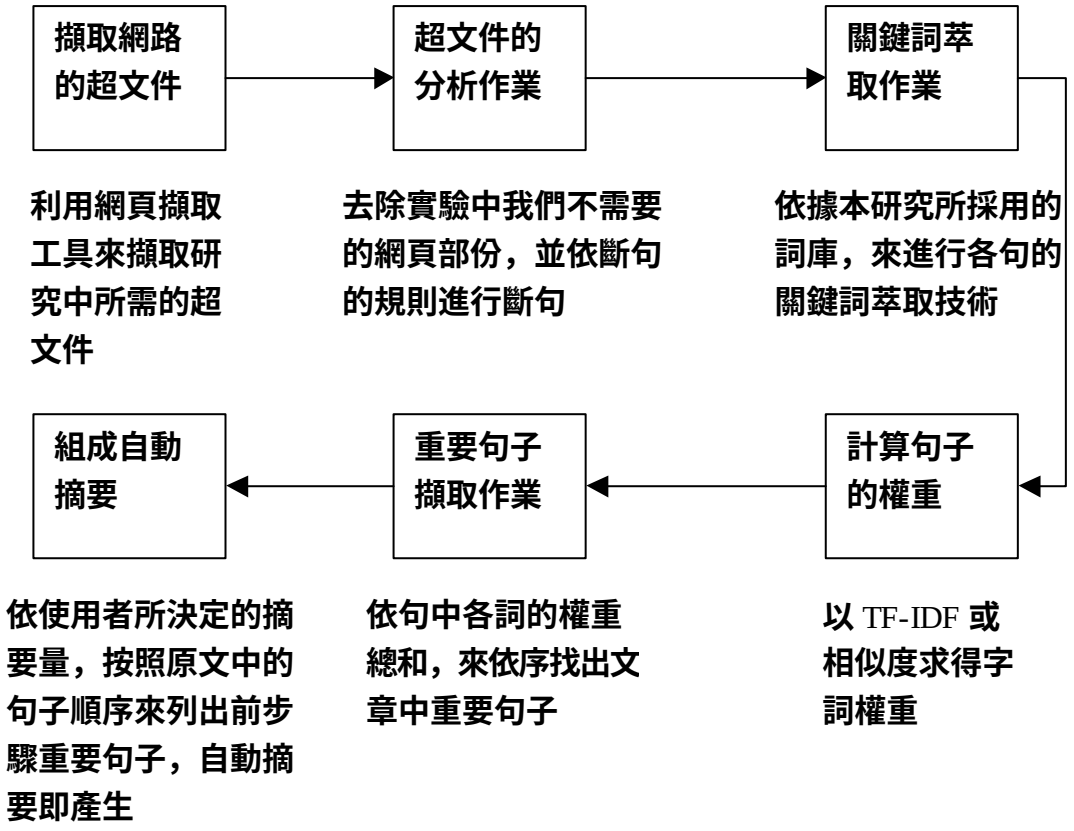


圖- 8 自動摘要的形成步驟

資料來源：邱立豐 民 91 年

另外，[22][23]**提出一個方法判斷句子在文章中的重要性。**

$$s\_imp(s,D) = \frac{1}{|s|} \sum_{w \in Keyw(s)} weight(w,D)$$

s_imp：**表示文章中句子的重要性**

$|S|$ ：**一個句子中，裡面共有幾個詞彙**

Keyw(s)：**在句子裡面，關鍵字詞的次數**

s：**表示句子**

w：**關鍵字詞的權重值**

D：**網頁文章**

第四節　向量空間模型

壹、向量空間模型

在傳統資訊檢索領域中，向量空間模型(Vector Space Model ,VSM)**是一種最簡單，且最具生產力的模型**[20]**。以下介紹向量空間模型基本概念。**

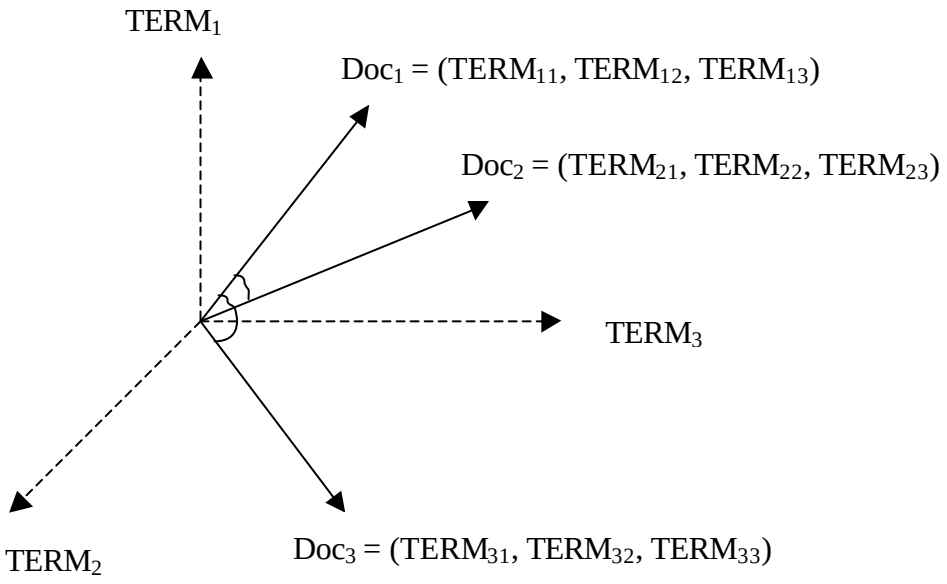


圖- 9 **向量空間模型**

資料來源：Salton & McGill 1983

參照圖-9 **向量空間模型概念主要是將每份文件或者段落、句子以向量來表示，而當中所包含的詞彙即為向量中的元素。因此，透過向量空間模型的概念，文件** D **可視為** D=(t1,t2,…,tn)**，其中** t **為文件中關鍵詞彙。而利用前述** TF-IDF[5][21]**字詞權重計算公式，每份文件** D **可視為** D=(w1,w2,…wn)**，其中** w **為該詞彙在文件** D **中之權重。**

透過向量空間模型概念，每篇文章出現的詞彙即可轉換成向量表

示法，參見圖-10 所示。利用建構的模型，便可輕易的計算出查詢語句與文件向量間的相似度，並進一步回饋給使用者。

	$Term_1$	$Term_2$	...	...	...	$Term_i$
Doc_1	W_{11}	W_{12}	...	...	...	W_{1i}
Doc_2	W_{21}	W_{22}	...	...	...	W_{2i}
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
Doc_k	W_{k1}	W_{k2}	...	...	...	W_{ki}

圖- 10 詞彙-文件矩陣

資料來源：Salton & McGill 1983

圖-10 **中的** Doc_1 **表示第** 1 **篇文章，**$Term_1$**、**$Term_2$**。。。至** $Term_{1i}$ **表示文章中的重要句子，**W_{11}**、**W_{12}**。。。至** W_{1i} **表示句子中每個詞彙的權重。利用詞彙-文件矩陣模型與相似度計算，可輕易找出文章中兩兩相似的句子並且自動產生摘要。**

貳、相似度計算

相似度計算是廣受運用的技術，不管是文件的群聚、分類和檢索等等，都需要利用相似度計算來進行處理。下表列出較常被大家運用的相似度公式。

表 1 相似度計算公式

Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Dice coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \bullet \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \bullet \sum_{i=1}^{t} y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i \cdot y_i}$

X= $(x_1, x_2, ..., x_t)$

$\lvert X \rvert$ =number of terms in X

$\lvert X \cap Y \rvert$ =number of terms appearing jointly in X and Y

資料來源：Salton **1988**

以下我們舉一簡單例子說明，請參見圖-11

	$Term_1$	$Term_2$	$Term_3$	$Term_4$
Doc_1	今天	天氣	非常	炎熱
Doc_2	下午	天氣	異常	炎熱
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
Doc_k	氣象	報告	天氣	炎熱

圖- 11 **詞彙-文件矩陣例子**

Doc_1**的句子：今天天氣非常炎熱**

Doc_2**的句子：下午天氣異常炎熱**

計算 Doc_1**詞彙權重值**:**今天**(0.02)**天氣**(0.2)**非常**(0.01)**炎熱**(0.18)

計算 Doc_2**詞彙權重值**:**下午**(0.01)**天氣**(0.2)**異常**(0.02)**炎熱**(0.18)

參照圖-11 **所示，本例子相似度的計算技術採用** Cosine Coefficient **方法，以利於計算** Doc_1**與** Doc_2**兩句子之間的夾角度數。**

$$Similarity = \frac{0.2 \times 0.2 + 0.18 \times 0.18}{\sqrt{0.02^2 + 0.2^2 + 0.01^2 + 0.18^2} \times \sqrt{0.01^2 + 0.2^2 + 0.02^2 + 0.18^2}} = 0.99$$

由此可知兩句子間的相似性 99%。**但是利用此相似度方法有一缺點，需要基底相同才能計算，也就是每個句子的詞彙長度必須一樣。**

第三章　網頁知識擷取機制架構

本論文於網際網路環境之下，發展一套網路知識擷取機制。其開發重點在於此系統應用文件資訊擷取技術以擷取文件內容等資訊；期望能藉由此系統模式與技術的發展，以減少傳統學習的時間並增加學習的效率。

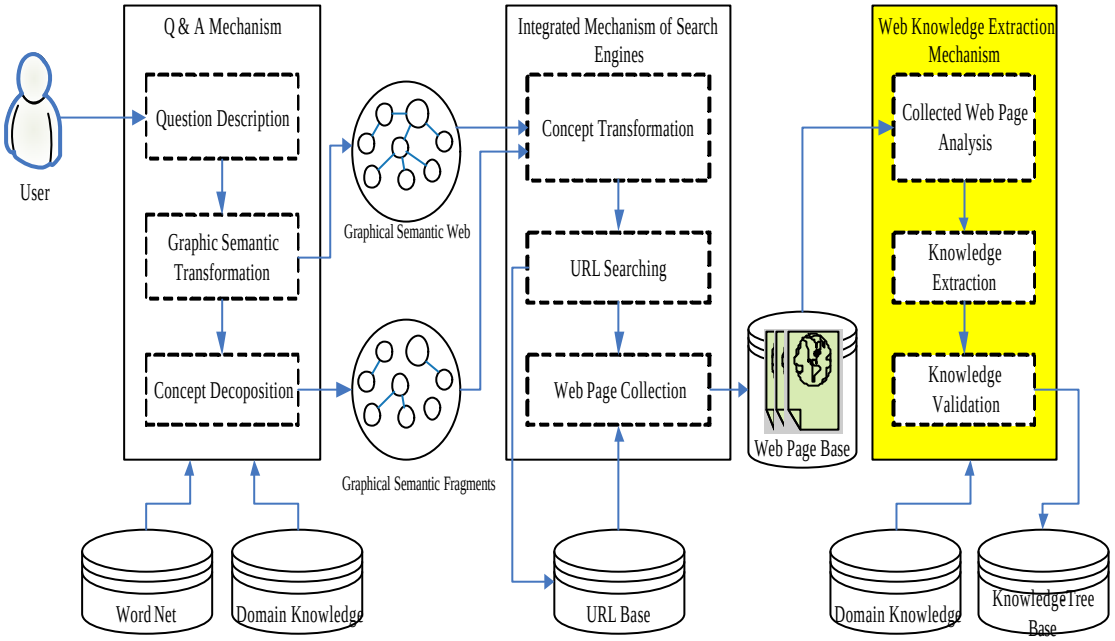


圖- 12　網頁知識擷取流程

圖-12 所示為整個網頁知識擷取流程，此流程分為三大部分。問與答機制(Q & A Mechanism)、整合式搜尋引擎機制(Integrated Mechanism of Search Engines)和網路知識擷取機制(Web Knowledge Extraction Mechanism)。本篇研究著重於網路知識擷取機制(Web Knowledge Extraction Mechanism)。

