A Study on Automated Text Summarization and Its Application on Chinese Documents

Student: Jen-Yuan Yeh Advisor: Dr. Hao-Ren Ke, Dr. Wei-Pang Yang.

Institute of Computer and Information Science

National Chiao Tung University

ABSTRACT

In this thesis, two novel approaches are proposed to extract important sentences from a document to create its summary. The first is a corpus-based approach using feature analysis. It brings up three new ideas: 1) to employ ranked position to emphasize the significance of sentence position, 2) to reshape word unit to achieve higher accuracy of keyword importance, and 3) to train a score function by the genetic algorithm for obtaining a suitable combination of feature weights. The second approach combines the ideas of latent semantic analysis and text relationship maps to interpret conceptual structures of a document. Both approaches are applied to Chinese text summarization. The two approaches were evaluated by using a data corpus composed of 100 articles about politics from New Taiwan Weekly, and when the compression ratio was 30%, average recalls of 52.0% and 45.6% were achieved respectively.

Keywords: Chinese Text Summarization; Corpus-based Approach; Latent Semantic Analysis; Text Relationship Map

文件自動化摘要方法之研究及其在中文文件的應用

研究生: **葉鎮源** **指導教授**: **柯皓仁博士，楊維邦博士**

國立交通大學資訊科學研究所

摘要

本論文提出了兩種新的文件摘要方法來摘錄原始文件中的重要語句。第一個方法屬於以文件集為基礎的摘要技術(Corpus-based Approach)**，此方法基於統計模型，利用特徵的分析來計算語句重要性。我們提出三個新的想法：**1) **利用語句位置重要性的分級以提高不同語句位置的重要性；**2)**利用詞彙相關程度**(Word Co-occurrence)**計算找出文件中的新詞，並將新詞加入關鍵詞重要性的計算，以得到更精確的關鍵詞權重特徵值；**3) **利用基因演算法訓練計算語句權重的** Score Function**，以期了解訓練文件集的特性。第二個方法，我們結合潛在語意分析**(Latent Semantic Analysis)**與主題相關地圖**(Text Relationship Map)**的概念，用來擷取文件中的概念結構**(Conceptual Structure)**以期得到語意層面的分析。實驗中，我們收集** 100 **篇新台灣週刊中關於政治類的文章，並將上述的兩種方法應用於中文文件的摘要實驗上。效益評估結果顯示，我們所提的方法都有不錯的表現，在壓縮比為** 30%**的情況下，平均來說，召回率分別為** 52.0%**及** 45.6%**。**

關鍵字：中文文件摘要、以文件集為基礎的摘要技術、潛在語意分析、主題關係地圖

致謝

本論文的完成，首先感謝指導教授柯皓仁老師及楊維邦老師。他們的指導，啟發我對於研究的興趣，並督導我學習研究的學問，更帶領我進入自然語言的研究中。除了課業與研究的寶貴指點，生活與做人處事上，也給我不少的影響。

感謝實驗室的夥伴們，由於你們對我的關懷與照顧，讓我的研究生活變為一種樂趣；並且感謝你們提供我寶貴的意見，讓我獲得許多新的想法。

謝謝圖書館計畫室的夥伴們在定期的報告討論時給我許多意見。另外，特別感謝何佳欣及鄭怡君幫忙作實驗評估的部分，使得我論文得以順利的完成。

最後，感謝我親愛的家人與朋友們長久以來的支持與鼓勵，使我能專心致力於研究，並得以順利完成學業。願以這篇論文與你們分享。

June 22, 2002

目錄

英文摘要 ..... **I**

中文摘要 ..... **II**

致謝 ..... **III**

目錄 ..... **IV**

圖目錄 ..... **VI**

表目錄 ..... **VII**

方程式目錄 ..... **VIII**

第一章　簡介 **1**

第一節　自動化資訊摘要 1
第二節　研究動機 4
第三節　研究目的 5
第四節　論文架構 5

第二章　相關研究工作 **7**

第一節　文件摘要相關研究 7
第二節　以文件集為基礎的摘要技術 12
第三節　以主題關係地圖(Text Relationship Map)**為基礎的摘要技術** 18
第四節　以語段模型(Discourse Model)**為基礎的摘要技術** 23

第三章　改良型語句權重摘要 **26**

第一節　基本特徵值分析 26
第二節　語句權重的計算與摘要生成 30

第四章　以潛在語意分析為基礎的語句摘要 **34**

第一節　潛在語意分析(Latent Semantic Analysis) 34
第二節　系統架構 39
第三節　語句分群與摘要生成 41

第五章　實驗結果分析與評估 **46**

第一節　實驗資料說明 46
第二節　評估方法 46
第三節　改良型語句權重摘要之效益評估 47
第四節　潛在語意分析語句摘要之可行性評估 53

第六章　結論與未來研究方向 **59**

第一節　結論與討論.....59
第二節　未來研究方向.....60
附錄一：實作系統展示.....**62**
附錄二：範例文件.....**63**
參考文獻.....**67**

圖目錄

圖 1：文件摘要系統架構概觀.....2
圖 2：文件摘錄的範例.....8
圖 3：相關研究工作.....10
圖 4：「以文件集為基礎的自動摘要技術」系統概觀.....13
圖 5：壓縮比對摘要系統正確率的影響.....16
圖 6：Text Relationship Map 的範例.....19
圖 7：Paragraph Relationship Map 與其對應的 Text Segmentation.....21
圖 8：計算 Aggregate Similarity 的概念圖示.....22
圖 11：m-n 基因交配方法.....31
圖 12：個體的基因組($M_{1,1}$, $M_{1,2}$, $M_{1,3}$, $M_{1,4}$, $M_{1,5}$)與其突變.....32
圖 13：Corpus-based 文件摘要生成演算法.....32
圖 14：LSA 工作原理示意圖.....35
圖 15：LSA 文件摘要系統架構.....40
圖 16：Word-by-Sentence 矩陣範例.....41
圖 17：LSA 文件摘要生成演算法.....45
圖 18：Modified Corpus-based Approach.....62
圖 19：LSA-based T.R.M. Approach.....62

表目錄

表格 1：以文件集為基礎的摘要方法研究的比較.....17
表格 2：Global Bushy Path, Depth-first Path 與 Segmented Bushy Path 的比較....22
表格 3：Characteristics of Discourse Model Approach.....25
表格 4：Lexical Chain 與 Co-reference Chain 的相異之處.....25
表格 5：實驗文件集的統計特性.....46
表格 6：考慮語句位置特徵時語句摘錄的召回率.....49
表格 7：考慮正面關鍵詞特徵時語句摘錄的召回率.....49
表格 8：考慮負面關鍵詞特徵時語句摘錄的召回率.....50
表格 9：考慮與標題的相似度特徵時語句摘錄的召回率.....50
表格 10：考慮向心性特徵時語句摘錄的召回率.....51
表格 11：利用詞彙相關程度所找到的部分新詞.....51
表格 12：Original 與 Modified 的實驗數據比較(考慮所有的特徵).....52
表格 13：Original 與 Modified 的實驗數據比較(不考慮負面關鍵詞).....52
表格 14：利用基因演算法所得到的特徵權重組(不考慮負面關鍵詞).....53
表格 15：Modified 與 Modified+GA 的實驗數據比較(不考慮負面關鍵詞).....53
表格 16：以 LSA 與 Keyword 向量表示法來實作 Global Bushy Path 摘要方法的比較.....54
表格 17：不同的維度約化比例對摘要結果的影響.....55
表格 18：LSA-based T.R.M.及 Keyword-based T.R.M.得到的主題相關地圖.....57
表格 19：各種摘要方法的綜合比較.....58

方程式目錄

方程式 1：給予 $F_1, \ldots, F_k$ 個特徵，語句 s 屬於摘要的機率 14
方程式 2：化簡後語句 s 屬於摘要的條件機率 14
方程式 3：當 s 屬於摘要的情形下，F_j 出現在摘要中的條件機率 14
方程式 4：訓練文件集中，特徵 F_j 的分佈機率 14
方程式 5：訓練文件集中，摘要語句的分佈機率 14
方程式 6：S_i, S_j 相似度的計算方式 22
方程式 7：S_i 的 Aggregate Similarity 的計算方式 23
方程式 8：s 的語句位置特徵值 27
方程式 9：A, B 關鍵詞的詞彙相關程度 28
方程式 10：s 的正面關鍵詞特徵值 28
方程式 11：給予負面關鍵詞 $Keyword_i$ 的條件下，s 不屬於摘要的機率 28
方程式 12：s 的負面關鍵詞特徵值 29
方程式 13：s 與標題的相似度特徵值 29
方程式 14：s 的向心性特徵值 30
方程式 15：計算語句權重值的 Score Function 30
方程式 16：K_{ij} 的計算公式 42
方程式 17：W_i 於 S_i 中的相對頻率 f_{ij} 43
方程式 18：W_i 於 D 中的資訊分佈值 43
方程式 19：W_i 於 S_j 中的總體權重 G_i 43
方程式 20：W_i 於 S_j 中的權重 L_{ij} 44
方程式 21：自動摘要系統的精確率評估 47
方程式 22：自動摘要系統的召回率評估 47

第一章　簡介

第一節　自動化資訊摘要

隨著電腦科技的進步與數位資訊技術的蓬勃發展，網際網路的存在儼然成為現代人生活中不可或缺的重要角色，並且帶動了人類文明往新的資訊紀元推進。拜科技之賜，大量的數位資訊在網路流通，網際網路無形中成為一個儲存各種資訊的大型倉儲；資訊的傳播不再完全藉由傳統平面媒體，人們漸漸地習慣在網路上找尋所要的資料，資訊的取得變成非常容易的事情。

隨手可得的資訊相對地也衍生許多問題，其中最大的問題是面對如此龐大的資訊時，人們無法快速且有效地得到真正符合自己需求的資料。究其原因，乃是因為大量的資訊使得搜尋及分辨是否為相關資訊的困難度大幅提昇。為了解決上述問題，人們需要藉助外在工具以便於在短時間內理解所取得資料中隱含的意義，迅速且正確地判斷真正符合自身需求的資料。

前述常用的工具主要分為兩大類：(1)搜尋引擎(Search Engine)，(2)自動資訊摘要(Automated Information Summarization)。[Gong01]對於上述二種工具做了以下詮釋：搜尋引擎所扮演的角色是『資訊過濾器(Information Filter)』，它的功能是分析使用者所下的檢索條件(Query)，並從資料倉儲中篩選出與檢索條件相關的資料；自動資訊摘要則是扮演『資訊監察者(Information Spotter) 』的角色，它的功能是將相關的資訊作統整，幫助使用者在最短時間內得知資訊內容的意義。

自動資訊摘要是由電腦自動地從原始資料中精練出最重要資訊的過程。根據原始資料的性質，自動資訊摘要大致上可分為以下三種：

- 文件摘要(Text Summarization)—原始資料為純文字；
- 多媒體摘要(Multimedia Summarization)—原始資料為影音等多媒體；

- **複合性摘要**(Hybrid Summarization)—**原始資料綜合了純文字與多媒體。**

[Mani99]**為文件摘要作了以下的定義：**

> **文件摘要是從原始文件中精練出最重要資訊的過程；其結果為足以代表該原始文件的精簡化版本，且可作為人們或其他資訊系統的判斷與決策依據。**
>
> Text summarization is *the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).*

圖 1 是文件摘要系統架構與流程圖，自動化文件摘要的過程可分為三個階段：首先是「分析原始文件(Analyze the input text)**」與「選取重要特徵**(Select salient features)**」；接著將分析的結果轉換為系統內部的摘要表示法**(Transform the input text into a summary representation)**；最後是評估內部摘要表示法的重要性，並挑選候選的表示法來合成摘要的輸出格式**(Synthesis an appropriate output form)**。**

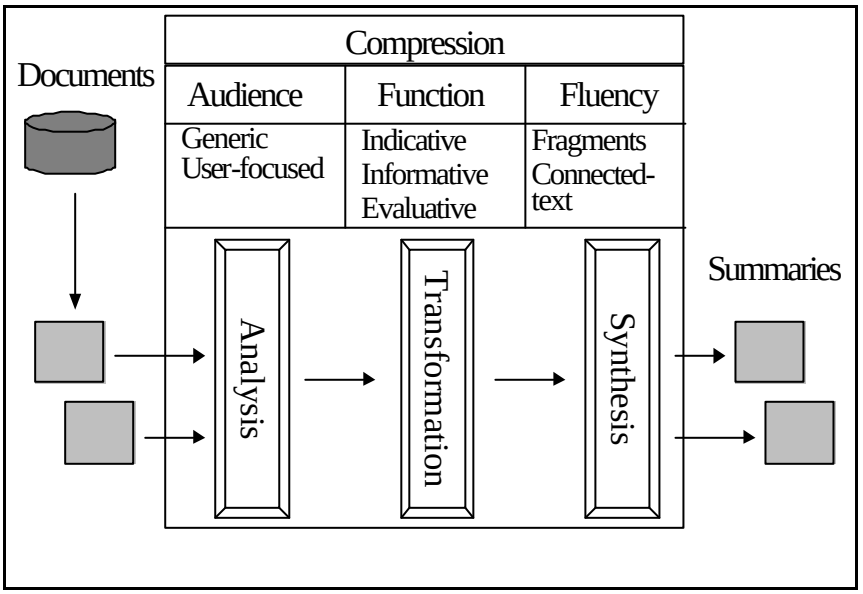

圖 1：文件摘要系統架構概觀 [Mani99]