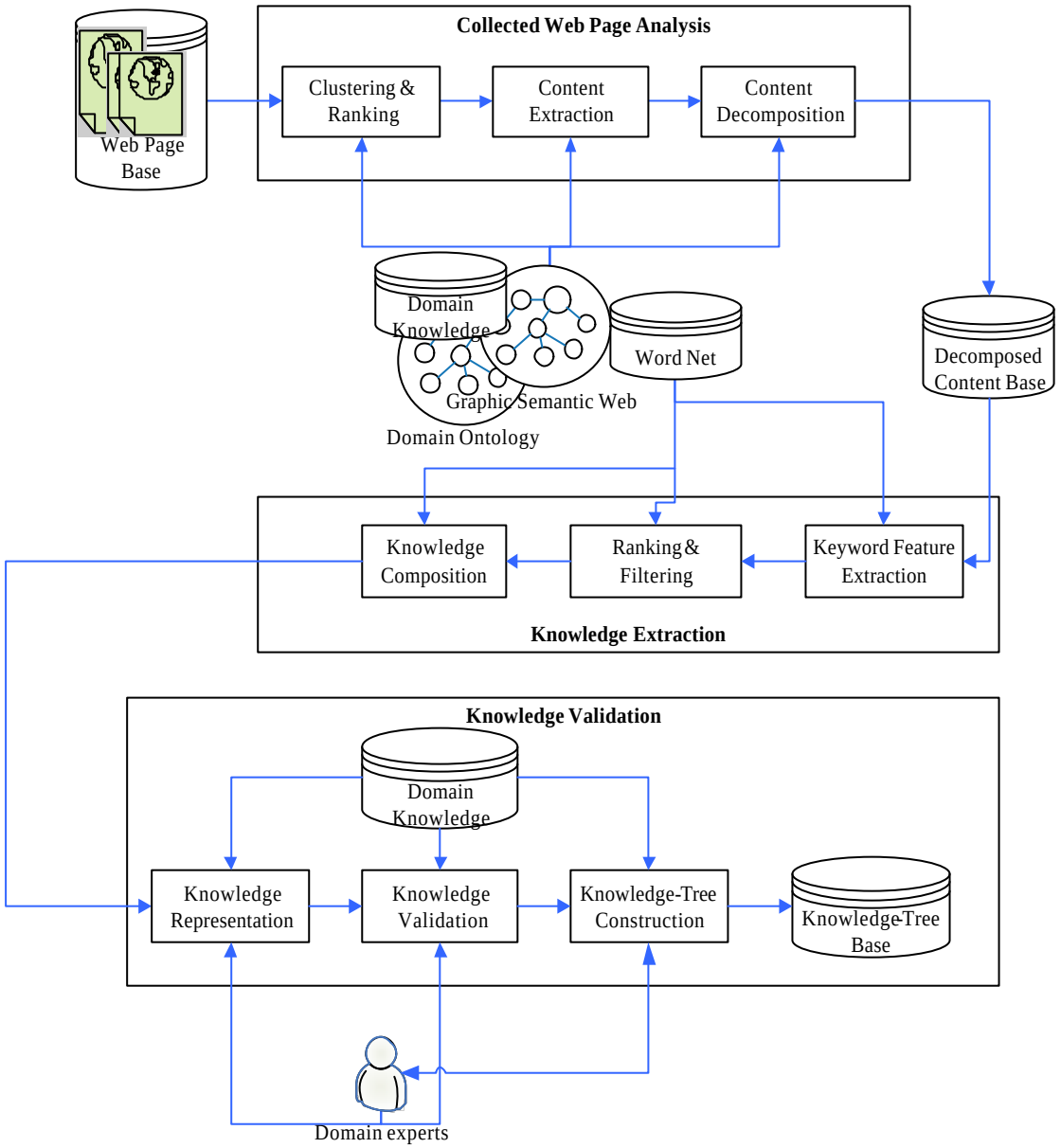


圖- 13 網路知識擷取機制架構圖

參照圖-13 所示，本研究架構網路知識擷取機制主要分三大部分。

一、分析蒐集的網頁(Collected Web Page Analysis)

二、知識擷取(Knowledge Extraction)

三、知識驗證(Knowledge Validation)

第一節　分析蒐集的網頁

功能分為三大部分：分群與排序(Clustering & Ranking)、內容擷取(Content Extraction)和內容拆解(Content Decomposition)。

壹、分群與排序

根據學習者語意搜尋的相關網頁存放於 web page base。把語意相近網頁內容群聚並且依照相似度高低依序排序，以利後續步驟進行。儲存在 web page base 之中的網頁利用向量空間模型(Vector Space Model, VSM)進行文件分群。

而文件分群的方法可分為「階層式分群法」（Hierarchical clustering）及「分割式分群法」（Partitional clustering）[9]。透過分群的結果，可以將web page base的描述或概念相同的文件歸類，這些類別可稱為「群集」（Cluster）。不管是階層式分群法或是分割式分群法，在web page base相當龐大的情況下，常有執行效能不佳的現象。故[20]提出一套較小成本的分群方法，稱為「Single-pass clustering」，本分群法所需的計算時間複雜度僅需要O(n log n)，其演算過程參見圖-14 所示

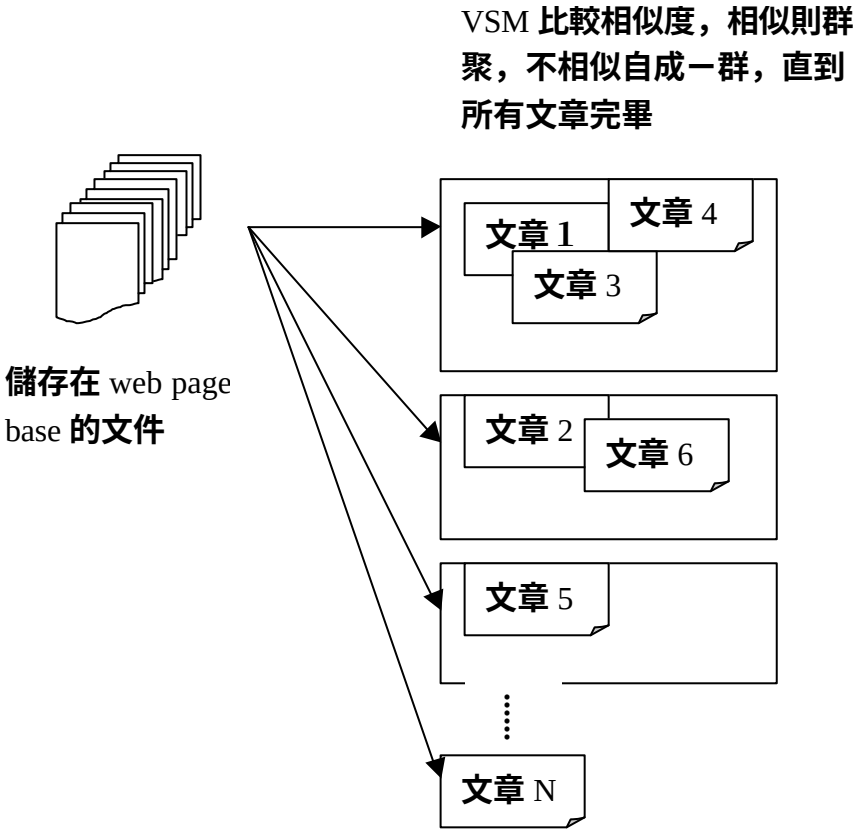


圖- 14 **文件群聚**

一、首先由web page base**之中取出第一份文件作為第一個群集。**

二、接下來再取出一份文件，並將之與現有的所有群集比較相似度。

三、指派該文件至適當的群集。

四、重新新增文件的群集向量。

五、若該文件並無與任何群集相似度高於門檻值，便自己形成一個新的群集。

六、重複步驟 2 **到** 6**，直至所有文件處理完畢。**

貳、內容擷取

把符合學習者語意的網頁進行內容擷取。參見圖-15 所示，網頁的形式大多為超文件(HyperText Markup Language, html)，透過超文件分析去除網頁標籤，擷取文件內容主體。

```
<html>
<body>

<p>
學習障礙，又稱為學習困難。
</p>

<p>
學習障礙（在現行的分類系統是指智能不足）有三個主要的部分—低智力表現；開始於出生或兒童早期；生活／適應技巧低下。
</p>

</body>
</html>
```

儲存在 web page base 網路文件格式如左，以人工校閱的方式去除多餘的文件標籤，使分析句子與段落變得更簡潔。

圖- 15 內容擷取

參、內容拆解

參見圖-16 所示，網頁文章由各個段落所組成，而各個段落又是由句子與句子之間所組成。而內容拆解目的即是把文章根據問號、句號、分號等等拆解成每個句子。依照 TF-IDF 與向量空間模式兩種方法，計算關鍵字在每一個之間彼此的重

要性，以利於下步驟「知識擷取」。

在內容拆解部分，我們運用領域本體論做語意處理的一種關鍵技術，在關鍵詞彙類似時能夠精確的對應一些意義相同的詞彙，倚靠由概念性模組所構成的本體論來精確的定義不同詞彙的意義。