整個過程中有幾項重要的因素需要考慮，如使用者對於摘要內容的需求、摘要內容的形式、摘要內容的流暢程度、可閱讀性以及文件摘要間的壓縮比(Compression Ratio)等等，都會影響所產出摘要結果的好壞 [Mani99]。

文件摘要的壓縮比是評估文件摘要系統優劣的重要指標之一，所謂的壓縮比係指摘要文件長度與原始文件長度的比例。壓縮比愈低的話，產出的文件摘要愈精練，但相對地也遺漏了愈多原始文件的資訊。相反地，壓縮比愈高，產出的文件摘要愈冗長，雖然包含的資訊愈多，但是相對地也包含愈多不重要的資訊。一般而言，壓縮比約在 1%—30%左右，便可以提供足夠的資訊給使用者作為決策判斷的依據 [Habn00] [Kupiec95] [Mani99]。

文件摘要系統最後所產生的摘要，可能是經過自然語言處理(Natural Language Processing, NLP)所潤飾過的文字(Connected-text)，也可能是原始文件的摘錄(Extract)—即直接由原始文件中節錄出足以提供該文件提及之事實或資訊的文句與段落片段(Fragments)。

文件摘要依其原始文件數量的多寡，可分為單文件摘要(Singular Document Summarization) [Aone99] [Edmundson68] [Hovy99] [Gong01] [Kim00] [Kupiec95] [Luhn59] [Myaeng99] [Salton97] [陳鈺瑾 00] 與多文件摘要(Multiple Documents Summarization) [黃聖傑 99] [翁鴻加 01] [蘇哲君 01]。單文件摘要把單篇文件的內容精簡化與重點化，注重的是能否有效地刪減沒有必要的資訊，並留下真正能代表文件內涵的資料；多文件摘要則是把多篇探討類似主題或事件的文件融合在一起，除了刪減無用的資料外，尚需有效率地過濾重複在多篇文章中所出現的資訊。

根據文件摘要所要達到的目的，產出的摘要結果可分為指示性摘要(Indicative Summary) [Aone99] [Edmundson68] [Gong01] [Kim00] [Kupiec95] [Luhn59] [Myaeng99] [Salton97] [陳鈺瑾 00] [黃聖傑 99] [翁鴻加 01] [蘇哲君 01]、

資訊性摘要(Informative Summary) [Hovy99]**與評論性摘要**(Evaluative Summary)**三種。指示性摘要提供閱讀者足夠的資訊，使其能夠根據這些資訊判斷並決定是否閱讀原始文件；資訊性摘要提供豐富的資訊內容，有時甚至可以取代原始文件；評論性摘要以摘要形式對原始文件作評論，可提供閱讀者不同角度的論斷。**

依照讀者需求的不同，文件摘要的結果可分為一般性摘要(Generic Summary) [Aone99] [Edmundson68] [Gong01] [Hovy99] [Kim00] [Kupiec95] [Luhn59] [Myaeng99] [Salton97] [**陳鈺瑾** 00] [**黃聖傑** 99] [**翁鴻加** 01] [**蘇哲君** 01] **及特定使用者導向**(User-oriented Summary)**的摘要等。前者針對較廣大的讀者群，摘要系統所產生的摘要以寫作者的角度出發，期能提供一般性的摘要給所有讀者閱覽；後者根據特定使用者的需求—如使用者感興趣的主題或是使用者所下的檢索條件—所產生的專屬摘要。隨著資訊爆炸時代的來臨，如何針對使用者的特定需求來產生摘要已經變得越來越重要。**

由語言的角度來看，摘要可分為單語言摘要(Mono-lingual Summarization) [Aone99] [Edmundson68] [Gong01] [Hovy99] [Kim00] [Kupiec95] [Luhn59] [Myaeng99] [Salton97] [**黃聖傑** 99] [**翁鴻加** 01] **與多語言摘要**(Multi-lingual Summarization) [**陳鈺瑾** 00] [**蘇哲君** 01]**。多語言摘要係指原始文件包含多國語言。這類研究，最大困難在於多國語言間字詞的用法、語句的表達方式及字義間了解與轉換所造成的語意模糊和誤解，若沒有領域知識**(Domain Knowledge)**與人工適時的介入，可能導致產出的摘要與原文件所要表達的意思南轅北轍。**

第二節　研究動機

好的文件摘要必須滿足以下兩個條件：

- **文件摘要必須要在真正了解文件內容之後而產生；**
- **文件摘要必須涵蓋原始文件所要表達的意涵；**

為了滿足上述的兩個要求，我們認為較佳的文件摘要系統必須要能夠理解文件的內容，並建構足以代表該原始文件意涵的知識模型，以便透過該知識模型來生成最後的摘要結果。

過去文件摘要的技術主要都是著重於英文文件摘要方面的研究，有鑑於英文文件與中文文件特性—比如關鍵詞的斷詞、語句切割、特徵值計算方式等—的不同，如果要將英文文件摘要的方法套用到中文文件摘要上，勢必要有所修正。

本論文研究的動機便是希望針對中文文件與英文文件特性的不同，修改過去文件摘要的技術應用於中文文件上，並提出一套知識模型來表達原始文件的意義，最後我們將以該知識模型為基礎提出一個文件摘要的演算法，並將其應用於中文文件。

第三節　研究目的

綜合上述的說明，本論文主要的研究在於單文件的自動摘要產生，且所著重的是如何產生具指示性、一般性與單文件的摘錄(Indicative and generic single-document extract)。

本研究擬達到以下三個目標：首先，將英文文件摘要的技術移植到中文文件摘要上；第二，採用潛在語意分析(Latent Semantic Analysis)來建構文件中內隱的知識模型，並以此知識模型做為摘要生成的表示法；最後，針對上述兩個目標設計適當的實驗，以比較過去文件摘要技術與本論文所提出方法間的差異性，並討論潛在語意分析(Latent Semantic Analysis)模型應用在文件摘要上的可能性。

第四節　論文架構

本論文共分為六章，第二章介紹自動化文件摘要的相關研究工作；第三章及第四章分別描述我們提出的兩種自動摘要方法：(1)以文件集為基礎的改良型摘要技術(Modified Corpus-based Approach)，(2)以潛在語意分析(Latent Semantic Analysis)為知識模型的摘要技術(Proposed LSA-based Approach)，第五章說明系統實作與實驗結果的分析討論，藉以驗證本論文所提研究方法的可行性。最後，第六章是結論與未來可行的研究方向。

第二章　相關研究工作

本章說明相關的研究工作。[Habn00]依照輔助資訊(Auxiliary Information)介入的多寡將自動文件摘要技術分為兩大類，一類為 Knowledge-rich approaches，例如[Mckeown95] [Barzilay97] [Aone99] [Azzam99] [Hovy99] [Silber00]，這類的論文著重於建構文件內容的表達模型；另一類則是 Knowledge-poor approaches，例 如 [Luhn59] [Edmundson68] [Kupiec95] [Myaeng95] [Salton97] [Aone99] [Hovy99] [Lin99] [Kim00] [Gong01]，這類論文討論直接評估文件中語句的重要性。

本章首先介紹文件摘要的相關研究，接著分別介紹三種不同觀點的研究技術：(1)以文件集為基礎的摘要方法(Corpus-based Approaches)、(2)以主題關係地圖為基礎的摘要方法(Text Relationship Map-based Approaches)與(3)以語段模型為基礎的摘要方法(Discourse-based Approaches)。

第一節　文件摘要相關研究

自動化文件摘要的研究起源於 1950 年代。受限於過去電腦技術的不發達，以及自然語言處理的高困難度，先前的研究方法僅僅著眼於計算文件中每個語句所提供的資訊量多寡或是判斷每個語句的重要性；此外，亦研究如何根據語句的重要性摘錄出足以代替原始文件的語句或段落，也就是所謂的語句(段落)摘錄[1] (Sentence/Paragraph Extraction) [Aone99] [Gong01] [Kim00] [Kupiec95] [Myaeng99] [Salton97]。

圖 2 舉例說明語句摘錄的範例，圖中的陰影部分即是範例文件的摘錄結果。摘錄類型的摘要作法是由原始文件計算每個語句的資訊量，並依照重要性的不同賦予每個語句權重；接著考慮使用者的需求(如壓縮比)，並依照語句權重挑

[1] 以下所提的文件自動摘要，所指的皆是語句或段落的摘錄。

選出候選的重要語句；最後再經過語句的排序與重組後即可作為該原始文件的摘錄。

三月四日一大早約九點出頭，前總統夫人曾文惠在女兒李安妮與隨扈的護送下，出現在台北地方法庭。在出發之前，前總統李登輝才對曾文惠表示了精神上的完全支持，但是她還是抵擋不住硬吞下眼淚的那種心情。

台灣有史以來，第一次出現前第一夫人到法院出庭的情況，曾文惠臉上沒有面對群眾時慣有的那種溫暖笑容，而是勉強擠出淺淺的笑，低著頭快速地進入法庭。只有在步出法庭時，看到熱情的支持群眾，她才露出親切溫柔的笑臉。

許多人都還記得，當然，李登輝一家人也都深深地記得。兩年前總統大選後的那幾天，許多「國民黨人士」包圍國民黨中央黨部，在民眾情緒激憤，要求李登輝下台的時候，謝啟大在宣傳車上，對著底下的群眾喊著「曾文惠帶了八千五百萬美金逃到美國」。

接下來，前立委馮滬祥以及前僑務委員戴錡更召開記者會，提出洋洋灑灑的「證據」，公開指稱曾文惠搭乘長榮航空，私運八千五百萬美元到美國，被美方拒絕入境，又緊急搭華航班機運回美元，於是引來了所謂的「八千五百萬元美金運送風波」。

小女兒李安妮不甘曾文惠被如此惡意誹謗，建議曾文會自訴謝啟大等三人涉嫌誹謗，並求償三億元賠償。但是，法官出身的謝啟大深閨司法，第一次出庭就採取反擊，反控曾文惠誣告，也要求三億元賠償，並且要求曾文惠出庭，也使得曾文惠必須在三月四日出庭應訊。

當天，曾文惠進入台北地院的北大門時，離開庭時間還有約半個小時，她快速地走上樓梯進入休息室，並準時出現在位於二樓的第七法庭。經過冗長的庭訊過程，從上午九點四十分開庭到中午一點休息，曾文惠完全沒有發言。經過短暫的休息之後，曾文惠才站在法庭前接受法官的詢問，否認運美金赴美。

在經過身體與精神的雙重煎熬之下，下午三點多，曾文惠終於承受不住心裡的委屈，趴在桌上偷偷地落淚，並在李安妮的攙扶下暫時離開法庭。在庭訊的過程中，曾文惠也不禁用紙張寫下她的心情，「上帝創造人的眼淚是流下來的，我的眼淚卻是吞進去的」。

實際上，基於對司法的尊重，曾文惠與家人也完全不願意對這件官司發表談話。而儘管曾文惠的高中校友鄭玉麗，曾經在二〇〇〇年三月二十二日下午打了通電話給她，並聊了將近半個小時，但基於自己沒有舉證責任的原則之下，曾文惠也不願鄭玉麗出面作證。

對曾文惠而言，這場官司是一種捍衛自己尊嚴的官司。看著老妻受到這麼大的委屈，李登輝心底絕對是相當心疼的。

圖 2：文件摘錄的範例

2.1.1 自動化文件摘要技術的發展

接下來介紹文件自動摘要技術的發展歷程。1950 年代到 1960 年代是文件摘要研究的開始，這個時期的研究重點著重於文件類型(Document Genre)的分析，例如：每一個段落的第一句話，通常都會直接點出接下來所要敘述的主題大綱；或是語句中出現某些常用的提示片語(Cue Phrase)—“in summary”、”in

conclusion”—等等，這些具有提示片語的語句通常是總結內容主題的說明，因此也具有高重要性。

文件摘要初期的研究，絕大多數都以分析文件類型與寫作風格的方式，以達到自動化摘要的目的。這類摘要技術的優點在於簡單容易，但這也是它最大的致命傷：摘要的方法和文件的類型與風格息息相關，導致同一技術在不同類型文件中的重複利用性不高。

1970 年代到 1980 年代初期，人工智慧的研究成果開始應用在文件自動摘要。這個時期的研究，重點在於如何建構知識的表達模型，用以辨析文件內容的主題與涵義，所使用的知識表達模型不外乎框架(Frame)及模板(Template)等。此類方法係利用自然語言的處理技巧來辨認出文件內容中人物、地點以及時間等基本要素(Entity)，並將之套用在事先定義好的模板或框架以取代原始文件中的語句，接著經由這些知識模型的推演來得知文件內容的主題並由模板來生成摘要。

此類技術的最大缺點在於模板的定義必須由專家進行，且因為模板的廣泛度不夠，使得有限數量的模板影響到文件涵義辨析的不正確性，導致產出的摘要內容在意義上的扭曲。

資訊擷取(Information Retrieval, IR)研究的議題在於如何從一文件集(Corpus)裡尋找與檢索條件有關聯的文件；若將資訊擷取的範圍縮小到單篇文件中，則文件摘要可以定義成如何在單篇文件中擷取出與內容主題相關的重要語句。

資訊擷取的技術從 1990 年代初期起大量地應用在文件摘要上，因為資訊擷取的分析著重於字層面(Word-Level)的分析，並未考慮到同義詞(Synonymy)與一詞多義(Polysemy)的詞義辨析、字詞與片語(Phrase)的辨析以及如何衡量字詞與字詞間的依屬(Term Dependency)程度等語意層面的分析，因而不能提供正確的摘要資訊。

除了上述幾種研究方法外，文件自動摘要的研究還有兩類不同的方法：以語言學(Linguistics)分析為主的摘要技術以及由認知心理學(Cognitive Psychology)來理解文件的摘要技術。它們的發展時期分別是 1960 年代到 1970 年代，以及 1970 年代到 1980 年代左右。

2.1.2 相關研究工作

第一章中曾經提到，自動化文件摘要系統的第一階段是分析原始文件，並擷取文件的特徵。究竟如何判斷所擷取特徵的重要性呢？[Habn00]根據詮釋知識(Meta-Knowledge)在特徵擷取過程中參與的程度，將使用的方法歸為以下三類：

- Knowledge-poor Approaches：
  [Luhn59] [Edmundson68] [Kupiec95] [Myaeng95] [Salton97] [Aone99] [Hovy99] [Lin99] [Kim00] [Gong01]
- Knowledge-rich Approaches：
  [Mckeown95] [Barzilay97] [Aone99] [Azzam99] [Hovy99] [Silber00]
- Hybrid Approaches：[Aone99] [Hovy99]

圖 3 中我們依年份及方法整理了這些相關的研究工作。

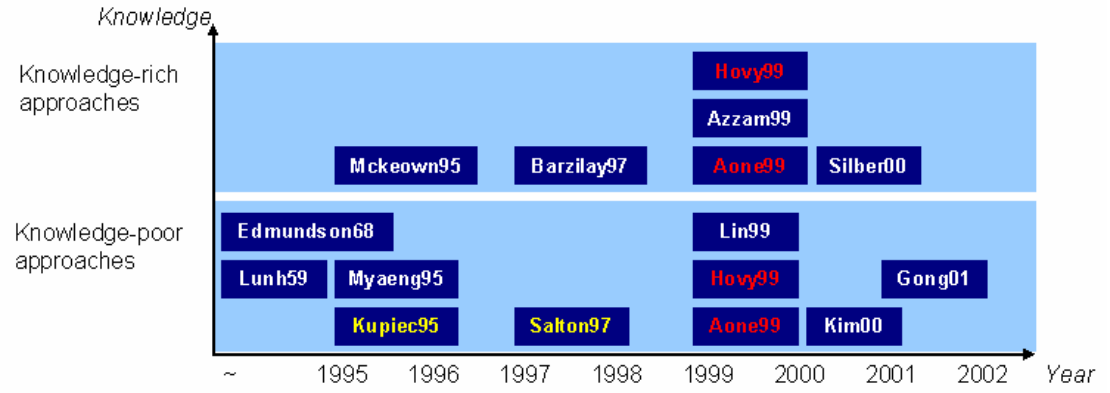

圖 3：相關研究工作

Knowledge-poor approach 是一種通用性的方法，不會因為所處理資料的不同而有所改變。其方法是擷取文件的實體特徵(Physical Features)作為分析的依據。

所謂的實體特徵可以是關鍵詞(Keyword)、語句位於文件中的位置或是提示詞語等等。這類方法通常是由資訊擷取的方法所衍生而來。

它的作法說明如下，首先分析文件中每個語句的特徵，並利用這些特徵作為語句的表示法；接著考慮特徵的重要性賦予每個語句不同的權重值(亦即代表該語句的資訊量或重要程度)；最後將文件裡的所有語句依照權重值由大而小排序，並挑選出權重值較高的數個語句成為原始文件的摘錄結果。

由上可知，利用實體特徵分析的方法，一般只著重於某些特定且較低層次的特徵分析，並沒有考慮到較高層次的語意，如知識概念(Knowledge Concepts)的分析。並且 Knowledge-poor 的方法用的是資訊擷取的技術，因此所擷取的特徵僅僅是建構在統計模型上的分析結果，並無法真正涵蓋到文件內容的意義。

為了彌補 Knowledge-poor 方法的缺陷，近年來關於文件自動摘要技術的研究已逐漸朝向 Knowledge-rich 的方法發展。所謂 Knowledge-rich 的方法除了分析文件結構與文件特徵之外，還加入領域知識(Domain Knowledge)輔助，以了解並表現出文件中所隱藏的主題和概念(Concepts)，進而達到語意層面(Semantic Level)的摘要目的。

此類方法引入額外的知識來分析文件的結構及其代表意義，以發掘出文件中包含的基本要素(Text Entity)和各個基本要素間的關聯性，從而建立文件的知識表示模型(Knowledge Representation Model)，最後精簡此模型(亦即保留此模型中具代表性的部分)，並利用精簡過後的模型來擷取文件中的語句以達到摘要的目的。

[Mani99]提出基本要素間的關聯性可能包含：

- 相似度(Similarity)：例如語彙的重複性(Vocabulary Overlap)；
- 鄰近度(Proximity)：二基本要素(如關鍵詞、人事時地物)在文件中的距

離；

- 同時出現(Co-occurrence)：基本要素是否在同一上下文(Context)中出現；
- 語彙在詞典中的關係(Thesaural Relationship)：如同義字(Synonym)、部分關係(Part-of relationship)等；
- 共同參照(Co-reference)：參照到共同的要素或者超鏈結(Hyperlink)；
- 邏輯上的相關性：如同義(Agreement)、矛盾性(Contradiction)與一致(Consistency)性等等；

舉例來說，新聞文件中的基本要素不外乎就是『人』、『事』、『時』、『地』、『物』五個要素所構成的，因此只要利用足夠的輔助知識，如人名的表格、地點的表格或是語料辭典等等，便可以辨認出該新聞文件中所存在的事件關係的模型；

有了知識模型後，更可以藉由邏輯推理來找出其中的隱性知識，最後挑選重要的知識概念用來當作文件中重要語句的擷取依據。然而此種方法最大的缺陷在於必須藉由外在知識的分析，因此可能導致字詞、語句、段落或文件層面的語意被誤解。

第二節　以文件集為基礎的摘要技術

2.2.1　以文件集為基礎的摘要技術說明

不同類型的文件，因為寫作方式及用字用詞等特性的不同，最後所產生的摘要形式也該有所差異；比如說科技論文與新聞文件的摘要在本質上就不會相同，科技論文的摘要著重於簡介(Introduction)以及結論(Conclusion)的部分，而新聞文件著重的是給閱讀者概觀性的敘述。然而，屬於同類型文件的摘要，就有可能具有某些共通的特性。

以文件集為基礎的自動摘要技術(Corpus-based Approaches)係利用機器學習(Machine Learning)，從已經具備摘要的同類型文件集中，探索出該類型文件摘要

所必備的共同特性，並應用這些共同特性於該類型文件之摘要的自動生成。圖 4 是「以文件集為基礎的自動摘要技術」之系統概觀圖。

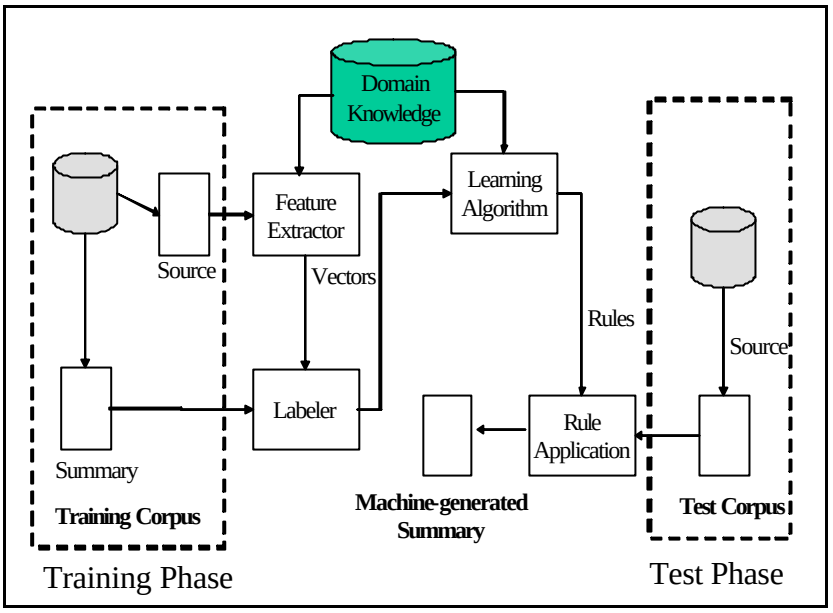

圖 4：「以文件集為基礎的自動摘要技術」系統概觀 [Kupiec95]

以文件集為基礎的自動學習摘要技術的流程分為兩個階段：(1)訓練階段(Training Phase)，(2)測試階段(Test Phase)。在訓練階段中，輸入事先由人工標示好摘要的訓練文件集(Training Corpus)，具有學習能力的摘要系統會自動從每篇訓練文件及其對應的摘要中擷取出具有代表性的特徵(Feature Extraction)；接著參考相關的領域知識，並選擇適當的學習演算法(Learning Algorithm)來產生相對應的摘要規則(Rule)。

在測試階段中，則是輸入同類型但不屬於訓練文件集的測試文件集(Test Corpus)，系統先根據學習得之摘要規則擷取出相關的特徵，並套用摘要規則產生屬於該測試文件的摘要。至於評估摘要系統優劣的方法，主要是比較系統產生的摘錄與人工標示的摘要間之準確率(Precision)和召回率(Recall)。

[Kupiec95]提出一個以貝式定理(Bayesian Rule)為基礎的 Corpus-Based 方法來計算每個語句的權重值。假設語句 s 是測試文件中的任一個語句，F_1 到 F_k 則

是系統中用來衡量語句重要性的 k 個不同的特徵，那麼語句 s 屬於摘要的機率如方程式 1：

$$P(s \in S \mid F_1, F_2, \ldots, F_k) = \frac{P(F_1, F_2, \ldots, F_k \mid s \in S) P(s \in S)}{P(F_1, F_2, \ldots, F_k)}$$

方程式 1：給予 $F_1, \ldots, F_k$ 個特徵，語句 s 屬於摘要的機率

假設每個特徵都是獨立事件的話，則方程式 1 可化簡成方程式 2：

$$P(s \in S \mid F_1, F_2, \ldots, F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

方程式 2：化簡後語句 s 屬於摘要的條件機率

$P(s \in S)$、$P(F_j \mid s \in S)$、$P(F_j)$ 是在訓練階段時由訓練文件集計算得知，其中 $P(s \in S)$ 代表訓練文件集中每個語句屬於摘要的機率，為一常數值；$P(F_j \mid s \in S)$ 代表當語句 s 屬於摘要的情形時，F_j 出現在摘要中的條件機率；$P(F_j)$ 代表訓練文件集中，特徵 F_j 的分佈機率。詳細的計算公式如方程式 3、方程式 4 和方程式 5：

$$P(F_j \mid s \in S) = \frac{\#(\text{sentence in summary, and has feature } F_j)}{\#(\text{sentence in summary})}$$

方程式 3：當 s 屬於摘要的情形下，F_j 出現在摘要中的條件機率

$$P(F_j) = \frac{\#(\text{sentence in training corpus, and has feature } F_j)}{\#(\text{sentence in training corpus})}$$

方程式 4：訓練文件集中，特徵 F_j 的分佈機率

$$P(s \in S) = \frac{\#(\text{sentence in summary})}{\#(\text{sentence in training corpus})}$$

方程式 5：訓練文件集中，摘要語句的分佈機率

[Kupiec95]所實作的系統中，用來判斷語句重要性的特徵主要為下列幾項：

- **語句長度**(Sentence Length)

語句的長度會影響到語句所涵蓋資訊量的多寡，較長的語句所包含的資訊通常比較短的語句所含的資訊量來得豐富。他們認為語句的長度至少必須要 5 **個字才可能屬於摘要。**

- **提示片語**(Fixed-Phrase)

文件中常用的提示片語，如"in summary"**以及**"in conclusion"**等等，這些片語往往會出現在介紹或總結主題敘述的語句中。他們認為文件中的語句如果包含這些常用的提示性片語，那麼該語句便有極高的可能性是屬於摘要。**

- **段落位置**(Paragraph)

他們將文件分為 paragraph-initial、paragraph-medial **以及** paragraph-final **等三個部分；並認為出現在** paragraph-initial **以及** paragraph-final **這兩個部份的語句，通常都是帶出主題或是總結主題的語句，所以，落於這兩個部份的語句具有較高的重要性。**

- **主題字詞**(Thematic Words)

一篇文件中，如果某個關鍵字重複出現許多次，則這篇文件的主題極可能與此關鍵字有關。他們認為擁有愈多出現頻率越高的關鍵詞的語句，愈有可能是屬於文件的摘要中。

- **大寫字詞**(Uppercase Words)

他們認為文件中大寫(Uppercase)**的字詞或是特殊的專有名詞**(Proper Nouns)**具有較高的重要性，因此擁有愈多大寫字詞或專有名詞的語句便愈可能屬於文件摘要。**