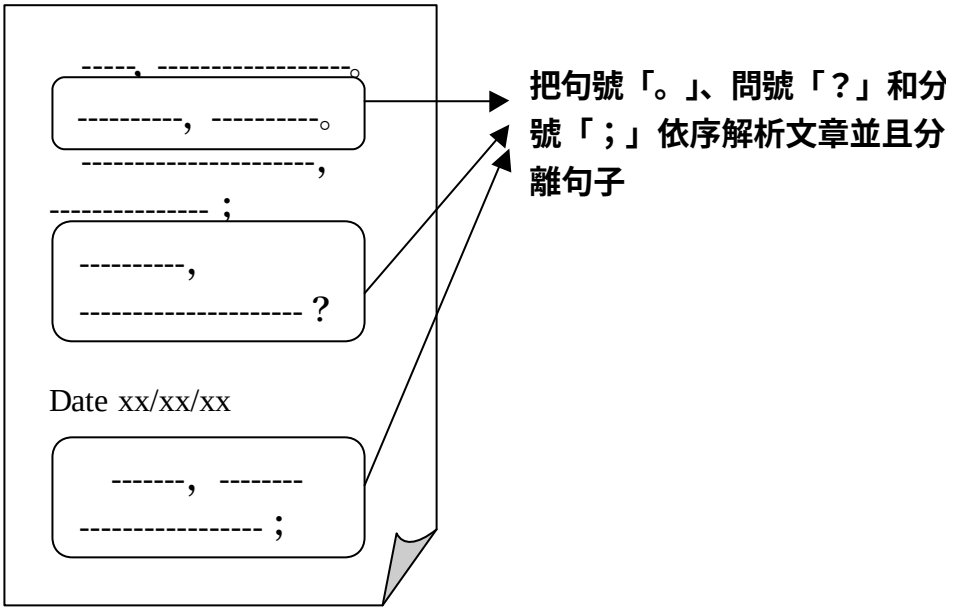


圖- 16 **內容拆解**

內容拆解儲存庫(Decomposed Content Base)**：參見圖**-17 **所示，用來存放文章經過拆解後的句子。**

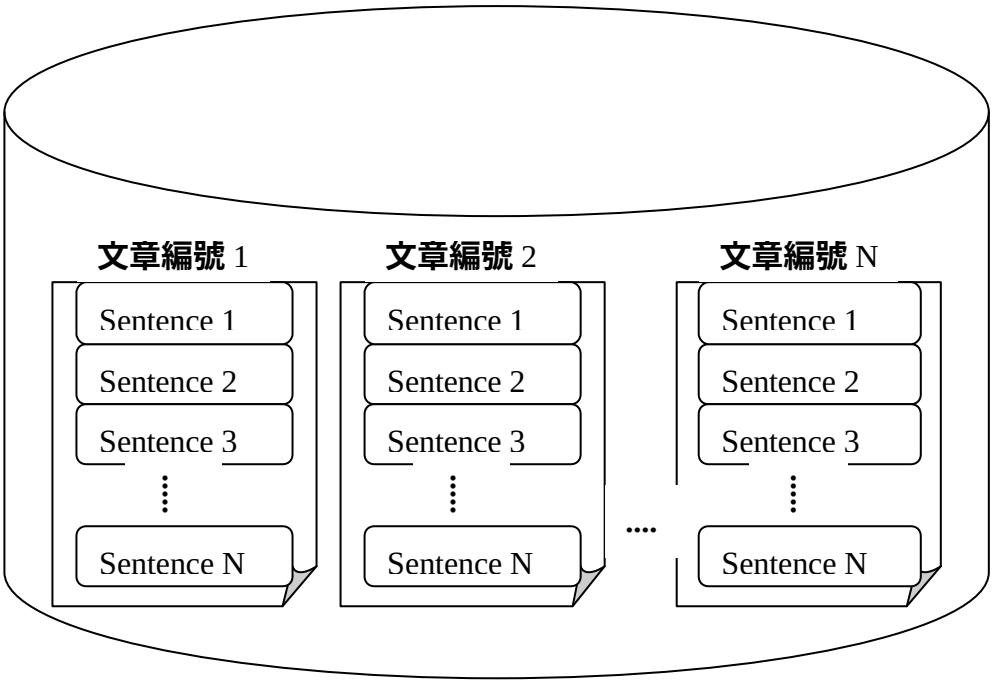


圖- 17 內容拆解儲存庫

第二節 知識擷取

功能分為三大部分，關鍵字特徵擷取(Keyword Feature Extraction)**、排序與過濾**(Ranking & Filtering)**和知識建構**(Knowledge Construction)**。**

一般網頁文章大多冗長，而學習者無法快速獲取所欲的知識，所以本研究「網路知識擷取機制」在於將冗長的文章濃縮成為簡潔的摘要，省去學習者自行搜尋與驗證網頁內容的時間，在第一時間內將濃縮的知識呈現給學習者是此機制的目的。

網頁文章內容經過前述的步驟已經拆解成段落，並分析出哪些段落是與學習者所要查詢的資料直接相關是知識擷取機制的主要目標；在這個機制中，基本的文章解析之斷字及斷詞是本機制的基本功

能，這些功能在前面已做簡單的介紹，所以本階段聚焦於下面主要功能的介紹。

壹、 關鍵字特徵擷取

關鍵字特徵擷取的任務在於減少資訊量，不重要的詞彙從特徵項空間中刪除，從而減少特徵項的個數。網頁內容拆解成為段落後，利用分號、句號、問候等等把段落分解成為句子，並且儲存至內容拆解儲存庫(Decomposed Content Base)。計算關鍵字在每個句子裡面的權重高低，選取權重值高於門檻值的句子。

貳、 排序與過濾

把權重值高的句子依照領域構詞規則與領域知識結合，一一把句子依序按照權重值由高至低排序。

參、 知識建構

透過句子與句子之間的領域構詞規則後，則可以產出整篇網頁文章的知識並且把這些知識總結。權重值較高的句子依序排列，配合本研究領域詞庫的混合斷詞法來產生每一篇文章的動態摘要，以兩百個字為預設值。

然而，根據單一關鍵字詞來判斷句子的重要性並無法適切的表達學習者的語意。原因如下：

一、單一關鍵字詞並無法確實的符合語意

例如：「腦科」與「電腦科學」。事實上，「腦科」這個詞彙泛指醫院內門診的用語，而「電腦科學」泛指資訊相關科系用語。在一般人認知中，這兩個詞彙完全是不相干的詞彙，但是在電腦中，卻認為是相同詞彙。所以單一關鍵字詞並無法確實的符合學習者語意。

二、句子與句子所組成，通常會有詞不達意且語句不通順的情形發生

由於文件摘要是由許多篇文章重要句子所組合而成，難免有語意不通順的情形發生，學習者進行研讀與學習會造成困難。

三、句子中只包含單一關鍵字所形成的摘要容易曲解使用者所查詢的原意

基於以上缺點並依據[22][23]**的方法，我們提出一個以多個關鍵字詞來評斷文章段落的方法。其方法如下：**

$$p\_imp(p,D)=\frac{1}{|s|}\sum_{Keyword=1}^{N} weight\ (FW_1+FW_2+..+FW_N,D) \qquad \text{公式}(7)$$

p_imp：**每個段落在文章中的重要性**

p：**文章段落**

D：**網頁文章**

$|s|$：**段落中所有的名詞**

Keyword **：經過** CKIP[16]**處理後所分析出來的關鍵字**

F **：關鍵字出現的次數**

W **：利用** TF-IDF **所建構出的權重值**

改善做法參見圖-18

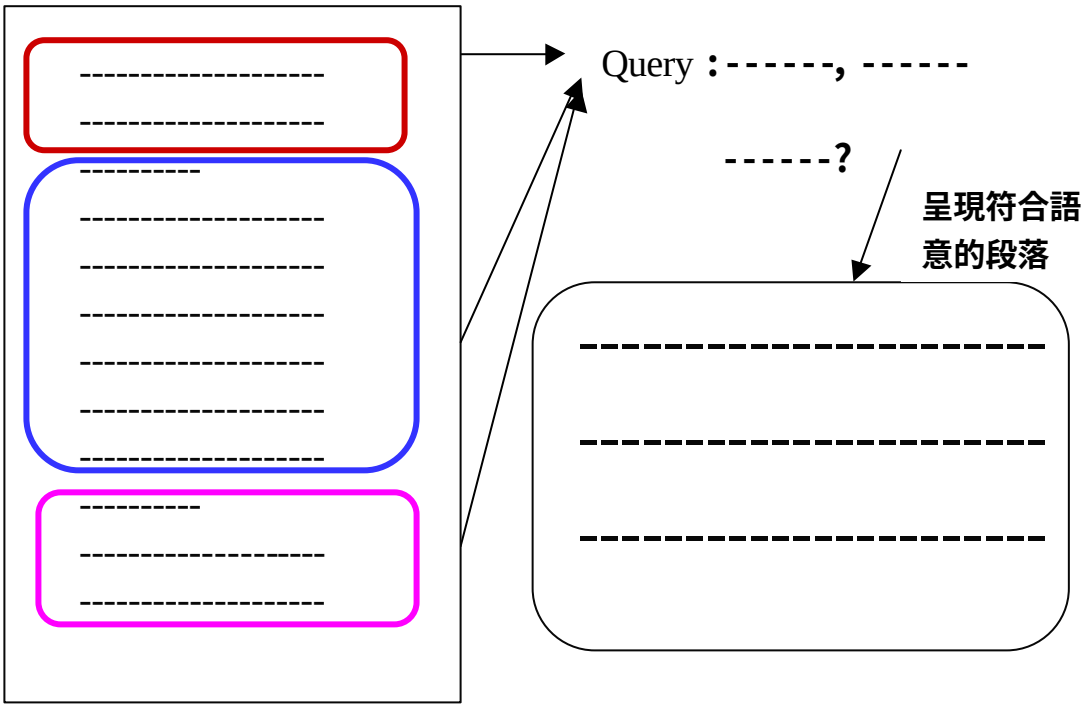


圖- 18 **改善作法**

此改善作法是根據學習者查詢的語意，系統根據文章中符合語意的段落計算哪個段落與學習者語意相近。此改善方法優點如下：

一、多個關鍵字詞可達成互相平衡語意上的差異。

二、增加語句順暢度與可讀性。

另外，我們舉一例子說明公式(7)

假設根據學習者查詢的語意經過解析後，權重值較高的關鍵詞為「學習」和「障礙」兩個詞彙。網頁原始文章段落內容如下：

一、學習障礙是一群學習異常現象的統稱，包括各種不同的類型。

二、學習障礙者一般智力在中等或中等以上。

三、學習障礙者在聽、說、讀、寫、推理、運算的學習上，會出現一項或多項的顯著困難。

四、這些學習上的異常是因為神經中樞的異常而導致，並不是由於智能障礙、感官缺陷、情緒困擾、環境文化等因素所造成。

五、學習障礙者雖然智力正常，但可能會出現學習成就和潛在能力之間有很大的差距，或是個體本身不同能力之間差異很大（亦即一項或數項能力特別低落，但是其他能力又表現良好），而產生令人難解的矛盾現象。

六、雖然學習障礙者因為腦部功能的不同，使其在資訊的接收和處理上異於常人，而無法在一般傳統教學下充分學習，但是透過有效的教學策略，學習障礙者一樣可以超越障礙，發揮潛在的能力。

此篇文章總共分為六個段落，利用 TF-IDF 方法計算本研究領域的關鍵字詞為「學習」和「障礙」兩字權重值。得知學習與障礙兩詞彙權重值各自為 0.233 和 0.2，之後依據先前提出改善的方法計算文章各段落重要性。

$P1=1/8 \times ((0.233\times 2) + (0.2\times 1)) = 0.083$

$P2=1/5 \times ((0.233\times 1) + (0.2\times 1)) = 0.087$

P3=1/6 × ((0.233×2) + (0.2×1)) =0.111
P4=1/12 × ((0.233×1) + (0.2×1)) =0.036
P5=1/14 × ((0.233×2) + (0.2×1)) =0.048
P6=1/17 × ((0.233×3) + (0.2×3)) =0.076

其中 P1 **表示文章中的第一個段落，分母則是句子包含的名詞與動詞的總數。文章符合學習者語意的重要性順序為** p3 > p2 > p1 > p6 > p5 > p4

第三節　知識驗證

這部份功能分為三大部份，分別為知識的呈現(Knowledge Representation)**、知識的驗證**(Knowledge Validation)**和知識樹的建構**(Knowledge Construction)**。**

當隱含再網頁的知識經過一連串的擷取程序形成摘要並組成知識之後，必須進行知識的呈現與知識驗證的工作，最後把知識建構成為知識樹[7]**。細部說明如下：**

壹、 知識的呈現

網頁的知識經過前述核心功能處理過後，可以把隱藏在網頁中的知識擷取出來，把擷取段落組成知識後並且配合領域知識將段落中意義不大或者較無相關的字詞去除，把簡潔的段落知識呈現給使用者。

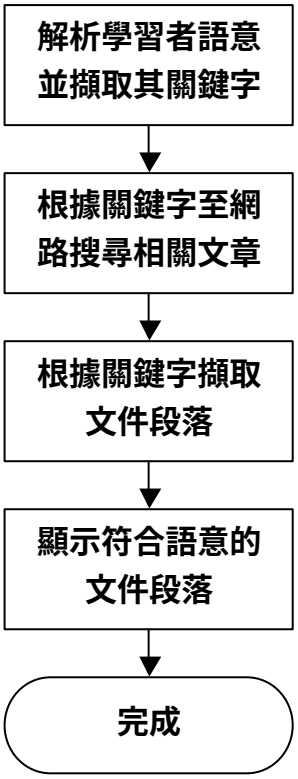


圖- 19 知識呈現流程

貳、知識的驗證

除了把隱藏於網頁中的知識擷取成為段落後，還必須確保知識的準確性，因此透過領域專家針對每篇文章的段落進行知識的驗證，然後把領域專家判斷無誤的段落輸出，作為建構知識樹的主要輸入來源。