這篇論文中有兩個很重要的結論：

1. 雖然使用上述五個特徵當作語句重要性的計算依據，但是，實驗的結果顯示，若只考慮 Paragraph、Fix-Phrase 以及 Sentence Length 的組合所得到的結果最佳。
2. 文件摘要的壓縮比會影響到自動摘要系統結果的正確率。從圖 5 中可知，當摘要系統所摘要出來的語句數目越多的話(代表壓縮比越高)，所得到的正確率就越高。

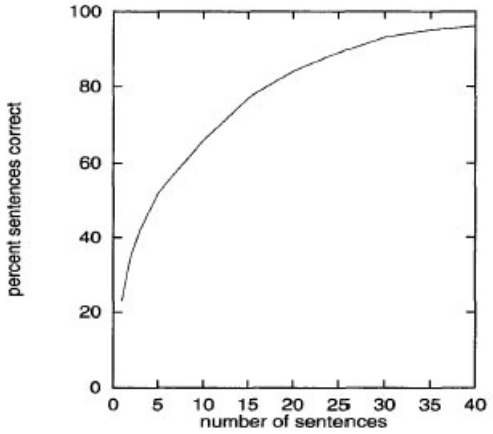

圖 5：壓縮比對摘要系統正確率的影響 [Kupiec95]

2.2.2 相關的研究成果比較

[Kupiec95]提出一個以貝式定理為核心的自動摘要方法，之後的研究都以此為中心而衍生，例如[Myaeng99]、[Aone99]與[Hovy99]。以下針對這幾篇論文的不同之處加以詳述，這幾篇論文的重點比較則列於表格 1 中。

[Myaeng99]認為文件摘要必須考慮到文件內容的架構。他們認為具有代表性的語句會出現在文件中 Introduction 及 Conclusion 這兩部分，且這兩個部分可進一步分割成四個組成結構—background, main theme, explanation of the document structure 及 future work，屬於各個部分的語句其重要性會有所差異。實驗結果顯示 Cue Word, Sentence Location 及 Resemblance to Title 最能夠代表語句的重要性。

[Aone99]**從解決資訊擷取的共通弊病來著手—語句的斷詞切字好壞會影響到摘要結果；亦即, 文件中的特殊片語或是專有名詞, 如果沒有正確地分辨的話, 很有可能會誤解文章的涵義。他們提出兩個原則來解決前述問題。第一，斷詞切字時盡量將可能是片語的字詞結合在一起；第二，利用** NameTag **工具來擷取專有名詞，並將具有相同意義的字詞視為相同，如**"IBM"**與**"International Business Machines"**在計算關鍵詞的權重時，這兩個字詞的出現頻率必須要同時考慮。**

[Hovy99]**集先前研究之大成，提出了一個重要的概念：*摘要*** *(Summarization)* ***＝主題辨認*** *(Topic Identification)+* ***概念融合*** *(Concept Fusion)+* ***摘要的生成*** *(Generation)***。亦即，輸入文件先經過主題的辨認以擷取出文件內容中描述的主題，接著將具有相同涵義的主題融合，最後再將這些主題所要表達的概念經過語句重組**(Sentence Planning)**後產生新的摘要。**

	Analysis Features	**Improvement (Compared with [Kupiec95])**	**No. of Training/ Testing Documents**	**Performance**	**Compression Rate**
[Kupiec95]	■ Sentence Length ■ Cue Phrases ■ Paragraph ■ Thematic Words ■ Uppercase Words ■ Proper Nouns	■ A statistical model based on Bayes' Rule	187/1	Recall: 42%	The same number of sentences as in the corresponding manual summary.
[Myaeng99]	■ Cue Words ■ Negative Words ■ Position ■ Theme Words ■ Centrality ■ Resemblance to Title	■ Thematic Structure Decomposition ■ Dempster-Shafer's Combination Rule ■ Use "text component" as filter	30/30	11-point average precision: 44%	5 sentences regardless of the size of source document.
[Aone99]	■ Thematic Words ■ Sentence Length ■ Position ■ Paragraph	■ To reshape the word unit ■ To acquire domain knowledge ■ To approximate text structure	100/100	Recall: 56% Precision: 51.4%	
[Hovy99]		■ Propose a new idea: Summarizat ion = Topic Identification + Interpretation + Generation ■ A method combines robust NLP and symbolic knowledge by concept fusion			

表格 1：以文件集為基礎的摘要方法研究的比較

綜合以上的說明，不難想像以文件集為基礎的摘要方法，它最大的問題在於只考慮到低層次(Low-Level)的特徵分析而已，其他較高層次的特徵，如語意索引(Semantic Index)、概念階層(Concept Hierarchy)等等語意層次的分析並沒有考慮在內。也就是說，利用這種技術來建構自動摘要系統可能導致所產生的文件摘要品質低劣，並且沒有辦法有效地涵蓋原始文件所要表達的意義。

2.2.3 以文件集為基礎的摘要技術延伸討論

以文件集為基礎的摘要技術還有一些其他的缺失，比如說 Anaphora Link 的問題等等。所謂 Anaphora Link 指的是某個語句中出現代名詞用以取代先前所提過的名詞個體，如『(1)***王老先生***有塊地。(2)***他***在這塊地上種了很多農作物。』上述語句中的『他』便是 Anaphora Link；假若摘要系統挑選了(2)當摘要，如此一來，第二句中的他便失去了原有的意義。為了解決這個問題，通常都是(1)(2)兩句一起挑選當作摘要，以保留原本 Anaphora Link 所代表的意思。

除此之外，以文件集為基礎的摘要方法，仍需要注意到以下幾點：

1. 當套用到不同寫作格式的文件集時，摘要系統該如何自動且有效地學習並發掘新的可利用特徵？
2. 當使用關鍵詞當作特徵時，摘要系統該運用何種技巧將關鍵詞層面(Term-Level)的涵義提昇到概念層面(Concept-Level)的涵義。
3. 如何利用輔助的資源如概念階層等來辨認各個關鍵詞所代表的語意。

第三節 以主題關係地圖(Text Relationship Map)為基礎的摘要技術

主題關係地圖(Text Relationship Map)由自動主題連結(Automatic Text Link)的研究延伸而來的。自動主題連結原本用在建構文件集中文件間的關聯，作法上將每篇文件以關鍵字詞的向量表示法表示，並計算所有文件兩兩間的相似度

(Similarity)；如果相似度大於系統內定的臨界值時，表示這兩篇文件具有相似的連結關係(Semantic Related Link)。依此原則可以建構出所有文件間的關係地圖。

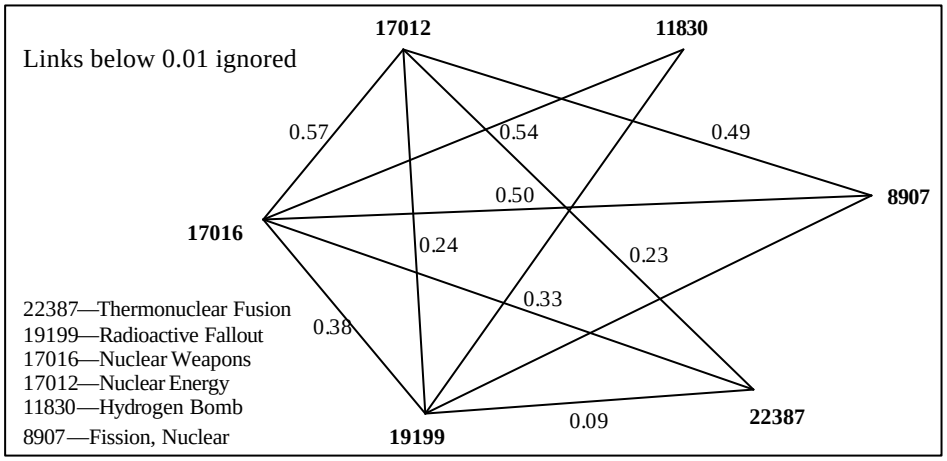

圖 6：Text Relationship Map 的範例 [Salton97]

舉例來說，圖 6 中編號 17012 及 17016 的文章，二者的相似程度約 0.57，大於臨界值 0.01，所以存在連結關係；而 8907 與 22387 這兩篇文章的相似度則因為低於臨界值，所以在 Text Relationship Map 中並沒有連結存在。具有連結的文章，可說是具有關聯性。

[Salton97]將 Text Relationship Map 的概念應用在文件摘要的研究上，並提出一個以段落(Paragraph)為摘錄單位的文件摘要系統。對於輸入的文件，以每個段落為單位計算兩兩段落間的相似度，建構 Paragraph Relationship Map。他們認為若某個段落與其他段落的連結數愈多，則代表該段落和整篇文章主題的相關性愈高。根據這個想法，連結數目愈多的段落則愈重要。

至於根據 Paragraph Relationship Map 來產生摘要，作法上分為兩個步驟。第一是判斷 Text Relationship Map 中每個段落的重要性；第二，根據 Text Relationship Map 中的連結數目來決定摘錄段落的先後順序。他們提出以下三種方法：

1. Global Bushy Path

首先定義 Text Relationship Map 上任一節點的 *Bushiness* 為該節點與其他節點間的連結數目，擁有越多關聯連結的節點，表示該段落與其他段落的寫作與用字方式相似，並且討論的主題也相似，因此，該段落視為討論文件主題的段落。Global Bushy Path 乃是將段落依照原本出現在文件中的順序以及其連結個數由大而小的排列結果。

定義 Global Bushy Path 之後，只要從 Global Bushy Path 中挑選排名最前面的 K 個段落(Top K)，即可當作該文件的摘要。此方法所摘錄出來的段落雖然涵蓋整篇文件所要表達的涵義，但是可能發生段落間語意不連續的問題，導致摘要的可閱讀性(Readability)降低；也就是說，所挑選出來的摘要中連續兩個段落雖然都是很重要的段落，但是所描述的事情可能截然不同。

2. Depth-first Path

Depth-first Path 方法可避免 Global Bushy Path 的問題。首先選取一個節點—可能是第一個節點或是具有最多連結的節點，接著每次選取在原始文件中順序與該節點最接近且與該節點相似度最高的節點當作下一個節點，依此原則選取出重要而且連續的段落以形成文件摘要。

這個方法挑選重要段落的時候也一併考慮到原始文件中的段落順序與關聯，因此可以避免類似 Global Bushy Path 的問題，同時使摘要的一致性(Coherence)與可閱讀性提高。然而，其最大的問題在於摘要內容的一致性提高，並不見得能夠涵蓋原始文件中所有主題與概念，原因乃是摘要的大小是固定的，為了要使摘要內容的連貫性提高，勢必要選取重複敘述的段落，如此便會造成篇幅的不足，而導致摘要內容的不完整。

3. Segmented Bushy Path

以上兩個方法共同的問題在於沒有考慮到文件的內容架構，舉例來說，根據文件的起承轉合，文件的內容可分為幾個不同的結構，如 Introduction、Main Them 以及 Conclusion 等等；如果套用上述的方法來挑選段落，很容易忽略掉屬於不同結構，但是重要性同樣很高的段落，最後導致摘要內容的完整性不足。Segmented Bush Path 可用來解決上述的問題。

Segmented Bushy Path 分為兩個步驟，第一個步驟是文件結構的切割 (Text Segmentation)，也就是分析文件內容並將文件內容切割成幾個具有代表的結構。Text Segmentation 利用 Paragraph Relationship Map 來分析文章的結構，圖 7 的左半部很明顯地發現 Map 上幾個節點之連結數目近乎相同，而形成可以分割的區段，其分割的結果如圖右半部，共分割成 5 個結構。

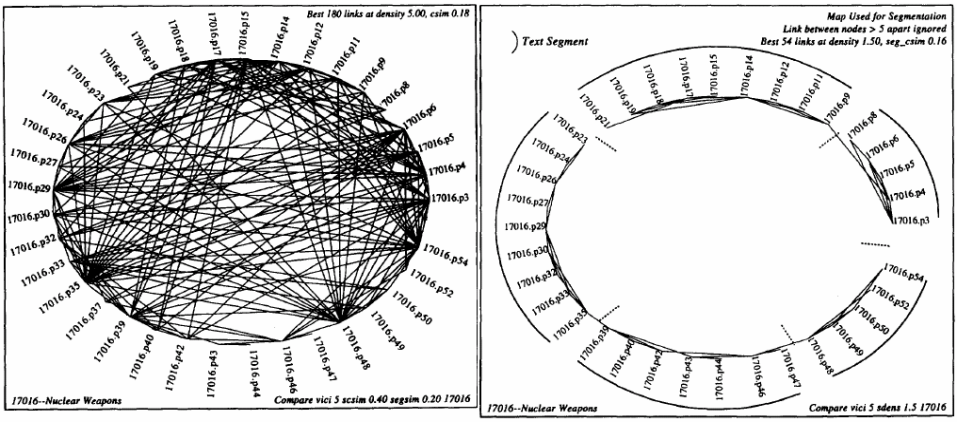

圖 7：Paragraph Relationship Map 與其對應的 Text Segmentation [Salton97]

接下來的工作便是針對每個 Segmentation 個別利用 Global Bushy Path 來選取重要的段落。為了保留每個 Segmentation 的涵義，每個 Segmentation 至少要挑選出一個段落納入最後的摘要。這樣做的好處是摘要可以涵蓋不同的主題，並使其完整性提高。

最後總結上述方法。第一，Global Bushy Path 所產生摘要的一致性最差，原因乃是挑選段落時沒有考慮到段落與段落間的連續性；第二，Depth-first Path 所