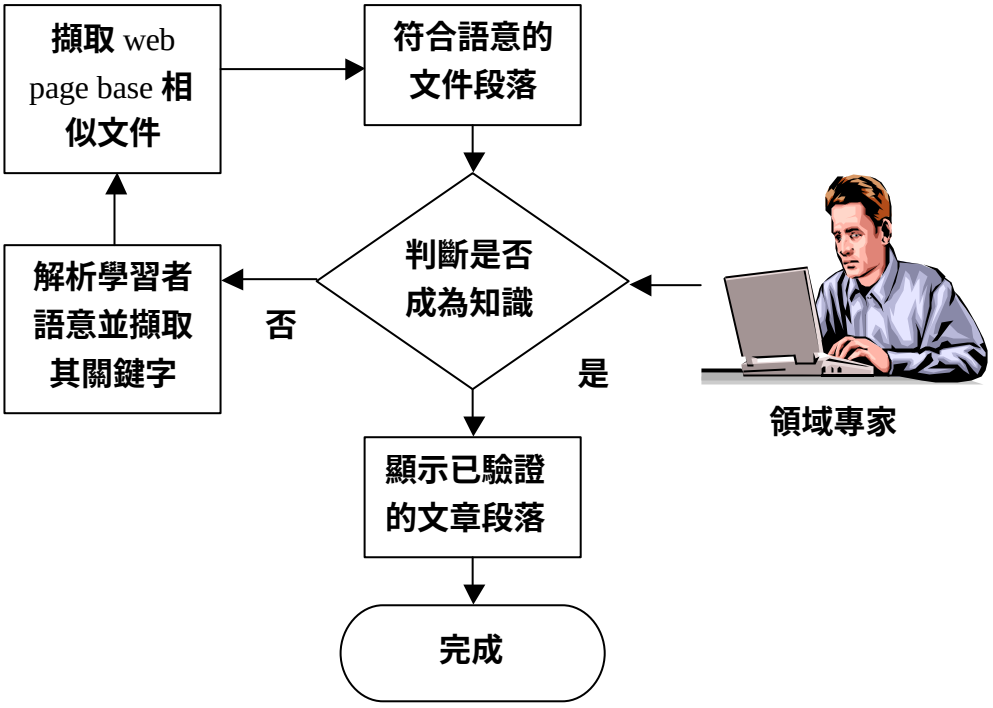


圖- 20 知識驗證流程

參、知識樹的建構

把每一篇網頁文章段落形成節點，節點中的檔頭儲存關鍵字詞，而其後面的儲存此關鍵字詞所擷取出來的文章段落。計算日後每一個新進來的節點與根節點的相似度，然後在決定知識樹的先後走訪順序。

隨著每一次學習者使用的次數越來越多，把這些龐大的知識摘要儲存至知識樹的資料庫。往後如果學習者對於目前的文章段落不如預期的時候，計算其他相似關鍵字詞的文章的節點並且呈現此節點內的段落。若學習者依舊不滿意當前知識的呈現，再經由上述步驟找出其他相似的節點，依此類推步驟以期望滿足學習者求知的需求。

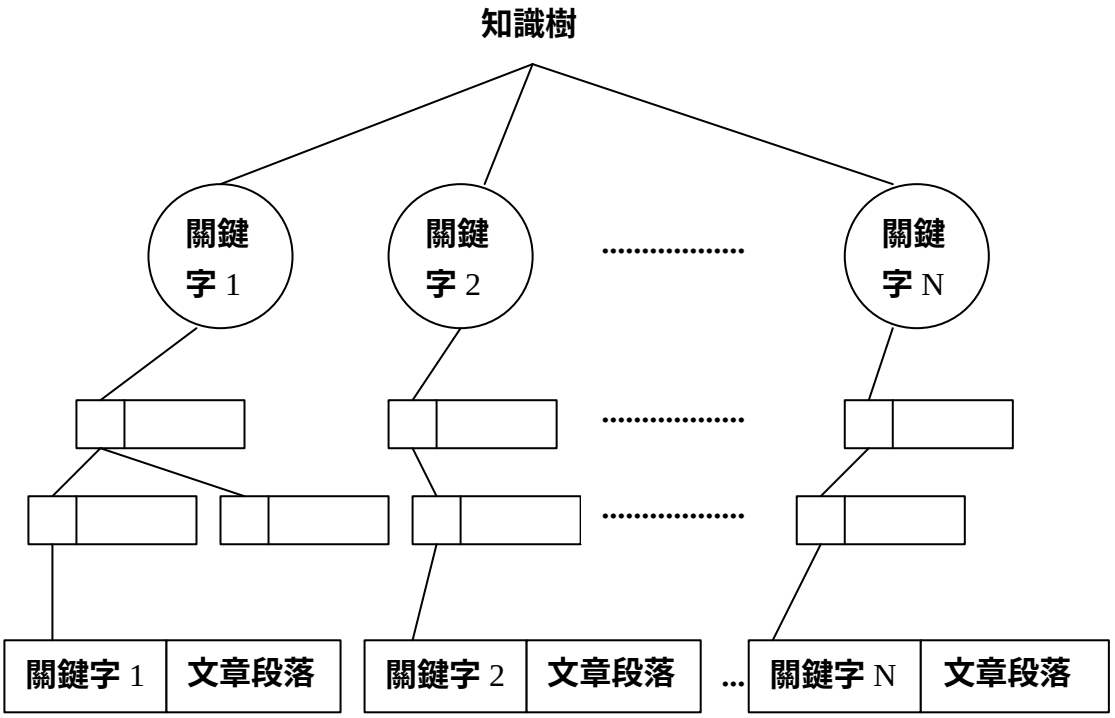


圖- 21 知識樹

隨著日後產生的擷取段落越來越多，知識庫勢必越來越龐大且搜尋速度隨之變慢，所以擷取知識建構成為知識樹是為了讓龐大的知識能夠有效率的搜尋、新增、刪除、修改。

參見圖-21 所示，知識樹的分支節點是學習者查詢的關鍵字 1、關鍵字 2 至關鍵字 N，而每個關鍵字又分支成各個子節點，這些子節點儲存了每個關鍵字所擷取出來的文章段落。

另外，系統設計流程參照圖 22

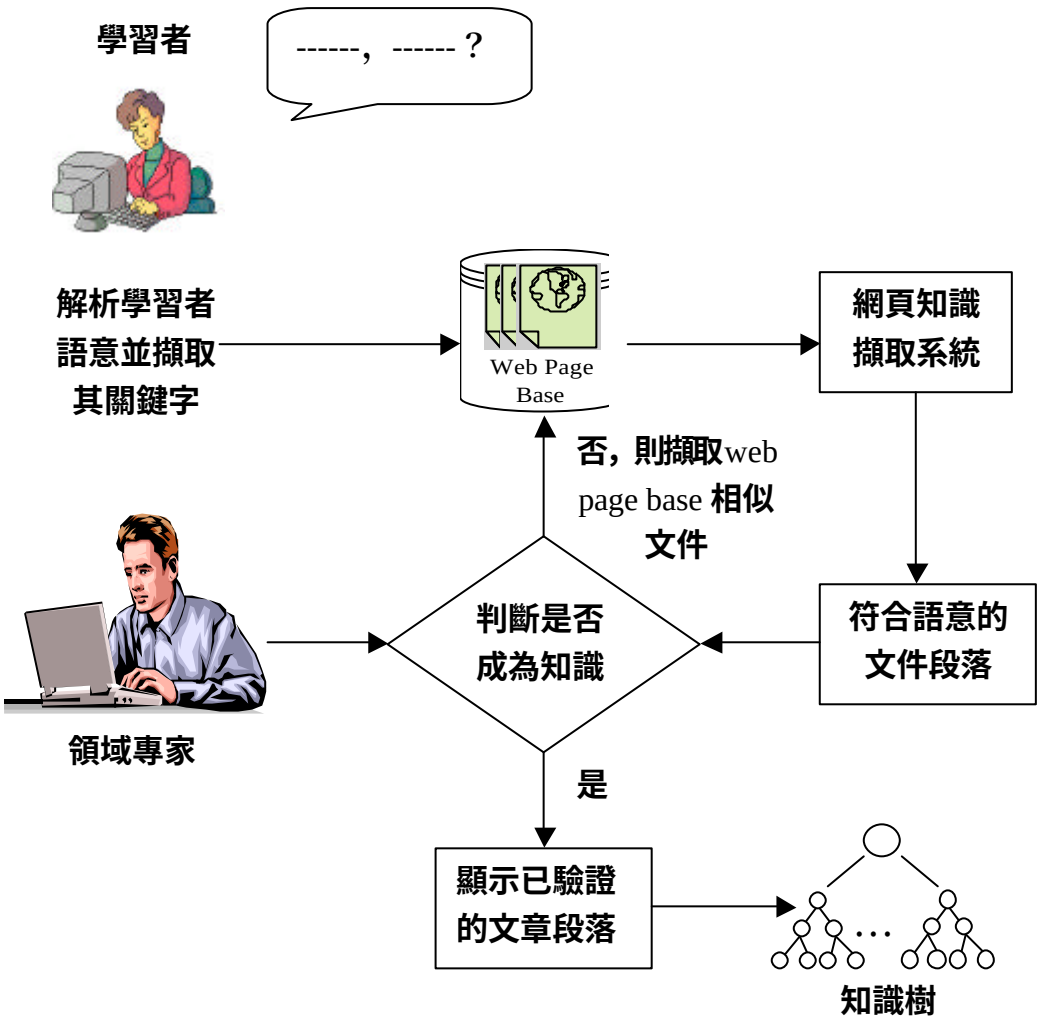


圖- 22 系統設計流程

1.解析學習者鍵入的語意並擷取出關鍵字詞。

2.在 web page base 之中尋找一篇與關鍵字詞相似文章。

3.使用網頁知識擷取系統擷取符合語意的文章段落。

4.符合語意的文章段落進行驗證。

5.商請領域專家進行知識的驗證。

5-1.若不符合領域專家之鑑定，重新至 web page base 之中尋找另外一篇與關鍵字詞相似文章。

6.**已驗證之知識建構成為知識樹，以利於知識庫日後的能夠有效率的搜尋、新增、刪除、修改。**

第四章　實例驗證

根據第三章所提出的雛型系統架構，本研究發展一套以網際網路為基之網路知識擷取機制，判斷文章中段落是否符合學習者語意並且呈現知識給學習者。

第一節　實驗環境介紹

為了實作本實驗，我們利用以下的環境來進行實作與測試

壹、軟體

一、作業平台：Windows XP Professional Edition

二、資料庫管理系統：Microsoft Access 2003

三、程式開發平台：JBuilder 2006 Enterprise Edition

四、程式語言：Java

貳、硬體

NB 主機：Intel Pentium M 1.7G CPU，512RAM

實驗目的主要探討在學習者語意查詢時，使用多個關鍵字所擷取的段落，是否比單一關鍵字擷取的段落更能貼切學習者語意。

第二節　資料來源

根據學習者查詢的語意，解析出關鍵字詞經由搜尋引擎搜尋相關文章並進行知識擷取。

假設學習者查詢的語句為「數學課堂上，對於老師所教導的九九

乘法似乎有學習障礙的問題。」

以上語句經過 CKIP 處理，擷取詞性為動詞與名詞兩者。因為根據[1][6]，所有詞性中，以名詞與動詞為數最多，所以本研究以擷取動詞與名詞作為評斷依據。

語意經過處理過的型態：數學(Na)、課堂(Nc)、老師(Na)、教導(VC)、九九乘法(Na)、學習(VC)、障礙(Na)、問題(Na)。

透過 TF-IDF 的計算，可以建構出領域詞庫的詞彙權重。本研究建構 50 個有關輕度數學學習障礙的領域詞庫。依據領域詞庫中關鍵字詞的權重高低，依序選取權重值較高的詞彙進行網路搜尋。

在本研究中，以「數學九九乘法學習障礙」這四個關鍵字詞權重值較高，以這四個關鍵字詞進行網路搜尋。本研究搜尋 25 篇有關於「數學九九乘法學習障礙」為關鍵字詞的文章。

本研究實驗步驟如下

一、單一關鍵字擷取文章段落

二、兩個關鍵字擷取文章段落

三、三個關鍵字擷取文章段落

四、擷取段落與使用者語意比較

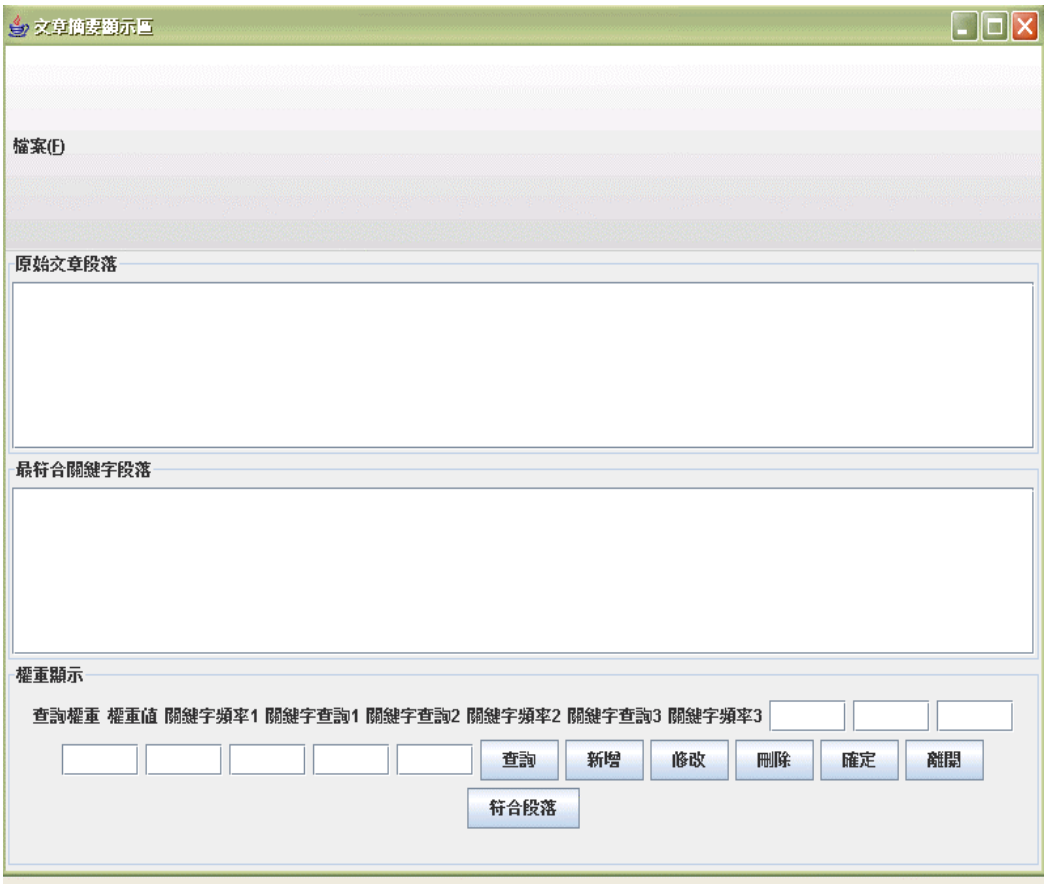


圖- 23 使用者介面

本系統提供予使用者使用之功能參見圖-23 所示，共提供關鍵字權重值查詢、新增、修改和刪除關鍵字等四種功能。

使用者介面分為三部份。依序為原始文章段落、最符合關鍵字段落和權重顯示等介面。以下就各項介面進行說明。

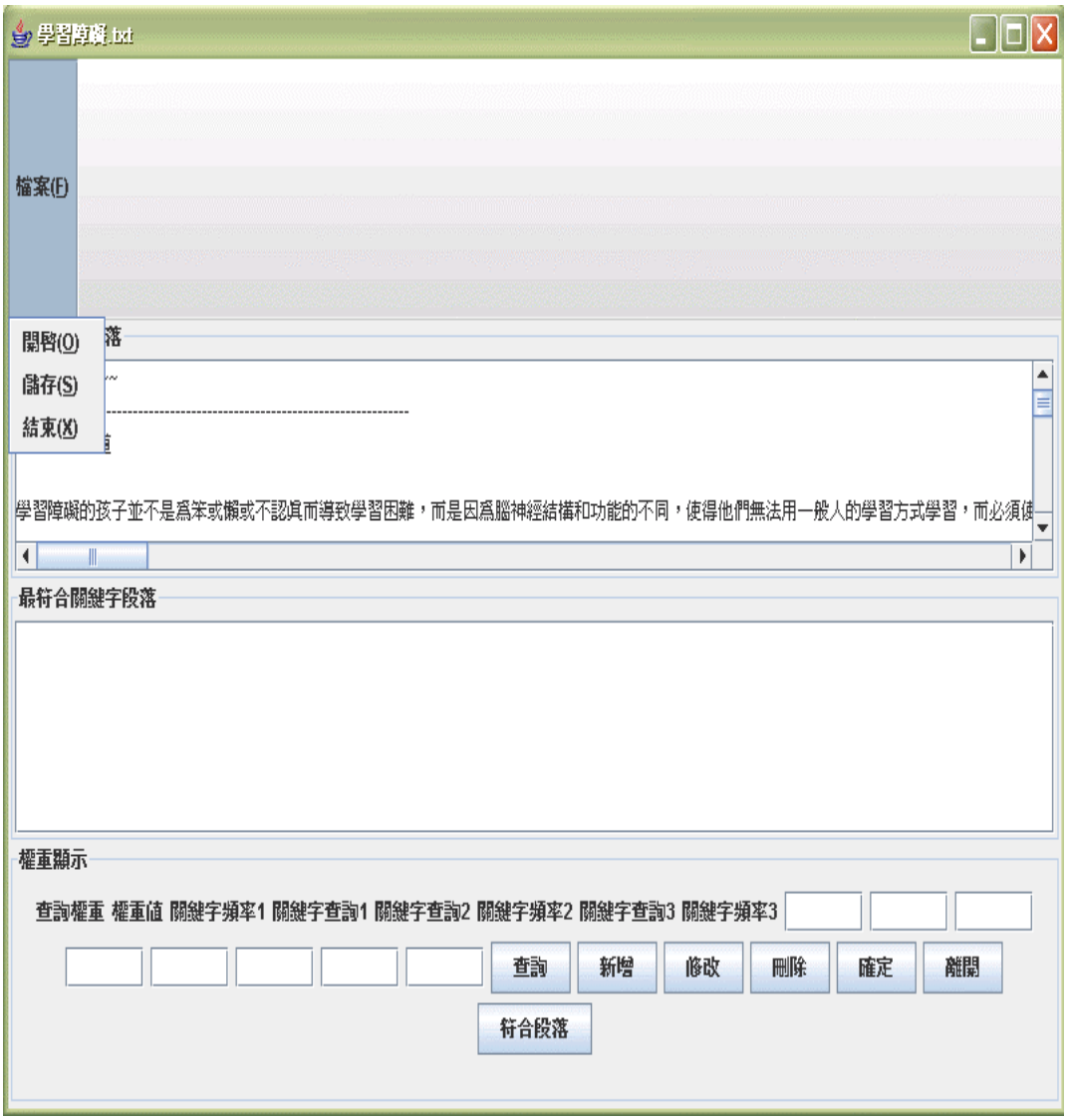


圖- 24 原始文章段落

參見圖-24 所示，畫面顯示把搜尋引擎搜尋到的檔案儲存至 web page base**。讀取檔案並顯示網路原始文章於原始文章段落。**

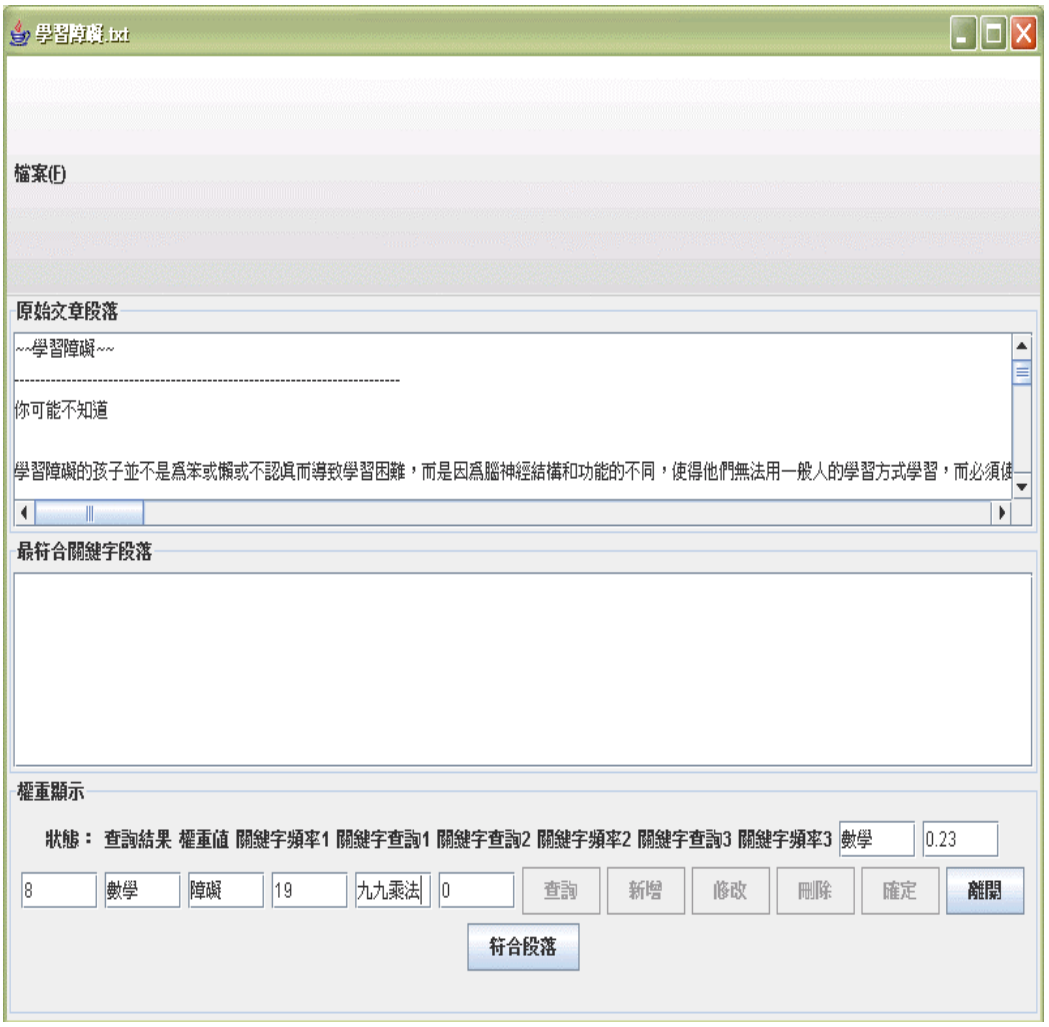


圖- 25 權重顯示

參見圖-25 所示，畫面顯示根據學習者的語意，在權重顯示功能中查詢關鍵字的次數以及權重值。找出原始文章裡面各關鍵字的權重以及頻率，以利於分析學習者的語意符合文章的哪一段落。