產生摘要一致性最好，對於內容的全盤涵蓋程度(Comprehension)最差，原因乃是因為所挑選到的段落集中於某幾個主題；第三，Segmented Bushy Path 的方法所產生的摘要考慮到文章內容的結構，因此對於內容的全盤涵蓋程度最好。表格 2 中整理上述三個方法的特性。

	Importance of initial paragraph	Coherence/comprehensiveness
Global bushy path	Usually starts with important early paragraph	Not coherent because adjacent paragraphs may be unrelated
Segmented bushy path	May lose important first paragraph because of need to include material from other segments	Not coherent but more comprehensive than global central path
Depth-first path	Starts with important first paragraph	Not comprehensive but more coherent than central paths, may be specialized to important subtopic

表格 2：Global Bushy Path, Depth-first Path 與 Segmented Bushy Path 的比較 [Salton97]

相對於[Salton97]只考慮到 Text Relationship Map 上每個節點的連結個數，[Kim00]認為若將每個連結的權重(語句的相似度)納入考慮，可產生更好的摘要，因此，他們提出一個以 Aggregate Similarity 計算每個語句重要性的方法。

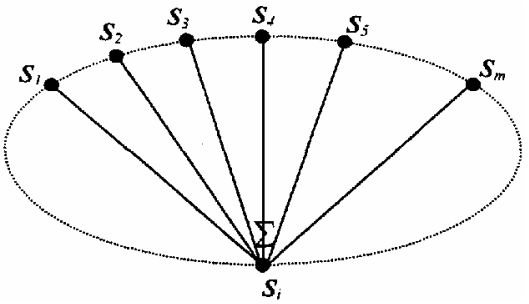

圖 8：計算 Aggregate Similarity 的概念圖示 [Kim00]

圖 8 是 Aggregate Similarity 的概念圖示。圖中的每個節點代表的是文件中某個語句的關鍵詞向量表示法，每個連結代表兩個語句間的相似度，任兩個語句的相似度即是計算相對應向量間的內積值，詳細的計算方法如方程式 6：

$$sim(i, j) = \sum_{k=1}^{n} s_{i,k} * s_{j,k}$$

方程式 6：S_i, S_j 相似度的計算方式

其中 n 表示出現在整份文件中相異的名詞個數，S_i 可以$(s_{i,1}, s_{i,2}, \ldots, s_{i,n})$表示，$s_{i,k}$ 是名詞 N_k 在語句 S_i 中出現的頻率。S_i 的 Aggregate Similarity 的計算方式如方程式 7：

$$asim(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} sim(i, j)$$

方程式 7：S_i 的 Aggregate Similarity 的計算方式 [Kim00]

對於某個節點而言，Aggregate Similarity 為此節點與其他節點之相似性的總和。計算每個語句的 Aggregate Similarity 的好處在於除了考慮到每個節點的連結個數，更考慮到每個連結的權重值。因此，Aggregate Similarity 的結果理論上會比 Global Bushy Path 的結果來得好。

第四節　以語段模型(Discourse Model)為基礎的摘要技術

認知心理學假設文件的作者在進行寫作的過程時，乃是由他本身所認知的概念空間(Conceptual Space)中去定義某個用詞的涵義，接著再組合這些定義好的詞句而成為一篇文章。當讀者閱讀文件的時候，他所作的事情便是試著去重組並建構當初該文件的作者所認知的概念空間，藉此得到相同語意的理解與認知。

[Barzilay97]以此想法為基礎，他們認為文件中所描述的概念，其實是由擁有該概念意義的所有字詞組成的結果；於是他們提出語意鏈結(Lexical Chains)的想法。所謂 Lexical Chains 即是一篇文章中相同意義的字詞所組成的集合，每個 Lexical Chain 代表這篇文件所要描述的一個概念，也就是對於這篇文件的一個認知；基於 Lexical Chains 所得到的摘要最能涵蓋該文件所要表達的意義。

作法上，首先將文件中的名詞詞彙都擷取出來，接著藉由 WordNet [24]來判斷每個字詞所代表的意義，並將具有相同詞義的字詞串接起來變成了 Lexical

Chains。美中不足的是，藉助 WordNet 的分析來建構字詞間的相似關係，可能因為其中某個字詞的意義辨認錯誤而導致產生錯誤的 Lexical Chain；如此一來，所得到的認知模型便可能偏離原文所要表達的意思。

圖 10 是圖 9 原始文件之 Lexical Chain 的視覺化示意圖。圖中清楚地看到 Mr.與 person被歸在同一個 Lexical Chain 中，這個 Lexical Chain 所表達的便是『人』這個概念；而 Machine，Micro-computer，Device 以及 Pump 則被歸屬於另外一個 Lexical Chain 中，這個 Lexical Chain 所要表達的是『機器』這個概念。我們可以發現，Lexical Chain 的確可以反映出文件中的知識概念。

Mr. Kenny is the ***person*** that invented an anesthetic ***machine*** which uses ***micro-computers*** to control the rate at which an anesthetic is pumped into the blood. Such ***machines*** are nothing new. But his ***device*** uses two ***micro-computers*** to achieve much closer monitoring of the ***pump*** feeding the anesthetic into the patient.

圖 9：原始文件範例

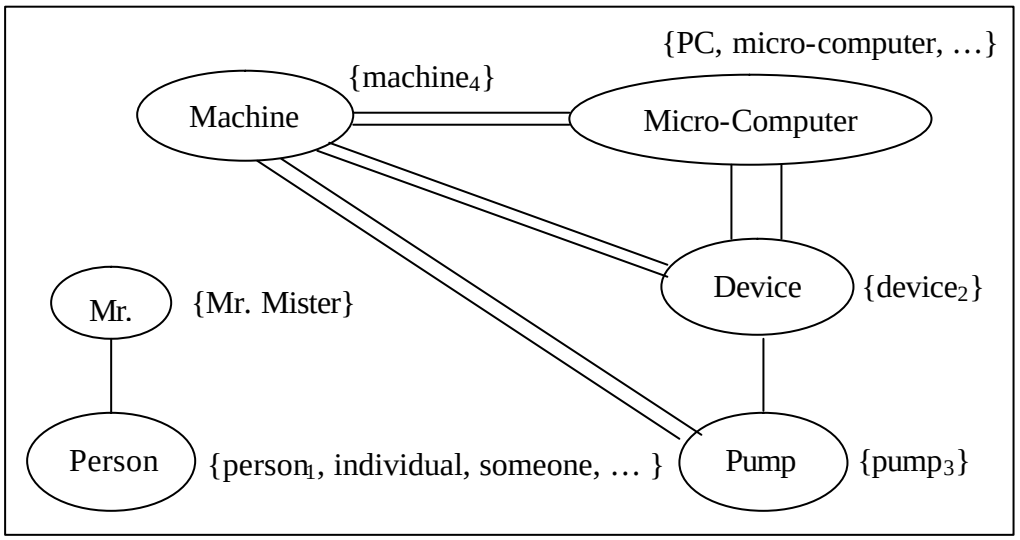

圖 10：語意鏈結的視覺化示意圖

[Azzam99]認為直接辨析文件內涵的一致性(Text Cohesion)可以更正確地認知文件。文件內涵依照[Halliday76]的定義可以由辨認下列四種關係來得到：(1)

相互參照(Co-reference)、(2)**取代與省略**(Substitution and Ellipsis)、(3)**關聯**(Conjunction)**以及**(4)**語詞相關**(Lexical Cohesion)。

[Azzam99]**利用相互參照的關係來建構文件的認知模型。其中相互參照用來描述文件中所提及的基本要素間的關係, 其與語詞相關最大的不同點在於它不像語詞相關必須藉由輔助的詮釋知識來建構關係, 而是直接分析文件中的基本要素來達到建立關係的目的。此方法最大的缺點是計算複雜度太高, 且必須要人工的介入才能正確地辨認出文件中基本要素間的關係。**

表格 3 **說明** [Barzilay97]**與**[Azzam99]**方法的特性；表格** 4 **則比較** Lexical **與** Co-reference Chain **的差異性。**

	Representation Model	**Performance**	**Comments**
[Barzilay97]	Lexical Chain	Recall: 64% Precision: 47%	The results indicate the strong potential of lexical chains as a knowledge source for sentence extraction.
[Azzam99]	Co-reference Chain	Recall: 30% Precision: 65%	The novelty is to combine the idea of a document extract based on co-reference chains with the idea of chains of related expressions serving to indicate sentences for inclusion in a generic summary.

表格 3：Characteristics of Discourse Model Approach

	Lexical Chains	**Co-reference Chain**
Differences	■ Easy to compute ■ Not rely on full text processing ■ Not always convey real "aboutness" of a text because of being indicated by an external resource	■ Require more complex techniques ■ Need to understand the meaning of texts ■ Hard to recognize relationships among objects correctly

表格 4：Lexical Chain **與** Co-reference Chain **的相異之處**

總結以上描述, 以外在知識輔助的方法所建構的知識模型並不能保證真正能夠包含文件中所敘述的意思, 原因乃是一詞多義與同義詞等語意混淆的現象, 導致了字詞的意義沒有辦法正確地被定義。此外, 使用自然語言處理的技術來分析文件, 如果沒有人工的適時介入, 便無法正確地建立基本要素關係的模型。

第三章　改良型語句權重摘要

過去以文件集為基礎的摘要技術主要應用於英文文件，若將英文摘要的研究成果應用在中文文件上，必須針對中英文的差異性加以修正，其中最大的差異在於中文斷詞比英文更容易影響到摘要結果的好壞；另外，針對過去方法的弱點，我們亦提出幾點改進：第一，針對每個語句位置的不同，我們賦予不同的權重以加強語句位置的重要性；第二利用詞彙相關程度(Word Co-occurrence)建構新字詞，並加入關鍵詞權重的計算；第三，利用基因演算法(Genetic Algorithm)得知每個特徵的重要性。

第一節　基本特徵值分析

先前提及以文件集為基礎的摘要技術，主要透過特徵來評估每個語句的重要性，本節說明我們所考慮的幾項特徵，分別為語句位置(Position)、正面關鍵詞(Positive Keyword)、負面關鍵詞(Negative Keyword)、與標題的相似度(Resemblance to the Title)以及向心性(Centrality)，以下逐一說明。

1. 語句位置(Position)

從文件內容結構的角度來看，文件中重要的語句通常出現於某幾個特定的位置；舉例來說，每一段落的第一句常常會點出該段落的主題，因此，它的重要性會比同段落中其他位置的語句還要高。

此外，即使是同樣屬於摘要的語句，我們認為他們的重要性也會因為位在文件中不同位置而有所不同；亦即，語句位置的特徵值不只是屬於或不屬於摘要的機率差別而已，每個位置都應賦予其不同的重要性。

為了加強語句位置的重要性，當人工挑選摘要語句的同時，我們也賦予每個屬於摘要的語句一個權重值；因此，概念上計算位置的特徵值，便相當

於計算某位置出現於摘要的期望權重。在實作上，對於位置的權重總共分為 6 個等級，分別為 0 到 5；其中 0 代表不屬於摘要，1 到 5 則表示該語句屬於摘要，且其重要性的強度 1 最弱，而 5 最強。

對於測試文件中的某個語句 s 來說，它的位置特徵值計算方式如方程式 8：

$$Score_{Position}(s) = P(s \in S \mid Position_i) \times \frac{Average\ weight\ of\ Position_i}{5.0}$$

where s comes from $Position_i$

方程式 8：s 的語句位置特徵值

2. **正面關鍵詞**(Positive Keyword)

資訊擷取(Information Retrieval)認為一份文件可由其含有的關鍵詞所組成的向量來表示；同樣地，對於每個語句而言，也可由其含有的關鍵詞向量來表示。基於這個想法，我們認為假若某個語句擁有越多重要的關鍵詞，那麼該語句便越可能屬於摘要。所謂的正面關鍵詞指的是常出現在摘要語句中的關鍵字詞。

考慮到中文的斷詞切字的困難程度，中文斷詞的正確與否會影響到關鍵詞出現在摘要語句的機率值；針對利用字典(Dictionary)作中文斷詞的缺點—新字詞無法辨認出來，我們應用詞彙相關程度(Word Co-occurrence) [Kowalski97]的技術來尋找文件中出現的新詞，並將找到的新詞加入計算以得到更正確的機率值。

假設 A, B, C 是三個關鍵詞組，且 C 是由 A, B 所組成的(亦即，C 為新詞)，$freq_a$ 表示 A 出現在文件集中的頻率，$freq_b$ 表示 B 出現在文件集中的頻率，

$freq_c$ 則表示 C 出現在文件集中的頻率，則 A, B 兩關鍵詞間的詞彙相關程度計算公式如方程式 9：

$$WC(A,B) = \frac{freq_c}{freq_a * freq_b}$$

方程式 9：A, B 關鍵詞的詞彙相關程度 [Kowalski97]

當 $WC(A, B)$的值大於某個臨界值時，我們認為 C 是具有意義的新詞，因而以 C 這個新詞來取代 A, B。

對於測試文件中某個語句 s 而言，假若它是由 $Keyword_1$, $Keyword_2$, …,$Keyword_n$ 所組成的，它的正面關鍵詞特徵值的計算方式如方程式 10：

$$Score_{PositiveKeyword}(s) = \sum_{k=1,2,\ldots,n} c_k \cdot P(s \in S \mid Keyword_k)$$

where c_k is the no. of $Keyword_k$ in s.