圖- 26 符合關鍵字段落

參見圖-26 所示，文章經過前述的步驟，在最符合關鍵字的段落出現符合語意的文章的段落。

第三節　實驗結果

表 2 顯示在 25 篇文章中，以單一關鍵字擷取的文章段落。

表 2 單一關鍵字擷取的段落

文章編號	單一關鍵字擷取的段落內容
第 1 篇	測驗過程中，我們看到學生用各種不同的方法答題。有的扳手指，有的口中唸唸有詞。有的先喵一眼全頁的題目馬上看出端倪。例如有一位學生在做十位數為 6 的借位減法時，將九題十位數都先填上 5，再算個位數。這些都是瞭解學生數學能力重要的資料
第 2 篇	今天，在報紙上又看到一些人反對國中升高中考作文。 大家說的都是技術上的理由，沒有一位主張廢考作文的人提出一個主張來如何加強國中生的寫作能力。我希望政府能成立一個跨部會的小組來提昇國民 「讀、寫、算」的能力。目標是要把每一個國民的讀、寫、算能力都提昇到一定的水準。把各級相關的教師好好的再培訓，讓他們能站在第一線上圓滿達成任務。把需要的預算編列出來，把計劃書寫出來，並且切實的執行。不要用廢除各種武功的辦法來「永遠不做什麼」 比方說，廢考作文，再廢考史地，再廢考數學，全廢了以後，大家用申請的方式進入學校，這時候，看看吧，看看是那個階級的同胞最佔便宜呢？
第 3 篇	記憶力問題；缺乏數學學習策略；處理歷程速度太慢；視知覺或視
第 4 篇	最後，我認為「九九乘法表」的練習無論從小二或是小三開始都無所謂，但重點應該是要在更早的階段有配套的措施。在學前教育中，數學力的養成是不容被忽視的，如何從具體量到半具體量，延伸到抽象量概念的形成，再配合「九九乘法表」學習，如此就能事半功倍，水到渠成。
第 5 篇	教育局長劉仲成說，相關問題的發生，是因九年一貫建構式數學不強調背頌九九乘法表所致，致學生家長反映學子計算能力變慢，縣府教育局十分重視，也向中央反映，目前要求學校建構式數學要教，但九九乘法表還是要背，因為背頌九九乘法，在生活上有其必要性。
第 6 篇	不可否認，九九乘法是學習數學計算相當重要的基礎，放眼世界各國，也是共通的教材。只是，會背九九乘法之外，怎麼教、怎麼理解，似乎比會背、不會背這個結果更重要。過往死背教出了一堆不明就裡的學生，後來極端建構又教出一群效率差、被害怕競爭力不足的學生，擺盪在死背和極端建構兩邊，正考驗老師們的教學專業。
第 7 篇	目前 DALE 的多媒體教學部分已完成數學部分的整數除法，此單元又分七個子單元，由九九乘法測驗、整數分群的概念至橫式及直式概念等，由簡而

	繁地呈現；單元中也包含遊戲部分，使用者除了需具備教學單元所建構的知識外，某些遊戲中也會測驗學生對於形狀及空間的概念。
第 8 篇	這種反思主要來自於認知到在不同的文化下，也可以生產出不一樣的數學系統，而這系統以適當的眼光來看，卻是充滿了趣味與前人的智慧
第 9 篇	第四件事是小三上老師要同學兩兩抽背九九乘法，我在小學的強記功夫算還不錯，但是抽背仍然帶給我不少困擾。一來因為我在壓力下，思考會變差，會更想不出來，二來我有點輕微顛倒症的毛病，對於二選一的抉擇特別感到難以擺脫，我記得 56 和 54 是最混淆的。這事一直到我確知 5＋4＝9 所以是 9 的倍數，才擺脫 7×8 和 6×9 的糾纏。一直到現在，我仍然對背九九乘法的事感到不怎麼舒服。後來，我至少看到兩個功成名就，但都不是在數學領域方面專才的外國學者，痛批背九九乘法帶給他們的痛苦（他們還至少是四年級才背的）
第 10 篇	注重思考訓練的建構式數學希望提昇孩子的數學興趣。
第 11 篇	國民小學數學新課程之精神 64 年課程處理乘法啟蒙教材的方式所衍生的問題
第 12 篇	我們也可以從兒童自己出的題目中，找到一些驗證--數學是生活化的，數學是人性化的
第 13 篇	李慶安則質疑建構式數學主導者黃敏晃與補習業者串聯，還點名曾參與制定建構式數學課程綱要的國科會科教處長林福來「球員兼裁判」。她指出，教育部評估建構式數學成效，卻找了參與制定綱要的教授來評估，林福來作的評估報告還聲稱接受建構式數學的小四學生能力沒有變差，這種「球員兼裁判」的研究報告根本不具公信力，教育部心態可議。
第 14 篇	一般而言，練習通常是指一定規則的重複演出，對於技能方面卻有其特定的功能。但是使用則強調把一項規則用在一個解題的情境之中。換句話說，情境中的意義是適於採用規則的。雖然在「分佈使用」中規則被重演，但卻是個有意義的重演。透過「分佈練習」而獲致的精熟學習，可能是沒有意義的熟練。機於數學教育的目的在於使兒童獲得數學的意義，採用「分佈練習」顯然較優於「分佈練習」。
第 15 篇	依據教育部日前修訂完成九年一貫課程數學學習領域綱要，九十四學年度起國小二年級學童必須學會九九乘法，而小學三年級學童則要會使用九九乘法。面對此教育新政策，長期關心孩子教育問題的功文文教機構總裁蔡雪泥博士，秉持著取之於社會，用之於社會的想法，再次發揮她的愛心，特印製九九乘法表及乘法 CD 的學習利器，免費贈送給全國需要的小朋友。本縣也由李清林議員發起「讓孩子學習更輕鬆」宣言，並於 4 月 6 日下午二時，於光復國小視聽室，由教育局文局長代表受贈，隨即轉送全縣二年級學童使用。

第 16 篇	基於學習障礙的多樣，輔導學習障礙者的老師也需要多樣的知識。他們要有：(一) 一般性的閱讀、數學、寫作等理論的知識。(二) 學習障礙者的特徵、學習能力、和認知能力的知識。(三) 鑑定和診斷學習障礙者的能力。(四) 矯正或補救閱讀、數學、寫作障礙的知識。
第 17 篇	數學原本就是生活的一部份，我們期待透過生活中的情境引發學生的學習興趣。讓我們牽引他走一小步，獨自大步向前走。
第 18 篇	林淑君老師說，即將於今年 9 月升小二的 30 萬名學童和家長不要緊張，為了避免學習壓力並增進學習效果，部編本數學課本總共安排長達一又 1/3 學期的課程，讓小二學童有足夠時間理解並熟練九九乘法，和早年的「死背而不重理解」或建構式數學「理解而不重背誦」不同。
第 19 篇	我認識一個國中生,她也是其他科都很好,但是就是數學成績不理想,我覺得她的問題常常是"記不住"和"粗心",像最近她在學"分數的四則運算",每次做錯一個題目,總能發現她是忘記要先乘除.後加減.括弧裡的要先算,不然就是計算錯誤.這些規則.口訣她都懂.也知道,但是就是在演算的過程中又會忘記,每次在她做題目之前提醒她,她也還是會忘記然後算錯,除了多提醒她以外,不知道有誰有什麼更好的方法??
第 20 篇	本教學活動編織的主題是「以符號代表數」，結合九年一貫課程數學學習領域「代數」部分的能力指標。並考慮下列原則：1.根據原住民的文化背景，使學生能在熟悉的生活情境中學習；2.提供弓箭、傳統服裝等，使學生有機會藉由具體操作以建立數學概念；3.提供合作學習環境，使學生有機會與同儕、老師溝通與互動，設計並實施教學活動以提供鷹架教學。
第 21 篇	貝爾伯強調，他要到開始當老師之後，才對數學有好感。他說：「小孩常常暴露在比自己心智早一、二年的東西裡。只要他們開始想，因為課程進展得太快，他們不專讀好某一科目，他們就會開始落後。」
第 22 篇	教育部主任秘書、前國教司長陳明印表示，82 年的建構式數學強調計算過程，不重視結果，而九九乘法則重視結果，不重視過程，現在的新課綱，則計算與過程並重，穩固學生數學能力。
第 23 篇	經過專家學者指出，這些智慧是經由參與某種相關活動而被激發出來的，雖然智慧的成長隨著智慧類別的不同而不同，但卻大致遵循一定的軌跡，即年幼時期開始發展，經過不同的顛峰發展階段，到了老年時期發展活力迅速或逐漸的下降。如語言智慧從兒童早期即開始發展，直到老年時期仍可持續緩慢發展；邏輯、數學智慧在青少年及成長早期達到高峰期；空間思考在兒童時期就已發展成熟，藝術眼光則持續發展到老年期；肢體、運作智慧隨著生理發展的日趨成熟而發展；音樂智慧的發展關鍵在兒童早期；人際及內省智慧的發展取決於幼兒經費。
第 24 篇	何謂學習障礙---腦神經發育異常，但智力正常；---約佔人口 3-5 %；---男女比例約 8：1；---先天的、生理的、不可逆的、終身的、外表看不出來的；---在學習上有單一或多重障礙（例如閱讀障礙、書寫障礙、聽的理

	解障礙、數學障礙、符號理解障礙、語言表達障礙、...)。
第 25 篇	我們的教科用書包含課本、習作、教學指引，以及配合學習相關之教學資源。課本或教科書，是師生上課使用的教材；其學習活動與內容之設計，應以學生為本位，藉由其生活經驗為中心，開展數學各主題之學習。習作或練習簿，提供學生上課時配合使用，或作為課後練習之用；在每一主題內容結束教學時，應有綜合學習單，以連結統整此主題內容或進行親子互動學習活動。除外，並依教材設計各種輔助學習之用具，可供學生具體操作和練習使用。教學指引或教師手冊，供教師教學或家長等指導教學時參考運用；每一冊中應臚列該冊教材綱要、敘明各主題名稱、教材內容、教學目標、時數配當和頁數等。其次，針對各主題，應有該主題內容之研究剖析和其設計理念說明，以及教學活動流程說明；並能提供各主題補充教學或親子活動，以期寓教於遊戲中，作為輔助課本活動之不足；最後臚列本主題相關研究文獻或參考資訊，作為進一步查詢使用。

表 3 列出 25 篇文章中，以兩個關鍵字擷取的段落優於以單一關鍵字擷取的文章段落。

表 3 兩個關鍵字擷取的段落

文章編號	兩個關鍵字擷取的段落內容
第 2 篇	現在，每當有人提出這樣做太繁複的時候，主張這樣搞的人就會說：「啊，那是你，你很聰明，但是有更多的人學不會呀，你要同情這些有學習障礙的人。」有的道理我同意，但是為什麼要以有學習障礙的人為主體來編課本呢?如果教育的目標很清楚，而受教的對象之間又存在很大的差異性，那麼就應該好好思考如何因材施教，而絕非大家都用同一套教材吧！
第 5 篇	針對目前國小新式教材，國語上課時數比以前少，學生因建構式數學不會背九九乘法表等問題，縣議員簡沛霖、莊文斌昨天在議會總質詢中，要求縣府教育局應重視改善，教育局長劉仲成表示，建構式數學導致學生計算能力變慢，教育局曾向中央反映，目前學校還是要求學生背頌九九乘法表，不僅學習有需要，未來生活中也可應用。
第 7 篇	根據研究者針對新竹市民富國小資源教室三至五年級學習障礙學生之國語及數學能力的觀察結果
第 8 篇	最後一個有趣的發現，就是不少同學眼中的數學「種類」增加了。筆者曾經訪談十四位大四同學（現已畢業），發現他（她）們在中、小學數學學習幾乎是以習題演算與獲得高分為核心，而進入數學系之後，所接觸的絕大多數是高度抽象化、形式化、以邏輯演繹為主的數學，也就是說，十幾年