方程式 10：s 的正面關鍵詞特徵值

3. 負面關鍵詞(Negative Keyword)

相對於正面關鍵詞而言，在文件集中常出現但不屬於摘要中的關鍵詞，我們稱作負面關鍵詞。考慮這個特徵的原因，主要是希望將負面的關鍵詞特性也一併考慮，以期能夠得到更正確的結果。對於某個負面關鍵詞 $Keyword_i$ 而言，擁有 $Keyword_i$ 的語句 s 不屬於摘要的機率，計算方式如方程式 11：

$$P(s \notin S \mid Keyword_i) = \frac{P(Keyword_i \mid s \notin S)P(s \notin S)}{P(Keyword_i)}$$

方程式 11：給予負面關鍵詞 $Keyword_i$ 的條件下，s 不屬於摘要的機率

對於測試文件中某個語句 s **而言，假若它是由** $Keyword_1$, $Keyword_2$, …,$Keyword_n$ **所組成的，它的負面關鍵詞特徵值的計算方式如方程式** 12**：**

$$Score_{NegativeKeyword}(s) = \sum_{k=1,2,\ldots,n} c_k \cdot P(s \notin S \mid Keyword_k)$$

where c_k *is the no. of* $Keyword_k$ *in s.*

方程式 12**：**s **的負面關鍵詞特徵值**

4. **與標題的相似度**(Resemblance to the Title)

這個特徵主要考慮每個語句與文件標題的相似程度。一般來說，標題通常可以代表文件中所要描述的主題，因此，假若文件中的語句包含越多標題中所出現的詞彙，則該語句與文件主題的相關程度便會越高。

對於測試文件中的某個語句 s **來說，與標題相似度的特徵值的計算方式如方程式** 13**：**

$$Score_{Resemblanæ\ to\ Title}(s) = \frac{|keywords\ in\ s \cap keywords\ in\ the\ title|}{|keywords\ in\ s \cup keywords\ in\ the\ title|}$$

方程式 13**：**s **與標題的相似度特徵值**

5. **向心性**(Centrality)

文件中的語句如果越能代表該文件所要表達的意思的話，那麼該語句的重要性便會越高，這就是所謂向心性的概念。計算某個語句的向心性即是計算語句的向量表示法與整份文件扣除該語句的向量表示法兩者間的相似度。具有最大向心性的語句越能代表該文件的中心(Centroid)**，換句話說，便是最具代表性的語句。**

對於測試文件中的某個語句 s 而言，向心性特徵值計算方式如方程式 14：

$$Score_{Centrality}(s)=\frac{|keywords\ in\ s \cap keywords\ in\ other\ sentences|}{|keywords\ in\ s \cup keywords\ in\ other\ sentences|}$$

方程式 14：s 的向心性特徵值

第二節　語句權重的計算與摘要生成

第一節中說明如何從訓練文件集(Training Corpus)中計算單一語句每個特徵值。在本節中我們介紹如何將單一語句的所有特徵值結合以得到該語句的真正權重值。

首先就每個特徵來討論，我們認為每個特徵的重要程度有所不同。針對這點，我們假設語句位置(Position)的權重是 w_1，正面關鍵詞(Positive Keyword)的權重是 w_2，負面關鍵詞(Negative Keyword)的權重是 w_3，與標題的相似度(Resemblance to the Title)的權重是 w_4，而向心性(Centrality)的權重是 w_5，最後設計了如方程式 15 的 Score Function 來計算每個語句的權重值。

對於測試文件中的某個語句 s 而言，它的整體權重值的計算方式如方程式 15：

$$Score_{Overall}(s) = w_1 \cdot Score_{Position}(s) + w_2 \cdot Score_{PositiveKeyword}(s) - w_3 \cdot Score_{NegativeKeyword}(s) + w_4 \cdot Score_{Resemblance\ to\ the\ Title}(s) + w_5 \cdot Score_{Centrality}(s)$$

方程式 15：計算語句權重值的 Score Function

其中 w_1, w_2, w_3, w_4, w_5 的數值大小，我們利用基因演算法(Genetic Algorithm)來訓練以得到適合該訓練文件集的最佳 Score Function。

訓練 Score Function 的做法，我們將每組 w_1, w_2, w_3, w_4, w_5 視為基因組(genome)。每次產生 1000 個個體(Element)當作一個世代(Generation)，接著計算每個個體對於該訓練文件集的摘要正確率—以召回率(Recall)為參考標準，並保留摘要召回率最高的 10 個個體當作下一世代的母體；每一個世代評估完後，依照保留下來的 10 個個體來交配產生下一個世代的部分個體，並隨機產生其他個體以補足每個世代的個體數目。個體交配的時候，我們以下面兩個原則來產生下一世代的個體。

1. 以圖 11 為例。E1, E2 分別代表母代的基因組，產生下一世代的時候，將 E1 的基因組($M_{1,1}$, $M_{1,2}$)與 E2 的基因組($M_{2,3}$, $M_{2,4}$, $M_{2,5}$)組合成為 E3，將 E2 的基因組($M_{2,1}$, $M_{2,2}$)與 E1 的基因組($M_{1,3}$, $M_{1,4}$, $M_{1,5}$)組合成為 E4，這樣的交配方法我們將之稱為"2-3 基因交換"。依照這個原則，我們實作了 1-4, 2-3, 3-2, 4-1 四種交配的方法。

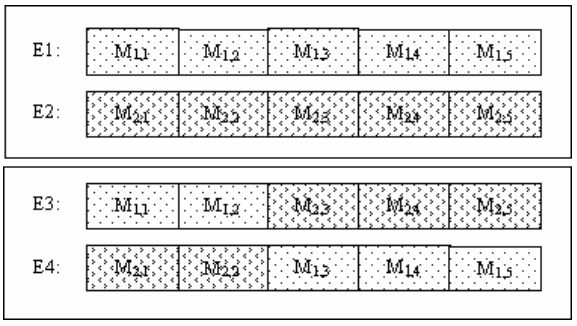

圖 11：m-n 基因交配方法

2. 為了增加基因的突變能力，以圖 12 為例，以隨機的方式保留下 E1 中的 $M_{1,2}$，$M_{1,4}$，$M_{1,5}$ 作為 E2 的基因(每次所保留的基因不同)，另外，E2 中的 $M_{1,1}$ 與 $M_{1,3}$ 則由系統隨機產生，便可以保留下部分優良的基因，以增加世代的突變能力。

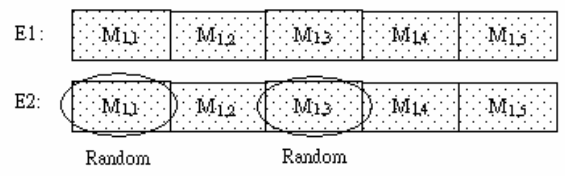

圖 12：個體的基因組($M_{1,1}, M_{1,2}, M_{1,3}, M_{1,4}, M_{1,5}$)與其突變

利用基因演算法的訓練方式，我們可以找到一組最適合該訓練文件集的特徵權重值。這樣的作法乃是因為對於不同的訓練文件集，很難有效地找到最適合的特徵權重值，然而，套用基因演算法的方式可以幫助找到一組適當的解，做為系統設計者調整系統好壞的依據。

值得一提的是，我們將基因演算法套用於訓練文件集上，因此對於測試文件集(Test Corpus)並不能保證所得到的權重值組也能得到相同的好結果。但是，假如測試文件集與訓練文件集的性質非常接近的話，此方法的結果與實際上真正適合該測試文件集的權重值組所計算出來的正確率並不會相差太多。

生成文件摘要的部份，以一篇測試文件而言，首先根據方程式 15 計算每個語句的整體權重值，當成是每個語句的分數(Score)，接著依據語句的分數將文件中的語句依分數由大至小的方式作排名(Ranking)，最後將 Top N 個語句擷取出來當作該文件的摘要結果。綜合上述，我們將此摘要的方法加以整理於圖 13。

Corpus-based 文件摘要生成演算法

1 針對測試文件中的每個語句依下列步驟計算它的權重值。
  1.1 利用方程式 8、方程式 10、方程式 12、方程式 13、方程式 14 計算每個特徵值的大小。
  1.2 利用方程式 15 計算該語句的整體權重大小。
2 將文件中的所有語句，依照語句的分數由大排到小形成一份語句重要性排名清單(Sentence Importance List)。
3 根據壓縮比(Compression Rate)計算欲摘要的語句個數 N。
4 從語句重要性排名清單中挑選出前面的 N 個語句即為該文件的摘要。

圖 13：Corpus-based 文件摘要生成演算法

總結來說，我們以[Kupiec95]**為本，提出了三項改進的方法：**

1. **引入權重的概念應用在語句位置的重要性計算上，以期得到更正確的語句位置特徵值的計算。**
2. **根據中文斷詞切字的問題，利用詞彙相關程度的技術來找到文件集中的新詞，以期能夠改進與關鍵詞相關的特徵值(正面關鍵詞、負面關鍵詞、與標題的相似度及向心性)的計算結果。**
3. **使用基因演算法來訓練** Score Function **中的** w_1, w_2, w_3, w_4 **以及** w_5**，以期能夠提供系統設計者調整** Score Function **的依據。**

第四章　以潛在語意分析為基礎的語句摘要

受到資訊擷取中潛在語意分析(Latent Semantic Analysis, LSA)與潛在語意索引(Latent Semantic Indexing, LSI)研究成果的影響，我們認為 LSA 可以正確地描述文件中字詞、語句與文件意義的整體關聯性；因此，本論文將 LSA 應用在自動摘要的研究上，並提出一套以 LSA 為知識核心的摘要方法。

本章中，首先介紹何謂 LSA 及其工作原理，接著說明本論文所提的以 LSA 為基礎的語句摘要技術的系統架構，最後描述我們的方法與相關討論。

第一節　潛在語意分析(Latent Semantic Analysis)

4.1.1　LSA 工作原理

[Landauer98]認為 LSA 除可作為文件的知識表示(Knowledge Representation)外，並可用來推演隱性的知識關聯；此外，LSA 的知識模型與知識推演過程接近於人腦用來理解文件知識的推演與認知機制模型。

LSA 是以數學統計為基礎的知識模型，其運作方式跟類神經網路(Neural Net)的極為相似，不同的是類神經網路以權重的傳遞與回饋來修正本身的學習，LSA 則以奇異值分解(Singular Value Decomposition, SVD)與維度約化(Dimension Reduction)為核心作為邏輯推演的方式。

LSA 的應用非常廣泛，主要集中在資訊擷取、同義詞建構、字詞與文句的相關性判斷標準、文件品質優劣的判別標準及文件理解與預測等各方面的研究。

LSA 的工作原理如圖 14 所示：利用 SVD 及維度約化將輸入的知識模型抽象化，整個過程除可以將隱含的語意顯現出來外，更能將原本輸入的知識模型提升到較高層次的語意層面。

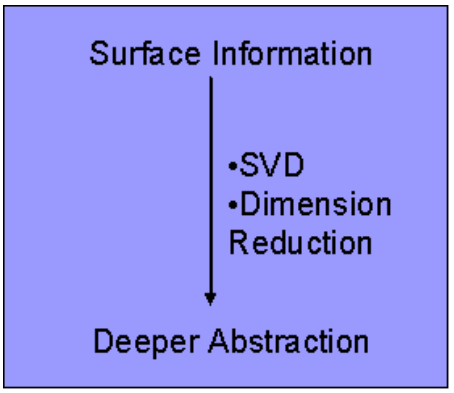

圖 14：LSA 工作原理示意圖

實際運作的過程中，首先將文件集(Corpus)中所有文件的 Context [2] 用 Word-by-Context 矩陣 M 來表示，矩陣中的每個元素即是某關鍵詞在某 Context 中的重要性或出現頻率。接著，將矩陣 M 經過 SVD 分解轉換得成新的矩陣乘積 LSU^T，亦即 $M=LSU^T$，其中 S 代表語意空間(Semantic Space)，L 代表關鍵詞在此語意空間中的表示法，U^T 則代表 Context 在此語意空間中的表示法。LSA 利用維度約化可更精確地描述語意空間的維度，並重建矩陣 M　$M'=L'S'U'^T$，更明確地探究出 Word-Word、Word-Context 或 Context-Context 間的關聯性。

總結上述說明：

1. LSA 假設經過 SVD 後所得到的對角線矩陣(即上述中的 S 與 S')所代表的意義是整份文件的語意空間。所謂的語意空間就是文件中每個字詞的定義空間，也就是說，每個字詞可以透過這個語意空間的定位來得到真正代表的意思。
2. 為了要將語意空間的真正維度定義出來，LSA 需要經過維度約化來重建最後的 Word-by-Sentence 矩陣。
3. M 經過 SVD 分解與維度約化後重建得到的新矩陣 M'中，S'代表語意空間，此語意空間比 S 可以更正確地定義且描述關鍵詞與 Context 所代表

[2] 所謂 Context 可視需求定義為 Sentence, Paragraph, Chapter, 或 Document 的層面來考量。

的意義。

4. **相較於使用外在資源以達到文件模型建構的方法，**LSA **提供直接的分析方式，更精確地建構文件的知識模型，且避免使用輔助知識可能發生的語意混淆的問題。**
5. LSA **與資訊擷取的不同在於** LSA **可以涵蓋字詞間關聯程度**(Co-occurrence)**，更可藉由維度約化將原** Context **中潛在的語意表現出來。**
6. LSA **具有知識推演的能力，如果將最原始矩陣中的任一個數值改變後，其結果會影響到最後重建的矩陣, 且影響的範圍不只是原先經過改變的數值，更可能影響到矩陣中的其他數值。**

4.1.2 LSA **實例說明**

接下來，我們以實例說明 LSA **的運作方式** [Landauer98]**。這個例子中共包含** 9 **個** Context**，分別為** c1、c2、c3、c4、c5、m1、m2、m3 **與** m4**，其中** c1 **至** c5 **是** Human-Computer Interface **領域的相關文件標題，而** m1 **至** m4 **則來自於** Mathematical Graph Theory **領域的相關文件標題。**

Exampl of text data: Titles of Some Technical Memos

[Human-Computer Interface]
c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of user perceived response time to error measurement

[Mathematical Graph Theory]
m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

我們挑選至少出現兩次的關鍵詞來建構 Word-by-Title **的矩陣***{X}*，*{X}***中的每一列**(Row)**代表在兩個或兩個以上的** Context **中出現過的關鍵詞，而每一行**(Column)**則代表一個** Context **；此外，***{X}***中每個元素代表特定關鍵詞在特定** Context **中出現的次數。***{X}***經過** SVD **分解後得到三個矩陣，分別為***{W}*,*{S}***以及** $\{P\}^T$**。其中***{X}***即是先前所說的** M**，另外***{W}***、***{S}***與** $\{P\}^T$ **分別為前面所說的的** L**、** S **與** U^T**。**

*{X'}***則是維度約化過程中取維度**(Dimension)**為** 2**，亦即取***{W}***、***{S}***與***{P}***的前二** Column(**相當於把其他** Column **的值均設為** 0)**後所重建回來的矩陣—***{X'}={W'}{S'}{P'}*^{T}(*{W'}, {S'}, {P'}***為將***{W}, {S}, {P}***取前二** Column **值，其餘** Column **值設為** 0 **的結果)。**

$$\{X\}=\begin{bmatrix} & c1 & c2 & c3 & c4 & c5 & m1 & m2 & m3 & m4 \\ human & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ interface & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ computer & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ user & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ system & 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ response & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ time & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ EPS & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ survey & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ trees & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ graph & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ minors & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \{W\}\{S\}\{P\}^T$$

$$\{W\}=\begin{bmatrix}
0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\
0.20 & -0.07 & 0.14 & -0.55 & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\
0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\
0.40 & 0.06 & -0.34 & 0.10 & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\
0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.30 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\
0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\
0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\
0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & -0.68 & 0.23 \\
0.03 & 0.45 & 0.14 & -0.01 & -0.30 & 0.28 & 0.34 & 0.68 & 0.18
\end{bmatrix}$$

$$\{S\}=\begin{bmatrix}
3.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2.35 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.50 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.31 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.85 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.56 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.36
\end{bmatrix}$$

$$\{P\}=\begin{bmatrix}
0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.01 & 0.02 & 0.08 \\
-0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \\
0.11 & -0.50 & 0.21 & 0.57 & -0.51 & 0.10 & 0.19 & 0.25 & 0.08 \\
-0.95 & -0.03 & 0.04 & 0.27 & 0.15 & 0.02 & 0.02 & 0.01 & -0.03 \\
0.05 & -0.21 & 0.38 & -0.21 & 0.33 & 0.39 & 0.35 & 0.15 & -0.60 \\
-0.08 & -0.26 & 0.72 & -0.37 & 0.03 & -0.30 & -0.21 & 0.00 & 0.36 \\
0.18 & -0.43 & -0.24 & 0.26 & 0.67 & -0.34 & -0.15 & 0.25 & 0.04 \\
-0.01 & 0.05 & 0.01 & -0.02 & -0.06 & 0.45 & -0.76 & 0.45 & -0.07 \\
-0.06 & 0.24 & 0.02 & -0.08 & -0.26 & -0.62 & 0.02 & 0.52 & -0.45
\end{bmatrix}$$

$$\{X'\}=\begin{bmatrix} & c1 & c2 & c3 & c4 & c5 & m1 & m2 & m3 & m4 \\ human & 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ interface & 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ computer & 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ user & 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ system & 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ response & 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ time & 0.16 & 1.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ EPS & 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ survey & 0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\ trees & -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\ graph & -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ minors & -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{bmatrix}=\{W'\}\{S'\}\{P'\}^T$$

比較*{X}*與*{X'}*在 m4 **中** ***survey*** **與** ***trees*** **的值。我們發現** ***trees*** **原先並未出現在** m4 **中，但** m4 **包含了** ***graph*** **及** ***minors*****，且這兩個詞與** ***trees*** **在** Graph Theory **領域中有語意的相關性；因此，** ***trees*** **的值原先在*{X}*中為** 0**，但是經過維度約化及矩陣重建後，它的值變成了*{X'}*中的** 0.66**，這個數值的象徵意義代表了** ***trees*** **在無限多篇** Graph Theory **領域文件的標題中可能會出現機率。**

相對地， ***survey*** **的值由*{X}*中的** 1 **降為*{X'}*中的** 0.42**，這代表** ***survey*** **出現在** Mathematical Graph Theory **這個領域中不具有特別的重大意義(直覺上來想，** ***survey*** **在任何領域中出現的機率應該是幾乎均等的)。**

假若將 Context **轉換成語意空間的表示法，便可以應用在許多領域。例如，計算*{X'}*中列向量(**Row-Vector**)—即關鍵詞在** Context **中的"出現機率"或"重要性"—的內積值(**Inner-Product**)，便可以推斷出關鍵詞間的語意相關程度；計算*{X'}*中行向量(**Column-Vector**)—即** Context **由關鍵詞所組成的整體資訊—的內積值，便可以推斷任意兩個** Context **的語意相關程度。**

第二節　系統架構

本論文整合潛在語意分析(LSA)及第二章中所提及的主題關係地圖(Text Relationship Map)概念，期望達到以語意為基礎的自動文件摘要。整個文件摘要系統分為四個模組，分別為(1) 前置處理，(2) LSA 語意模型建構，(3) Text Relationship Map 建構，以及(4) 語句選取等， 圖 15 為系統架構示意圖。

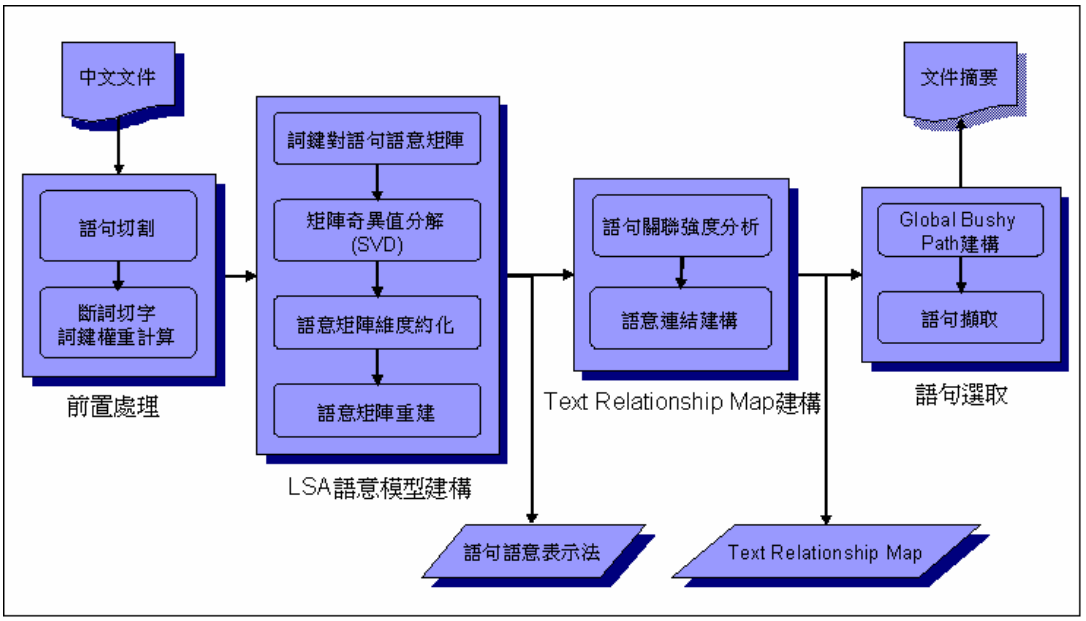

圖 15：LSA 文件摘要系統架構

前置處理將中文文件依照段落及語句切割，判斷語句結束是依據標點符號『。』、『？』以及『！』作為語句的結束。斷詞切字的部分則利用中央研究院所研發的『CKIP 中文自動斷詞系統』 [23]來處理，並計算每個關鍵詞的出現頻率與權重值。CKIP 中文自動斷詞系統除了具有中文自動斷詞的功能外，更可以標示每個字詞的中文詞類；同時，CKIP 允許使用者根據自己的需求選擇不同的詞典，作為斷詞與及標記的參考。

LSA 語意模型建構將前置處理所得到的關鍵詞與語句建構成 Word-by-Sentence 的矩陣(亦即每個 Sentence 為一 Context)，接著將該矩陣作奇異值分解(SVD)，再經過矩陣維度約化(Dimension Reduction)，最後重建具有語意的語意矩陣(Semantic Matrix)。這個模組將隱藏在文章中的語意透過 LSA 表現出

來，其產出是詞鍵的語意向量表示法—列向量(Row-Vector)以及語句的語意向量表示法—行向量(Column-Vector)。

Text Relationship Map 建構以 LSA 語意模型建構步驟的產出—語句的語意表示向量—為基礎，計算兩兩語句間的相似度，以分析語句間的連結關係，並建構主題關係地圖。

語句選取步驟則由主題關係地圖中的連結來衡量每個節點的重要性，並且依照每個節點的重要程度建構 Global Bushy Path [Salton97]；最後，摘要的生成係由 Global Bushy Path 中挑選語句當成摘要的產出。

第三節　語句分群與摘要生成

本節中說明如何將潛在語意分析(LSA)的概念應用於以主題關係地圖(Text Relationship)為基礎的語句摘錄系統上。LSA 除考慮到文件中關鍵字的相關程度之外，更能將關鍵字之間的語意關係顯現出來；文件中的每個語句透過 LSA 可更精確地定義出語句涵義與相互間的關聯，因此可達到更好的摘要結果。

作法上以一篇文件為單位建構 Word-by-Sentence 矩陣，假設此文件為 D，文件中擁有 N 個語句，且重要的關鍵詞有 M 個，最後會得到如圖 16 的矩陣，其中 W_i 代表文件中的關鍵詞，S_i 代表文件中的語句。

	S_1	S_2	⋯	S_N
W_1	K_{11}	K_{12}	⋯	K_{1N}
W_2	K_{21}	K_{22}	⋯	K_{2N}
⋮	⋮	⋮	⋱	⋮
W_M	K_{M1}	K_{M2}	⋯	K_{MN}

圖 16：Word-by-Sentence 矩陣範例

4.3.1 關鍵詞的選取

文件中並非所有的關鍵詞都具有同樣的重要性，一般來說，名詞與動詞的重要性就比冠詞、副詞或是介系詞的重要性高很多；加上每個語句都是由詞彙所組成的，因此，假如建構矩陣的詞彙選擇不好的話，LSA **的結果便會受到影響。**

由於大多數語句都是"主詞-述詞-受詞"的結構 [陳光華 98**]，且文件裡的主詞與受詞往往就是名詞，而述詞往往是動詞。因此，對於每個語句來說，便可以單由名詞和動詞來理解其語意。**

據了解(Dk)**，國台辦**(Nc)**高層**(Na)**官員**(Na)**還**(D)**曾**(D)**設宴**(VA)**款待**(VC)**該**(D)**參訪團**(Na)**成員**(Na)**。**

上述例子中，屬於名詞的關鍵字有『國台辦』、『高層』、『官員』、『參訪團』、『成員』，屬於動詞的關鍵字有『設宴』、『款待』；如果保留這些關鍵詞，依舊可以推測整個語句所要表達的意思。因此，我們只保留名詞及動詞作為建構 Word-by-Sentence **矩陣的詞彙。**

4.3.2 矩陣中數值的計算方式

為了精確地掌握每個關鍵詞的重要性，除了計算每個關鍵詞在每個語句中出現的頻率之外，我們亦考慮每個關鍵詞於整份文件中的重要程度。因此，圖 16 **中的每個** K_{ij} **的計算方式便如方程式** 16**：**

$$K_{ij} = G_i * L_{ij}$$

方程式 16**：**K_{ij} **的計算公式**

其中 G_i 代表關鍵詞 W_i 於 D 中的分佈權重，L_{ij} 代表 W_i 在 S_j 中的分佈權重。假設 c_{ij} 為 W_i 出現在 S_j 中的次數，t_j 為 W_i 出現在 D 中的次數，則 W_i 在 S_j 中的相對頻率計算方式如方程式 17：

$$f_{ij} = \frac{c_{ij}}{t_i}$$

方程式 17：W_i 於 S_i 中的相對頻率 f_{ij} [Bellegarda96]