	的數學學習經驗讓他（她）們認為數學是永恆、普遍並且具有絕對確定性的客觀真理。
第 9 篇	我覺得數學後來學得還不錯的人，沒有不痛斥反覆練習耽誤他們學習的進程，殘害他們解題能力以及對數學整體認識的發展。我還記得向武義教授就痛斥和差化積、積化和差公式在三角恆等式的應用：如：若 A＋B＋C＝180°，則 sinA＋sinB＋sinC＝4cosA/2 cos B/2 cos C/2
第 10 篇	曾輔導資優生、學習障礙學生的台大數學系副教授朱建正認為，這個時候問題會浮出檯面，原因在於高年級老師的數學能力通常較強，他們發現學生的計算熟練度大不如前，不能接受，才會反彈。
第 12 篇	這次本校二年級的老師在數學科的紙筆測驗上，內容稍作改變，藉此檢驗孩子的學習成果，卻發現了一些值得探討的問題，僅提出供家長、老師們一起來深思。
第 13 篇	至於教育部要對這些學習建構數學的學生進行「補救教學」的說法，更是讓人哭笑不得，他指出，補救教學是針對學習有障礙的學生作課後輔導，現在，補救教學卻必須普及到每一個學生。
第 19 篇	以心理學的角度看,自我暗示是建立信心的好法子,不僅是數學學習有效,其他學習科目亦同,日常生活效果亦是非凡,甚至於生病時,也有相當程度的效果
第 21 篇	喬治城日間小學（Georgetown Day School）三年級老師貝爾伯（Anthony Belber）表示，有些小孩在心智發展成熟前，就被要求學習一些概念，他說：「我還記得，在讀聖亞班斯小學（St. Albans School）時，對數學的慌亂感，我老是想勉強地混過數學課，也從來搞不清楚自己在幹什麼。」
第 23 篇	用邏輯—數學學習鳥類:收集鳥類統計數字，回答鳥如何飛行、什麼因素會影響鳥類飛行等等問題

表 4 列出 25 篇文章中，以三個關鍵字擷取的段落優於以兩個關鍵字擷取的文章段落。

表 4 三個關鍵字擷取的段落

文章篇數	三個關鍵字擷取的段落內容
第 3 篇	學習障礙的定義是以學業性學習障礙為主，將學習障礙分為閱讀障礙、數學
第 17 篇	「事前的預防，勝於事後的補救」學習是行為知識改變的歷程。平日教學活動多留意學生的個別差異，透過選擇合適的教材，活潑的教法，積極的教學態度減少學習障礙的發生，學生學習的困難必可減少，即使發現徵候立即予以矯正補救，相信數學學習輕鬆又愉快，師生都可收到事半功倍的

	效果。
第 19 篇	一、數學學習障礙：指個體智力正常，但在學習與運作數學符號的能力有困難，而導致數學成就低下的現象；其數學困難可能與下列因素有關：1．認知的缺陷，如解碼的能力不佳、後設記憶差等。 2．學科基本知識不足，可能來自發展與文化不利因素的影響所導致。3．非認知因素的影響，如成敗歸因、信心系統偏差等。
第 21 篇	巴洛帝說：「在有數學學習困難，甚至被貼上『學習障礙』標籤的小孩中，有器質性功能失調（organic dysfunction）的小孩也許佔了一小部分。」至於學習數學是否永不嫌遲，巴洛帝說：「如果學習精神猶在的話，也許永不嫌遲。」

另外，我們根據擷取的段落，進行與學習者查詢的語意做比較，評斷多關鍵字詞所擷取的段落是否比單一關鍵字詞有效率。採取的評斷方法為 F-measure 值。

表 5 說明 web page base 當中的 25 篇文章 F-measure 值。

表 5 實驗數據

篇數	單一關鍵字	兩個關鍵字	三個關鍵字
1	0.05	0.05	0.05
2	0.11	0.11	0.11
3	0.12	0.12	0.12
4	0.15	0.2	0.2
5	0.18	0.18	0.18
6	0.09	0.27	0.27
7	0.08	0.08	0.08
8	0.15	0.11	0.11
9	0.11	0.23	0.23
10	0.19	0.19	0.19
11	0.11	0.11	0.11
12	0.03	0.03	0.03
13	0.07	0.07	0.07
14	0.08	0.08	0.08
15	0.13	0.13	0.13
16	0.18	0.18	0.15

17	0.13	0.13	0.13
18	0.08	0.15	0.16
19	0.1	0.1	0.1
20	0.06	0.2	0.3
21	0.12	0.12	0.12
22	0.03	0.18	0.18
23	0.18	0.18	0.18
24	0.04	0.04	0.04
25	0.08	0.08	0.08

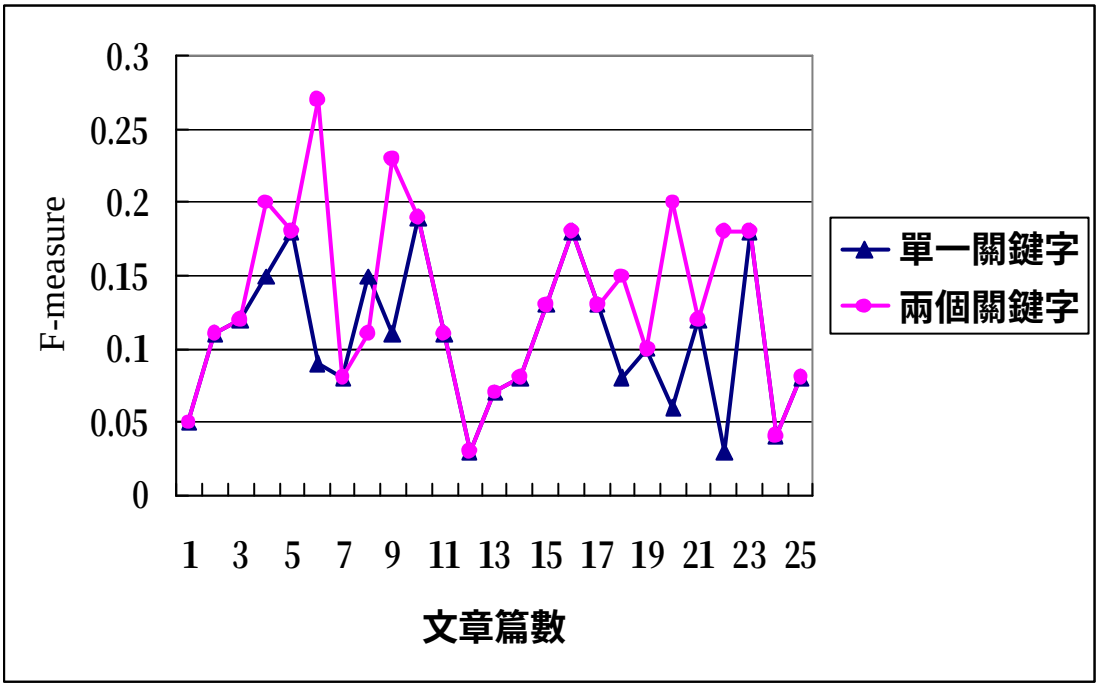


圖- 27 實驗結果(1)

參照圖-27 所示，進行單一關鍵字詞與兩個單一關鍵字詞所擷取出來的文章段落進行與學習者查詢的語意做比較。

經由圖表數據發現，其中文章第 4、6、9、18、20 和 22 篇兩個關鍵字所擷取出來的文章段落比單一關鍵字擷取較能符合學習者的語意。數據顯示文章編號第 8 篇則是單一關鍵字擷取表現比兩個關鍵字為佳。

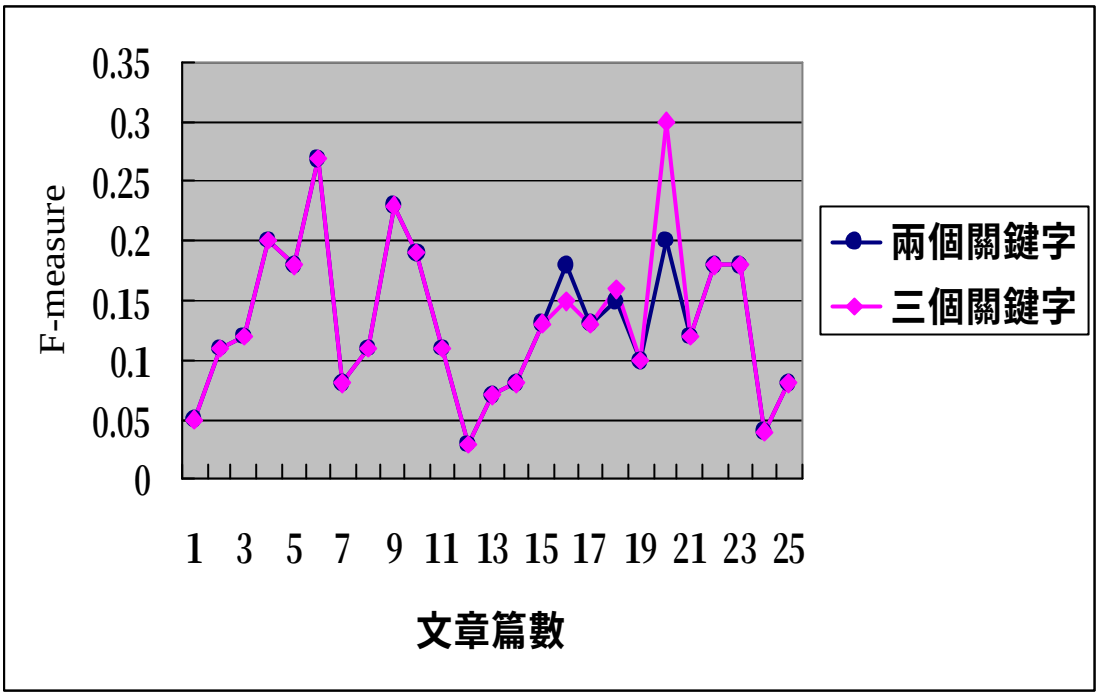


圖- 28 實驗結果(2)

參照圖-28 所示，進行兩個關鍵字詞與三個關鍵字詞所擷取出來的文章段落進行與學習者查詢的語意做比較。

經由圖表數據發現，其中文章第 18 和 20 篇三個關鍵字所擷取出來的文章段落比兩個關鍵字擷取較能符合學習者的語意。數據顯示文章編號第 16 篇則是兩個關鍵字擷取表現比三個關鍵字為佳。

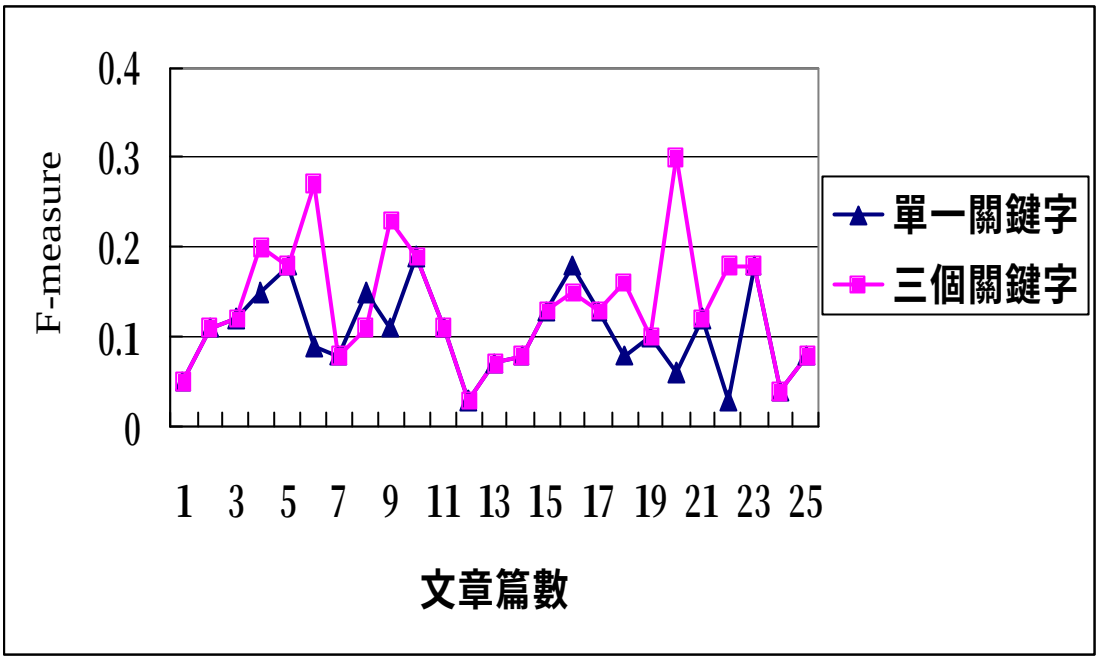


圖- 29 實驗結果(3)

參照圖-29 所示，進行單一關鍵字詞與三個關鍵字詞所擷取出來的文章段落進行與學習者查詢的語意做比較。

經由圖表數據發現，其中文章第 4、6、9、18、20 和 22 篇兩個關鍵字所擷取出來的文章段落比單一關鍵字擷取較能符合學習者的語意。數據顯示文章編號第 8 和第 16 篇則是以單一關鍵字擷取表現比三個關鍵字為佳。

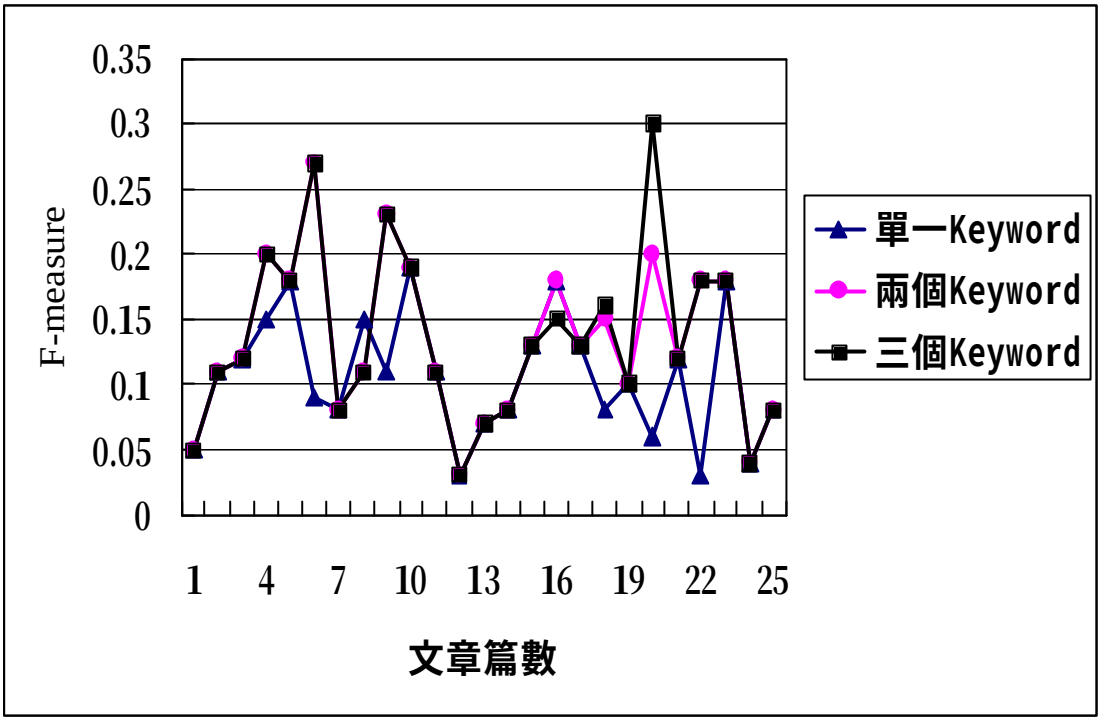


圖- 30 實驗結果(4)

參照圖-30 實驗結果，我們可以發現 25 篇文章之中，以單一關鍵字擷取的文章段落平均的 F-measure 值為 0.114，以兩個關鍵字擷取的文章段落平均的 F-measure 值為 0.141，以三個關鍵字擷取出來的文章段落平均的 F-measure 值為 0.144。

本研究提出以多個關鍵字所擷取的段落，比單一關鍵字能更符合學習者語意。因此，在多重關鍵字的環境下所擷取的段落更能比傳統單一關鍵字有不錯的表現。

在現今講求快速正確的年代，往往很難求出能夠滿足學習者語意的知識，有時候會將一些具有高價值性的資訊遺漏而不自知。為了避

免這種遺憾發生與語言上同字詞卻不同意義的情況，我們提出了一個根據多個關鍵字來判斷文章段落。幫助初次進入的學習者迅速的獲取相關的知識，省去繁雜的內容檢視時間。

第五章　結論與未來展望

第一節　　結論

在本篇論文中，我們提出了以多個關鍵字詞來判斷網頁文章段落是否符合使用者語意的方法，並且依據使用者的語意找出符合語意的段落。在我們的實驗結果顯示以多個關鍵字詞擷取文章段落，確實能符合學習者所查詢的問題。而以多個關鍵字詞所擷取的文章段落也確實比單一關鍵字詞為佳且符合語意。

本研究資料來源是根據學習者的語意並經由搜尋引擎所蒐集 25 篇文章，由於樣本數 25 篇文章略嫌少量，期望未來可以蒐集更多的樣本數文章，以利於更進一步的探討與研究。

形成多重網路文件段落：未來隨著系統的使用，會產生越來越多符合語意的段落，這時把這些符合語意的段落濃縮成為一篇多重網頁文件的摘要。把多重摘要交付給領域專家進行知識的驗證，確定正確無誤後，則進行知識的產出。多重摘要包含多個領域問題的解答，提供使用者領域知識，免除上網搜尋的時間，增加瀏覽效率。

隨著知識的新增與刪除，管理知識庫的工作也越來越煩雜。因此如何把網路上擷取與問題相關的知識，建立知識彼此間的關聯，將所獲取的知識建構成知識樹。知識樹除了記錄知識實體外也記錄知識實體間的關係，這樣能提供使用者認知中所需的知識，進一步提供使用

者未知且與所認知的知識相關的知識，如此使得學習能夠向外延伸。

第二節　未來展望

壹、網頁文件詞彙權重值計算的加強

本研究主要是利用TF-IDF 計算字詞權重，但透過研究結果檢視，部份重要詞彙並無法在此公式顯露其重要性，如研究中「障礙」二字詞，在部分領域網頁文件內其重要性與「殘障」幾近相同。故為了加強多文件摘要效益，日後可針對詞彙權重值計算予以改良。

貳、多文件段落語意改良

對於多文件段落形成，目前華語系國家仍無法結合中文語意，藉此提升段落可讀性。其最主要困難點在於中國字博大精深，相同的中文字，可以描述不同的情境主題，故有效發展中文語意知識庫，將可以大幅提升多文件段落的可讀性。

參、多文件段落區隔

研究中多文件網頁段落之生成，因其語句來源來自不同網頁文件，故造成多文件段落內容無法區分段落，進而造成使用者閱讀上的困難。

參考文獻

中文部份：

[1] 杜海倫，「以標題進行新聞自動分類」，清華大學資訊工程學系碩士論文，87 年 6 月。

[2] 邱立豐，「互動式概念查詢應用於網路文件摘要之效益」，雲林科技大學資訊管理學系碩士論文，91 年 6 月。

[3] 吳郁瑩，「網路中文超文件自動摘要之研究與實作」，雲林科技大學資訊管理學系碩士論文，88 年 6 月。

[4] 林俊佑，「在數位圖書館多代理人系統中以本體論為基礎的內容檢索」，清華大學資訊工程學系碩士論文，89 年 6 月。

[5] 郭家良，「新聞事件群聚及摘要檢索研究」，雲林科技大學資訊管理學系碩士論文，93 年 6 月。

[6] 許雅芬，「新聞文件自動分類之研究」，東吳大學資訊科學系碩士論文，90 年 6 月

[7] 黃美珠，「一個知識分類與搜尋相關資訊的架構」，中原大學資訊工程學系碩士論文，90 年 7 月。

[8] 黃耀明，「以子句擷取為基礎並應用於文件分類之自動摘要之研究」，台灣師範大學資訊工程研究所碩士論文，94 年 6 月。

[9] 曾憲雄等編著，資料探勘，台北，旗標出版股份有限公司，民國 94 年。

[10] 廖嘉欣，「實體論自動建構技術與其在資訊分類上之應用」，成功大學資訊工程學系碩士論文，91 年 7 月。

[11] 鐘明強，「基於 Ontology 架構之文件分類網路服務研究與架構」，成功大學資訊工程學系碩士論文，93 年 7 月。

[12] 顧浩光，「網路文件自動分類」，台灣大學資訊管理研究所，86 年 7 月。

西文部份：

[13] Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents," IEEE Intelligent Systems, Vol. 18, pp. 14-21, 2003.

[14] Berners-Lee, T., H. James, and L. Ora, "The Semantic Web", Scientific American, Vol. 284, pp.34-43, 2001

[15] Chakrabarti, S.,Mining the web: Mining The Web - Discovering Knowledge From Hypertext Data, Morgan Kaufmann, 2003.

[16] CKIP(Chinese Knowledge Information Processing), http://godel.iis.sinica.edu.tw/.

[17] Gomes-Perez, A., and O.Corcho, "Ontology Languages for the semantic web," IEEE Intelligent System, Vol. 17, pp．54-60, 2002.

[18] Gruber, T. R., "A translation approach to portable ontology specifications," Knowledge Acquisition, Vol. 5, pp. 199-220, 1993.

[19] Maedche A., M. Boris, S. Ljiljana, S.Rudi, and V. Raphael, "Ontology for Enterprise Knowledge Management," IEEE Intelligent Systems, Vol. 18, pp. 26-33, 2003.

[20] Noy, N. F., and D. L. McGuinness, "Ontology Development 101:A Guide To Creating Your First Ontology," Stanford Knowledge System Laboratory Technical Report KSL-01-05 and Stanford

Medical Informatics Technical Report SMI-2001-0880, 2001.

[21] Salton, G., and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc, 1983

[22] Tatsunori M., "Information Gain Ratio as Term Weight-The case of Summarization of IR Results", In Proceedings of the 19th International Conference on Computational Linguistics, pp. 688-694, 2002.

[23] Tatsunori, M., K. Miwa, and Y. Kazufumi, "Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems," In Proceedings of NTCIR Workshop 2 Meeting, pp 205-212, 2001.