此外，考慮文件 D 中 W_i 的資訊分佈量(Entropy)計算方式如方程式 18：

$$E_i = -\frac{1}{\log(N)} \sum_{j=1}^{N} f_{ij} * \log(f_{ij})$$

方程式 18：W_i 於 D 中的資訊分佈值 [Bellegarda96]

由方程式 18 可知當 f_{ij} 等於 1 的時候，E_i 的值為 0；當 f_{ij} 等於 $1/N$ 的時候，E_i 的值為 1。當 E_i 的值越接近於 1 的時候，表示 W_i 在文件 D 中的分佈越平均，W_i 的重要性便會降低；相反地，如果 E_i 的值越接近 0 的時候，表示 W_i 只出現在某些語句中而已，W_i 的重要性便比平均分布在文件 D 的關鍵字來得高。舉例來說，如果文件 D 是討論資料庫系統效能的文章，因為文件內容中常常提到『資料庫』這個詞語，『資料庫』在整份文件中的分佈便會非常地平均，因此，它的重要性就會降低。

定義了 E_i 之後，我們定義 W_i 於 S_j 中的總體權重 G_i 如方程式 19：

$$G_i = 1 - E_i$$

方程式 19：W_i 於 S_j 中的總體權重 G_i [Bellegarda96]

此外，定義 W_i 於 S_j 中的權重 L_{ij} 如方程式 20：

$$Li_j = \log_2\left(1+\frac{c_{ij}}{n_j}\right)$$

方程式 20：W_i **於** S_j **中的權重** L_{ij} [Bellegarda96]

其中 n_j **代表** S_j **中所含的關鍵詞總數。**

接著建構 Word-by-Sentence **的矩陣。假設該矩陣為** A**，接下來將矩陣** A **作奇異值分解**(SVD)**使得** $A=USV^T$**。對於** S**，經過維度約化**(Dimension Reduction)**取適當的維度後重新建構矩陣** $A'=U'S'V'^T$**；此時, 便得到具有語意的** Word-by-Sentence **矩陣表示法，其中，每個列向量**(Row-Vector)**代表了該關鍵詞在每個語句中的權重，而每個行向量**(Column-Vector)**代表該語句由各個關鍵字所組成的意義。**

4.3.3 摘要的生成

由於 LSA **可以將文章中的隱性語意表現出來，因此，若以** LSA **產生的語句表示方式來計算語句間的相似度, 其結果會比單純使用關鍵字出現頻率權重的表示法來得好。基於這個想法，我們將** LSA **所得到的語句表示式—行向量**(Column Vector)**套用在主題相關地圖**(Text Relationship Map)**上，並衡量** LSA **對於摘要結果的影響。**

接下來計算每個語句的相似度，並建構主題相關地圖。我們以 LSA **重建之後得到的行向量當作語句的表示法，並計算兩向量間的** Cosine **值來衡量計算語句間的相似度。建構主題相關地圖時，只保留約** 1.5 **倍語句數目的連結；亦即該文件中若有** n **個語句的話，那麼總共的連結數目會是** C(n,2)**個，而最後只保留相似度高的前** 1.5*n **個連結。**

我們採用 Global Bushy Path [Salton97]**來產生摘要，統計主題相關地圖上每個節點的連結數目, 依照每個語句在原始文件中的先後順序以及每個語句所擁有**

的連結數目由大而小排列；最後，挑選排名前面的 Top K **個語句即是該文件的摘要。綜合上述，我們將此摘要的方法加以整理並列出於圖** 17。

LSA 文件摘要生成演算法

1 **中文文件前置處理**
  1.1 **語句切割**
  1.2 **斷詞切字與詞鍵權重計算**
2 LSA **語意模型建構**
  2.1 Word-by-Sentence **語意矩陣建構**
  2.2 **語意矩陣的奇異值分解**(SVD)
  2.3 **語意矩陣維度約化**(Dimension Reduction)
  2.4 **語意矩陣重建：語句與詞鍵語意知識模型的形成**
3 **利用** 2 **中所得到的語句語意表示法計算兩兩語句間的相似度** $sim(s_i,s_j)$**，並保留** 1.5n **的連結數以建構** Text Relationship Map
4 **排序** Text Relationship Map **上的節點並形成** Global Bushy Path
5 **由** Global Bushy Path **上取** Top K **個節點語句當作最後摘要的集合**

圖 17：LSA **文件摘要生成演算法**

第五章　實驗結果分析與評估

本章闡述第三、第四章所提出方法的實驗結果及相關討論。第一節介紹實驗文件集的特性，第二節說明評估系統優劣的方法，第三節討論改良型語句權重摘要系統的效益評估，第四節討論以潛在語意分析語句摘要之可行性評估。

第一節　實驗資料說明

我們從新台灣新聞週刊(New Taiwan Weekly)雜誌 [31]的政壇話題中收集了 100 份新聞文件，並將這 100 份文件分成 5 個大小相同的集合，分別為 Set 1、Set 2、Set 3、Set 4 及 Set 5，平均每個集合中有 20 份文件，每份文件的平均長度為 27.4 個語句。

因為所收集到的新聞文件並沒有相對應的摘要，因此，每份文件都經由人工挑選重要的語句當成該份文件的摘要，用來評估摘要系統的好壞；實驗文件集的統計結果，平均每份文件中由人工所挑選出來的摘要語句約為 8.7 句，平均摘要壓縮比為 31.8%。表格 5 中列出前述新聞文件集的特性。

	Set 1	Set 2	Set 3	Set 4	Set 5
文件篇數	20	20	20	20	20
每篇文件的平均語句數目	27.5	24.8	26.7	31.5	26.4
每篇文件的平均人工摘錄語句數目	8.8	8.0	8.5	9.8	8.4
每篇文件的平均人工摘錄壓縮比	32%	32%	32%	31%	32%

表格 5：實驗文件集的統計特性

第二節　評估方法

[Mani99]提到評估自動文件摘要系統優劣的方法，依照摘要本身的品質來看，可分為兩種不同的評估方式，分別為外在評估(Extrinsic-Level Evaluation)及內在評估(Intrinsic-Level Evaluation)。

外在評估將摘要結果當成其他資訊系統的輸入，並衡量摘要的品質對於其他資訊系統整體表現的影響程度。舉例來說，自動摘要結果可以當成文件分類系統(Document Classification System)的輸入，假如摘要結果可以反映原始文件所要表達的涵義，那麼以摘要結果作為分類判斷標準的系統也會有相當的表現。

內在評估則比較自動摘要與人工摘要結果品質的好壞，亦即，比較自動摘要結果與人工摘要結果間的相似程度；此外，也可經由人工的直接閱讀，主觀地判斷自動摘要結果的好壞。舉例來說，對於產生摘錄(Extract)的資訊系統而言，主要以自動產生的摘錄與人工產生的摘錄來比較其準確率(Precision)和召回率(Recall)。

本論文所採用的評估方法屬於內在評估。考慮一篇文件 D，假設人工摘錄的語句集合是 A，自動摘要系統摘錄的語句集合是 B，則精確率與召回率的定義如方程式 21 與方程式 22。

$$Precision = \frac{|A \cap B|}{|B|}$$

方程式 21：自動摘要系統的精確率評估

$$Recall = \frac{|A \cap B|}{|A|}$$

方程式 22：自動摘要系統的召回率評估

本篇論文中，自動摘要與人工摘要所挑選的語句數目皆固定壓縮比為 30%左右。因此，上述兩個方程式中，|A|和|B|是相同的。也就是說，在這樣的條件下所得到的準確率與召回率大小是一樣的。以下的討論中，我們只評估召回率而不另外列出準確率的數值。

第三節　改良型語句權重摘要之效益評估

5.3.1 K. Cross-Validation

機器學習的方法可以分為訓練(Training)及測試(Test)兩階段，而評估機器學習方法的好壞則是利用異於訓練資料的新資料來作測試，以真正測出系統學習能力的好壞。常見的評估方法主要有 Holdout 及 K. Cross-Validation。在本實驗中，我們以 K. Cross-Validation 來驗證摘要系統的好壞。

首先解釋 K. Cross-Validation [Han01]？一般而言，當資料集的大小有限時，常常會利用 K. Cross-Validation 來交錯驗證，以期得到較精準的評估結果。作法上將取得的資料集切割成 K 等份，每次取其中一個集合當作測試資料，並以其他 K-1 個集合當作訓練資料。這樣作的好處在於，當取得的資料不足時，依舊可以驗證出系統的好壞；此外，也可將每個資料集合的特性融合，以得到較一般化且具說服力的測試結果。

在訓練階段的時候，每次選取 4 個集合當作訓練用的文件集，剩下的 1 個集合則用來測試訓練結果的好壞。舉例來說，如果測試的集合為 Set 5，那麼便拿 Set 1、Set 2、Set 3 及 Set 4 來訓練 Score Function。

5.3.2 特徵值的影響探討

為了解每個特徵對於摘要結果的影響，首先針對每個特徵來進行實驗，表格 6 到表格 10 列出各個特徵的影響數據，表中 Original 代表未經改良過的方法，Modified 代表我們提出的方法。從這幾個表格中可以發現，語句位置(Position)、正面關鍵詞(Positive Keyword)、與標題的相似度(Resemblance to the Title)以及向心性(Centrality)對摘要系統的影響較為重要，而負面關鍵詞(Negative Keyword)對於摘要系統的影響不是很重要，甚至有可能導致不好的結果。

首先以語句位置(Position)**來說，由表格** 6 **可知** Modified **比** Original **表現來得好；此實驗結果反映出先前的假設：存在有重要資訊的語句位置，其重要程度會因為所在位置不同而有改變，並不是單純考量是否屬於摘要的機率而已。**

	Original	**Modified**
Set 1	0.4415	0.4788
Set 2	0.4639	0.4924
Set 3	0.4648	0.4844
Set 4	0.4796	0.4955
Set 5	0.4286	0.4632
Average	0.4557	0.4829

表格 6**：考慮語句位置特徵時語句摘錄的召回率**

以正面關鍵詞(Positive Keyword)**來說，關鍵詞是組成語句的重要因素，因此，假若某個語句擁有越多重要的關鍵詞，該語句便越有可能屬於摘要中。實驗的結果反映出單純考慮正面關鍵詞時，亦可得到不錯的結果。**

對於負面關鍵詞(Negative Keyword)**來說，實驗結果說明當考慮負面關鍵詞時，對於摘要結果的影響並不大。我們推測可能因為某些語句特別長，若其含有較多的負面關鍵詞時，會導致累計結果造成該語句的重要性降低，這個現象與長度越長的語句，其重要性越高的假設相矛盾；因此，負面關鍵詞的影響力便會有所偏差，而導致結果不好。**

	Original	**Modified**
Set 1	0.4647	0.4839
Set 2	0.3648	0.3865
Set 3	0.4381	0.4190
Set 4	0.4912	0.5030
Set 5	0.5399	0.5410
Average	0.4597	0.4667

表格 7**：考慮正面關鍵詞特徵時語句摘錄的召回率**

	Original	Modified
Set 1	0.1982	0.1936
Set 2	0.2972	0.2771
Set 3	0.2301	0.2000
Set 4	0.1746	0.1826
Set 5	0.1739	0.1800
Average	0.2148	0.2067

表格 8：考慮負面關鍵詞特徵時語句摘錄的召回率

對與標題的相似度(Resemblance to the Title)**而言，文件的標題通常是文件主題的縮影；因此，與標題相似度越高的語句，該語句的重要性也就會越高。對向心性**(Centrality)**來說，文件的內容通常都是都是圍繞幾個重要的主題陳述，如果文件中語句的向心性越高的話，該語句與整體文件內容所要表達的主題意義便會越相近，它的重要性也就跟著提高。表格 9 及表格 10 證實了與標題的相似度及向心性的確是很重要的特徵。**

對於中文斷詞切字不明確的問題，我們所加入的新詞的確影響到摘要結果的好壞。由正面關鍵詞、負面關鍵詞、與標題的相似度及向心性來看，加入新詞計算的結果皆比原本的結果來得好，尤其以與標題的相似度及向心性較為明顯；但是，因為整個系統中利用詞彙相關程度(Word Co-occurrence)**所找出來的新詞大約只有 1600 個左右，整體表現並非很明顯，雖有進步，可是進步並不多。最後，表格 11 中列出系統中利用詞彙相關程度所找到的部分新詞以供參考。**

	Original	Modified
Set 1	0.4274	0.4370
Set 2	0.4217	0.4487
Set 3	0.3644	0.3716
Set 4	0.4557	0.4628
Set 5	0.3817	0.3895
Average	0.4102	0.4219

表格 9：考慮與標題的相似度特徵時語句摘錄的召回率

	Original	Modified
Set 1	0.4511	0.4798
Set 2	0.3944	0.3980
Set 3	0.4723	0.5217
Set 4	0.4777	0.4967
Set 5	0.5024	0.5229
Average	0.4596	0.4838

表格 10：考慮向心性特徵時語句摘錄的召回率

編號	新詞	編號	新詞
1	**ｅ世代**	**11**	**納莉風災**
2	**五一大執法**	**12**	**國營事業**
3	**水資局**	**13**	**張昭雄**
4	**台聯黨**	**14**	**許家班**
5	**台灣水**	**15**	**尊李**
6	**台灣正名**	**16**	**跑票案**
7	**宋氏兵法**	**17**	**新系**
8	**客家文學**	**18**	**新政治**
9	**政黨化**	**19**	**翡翠水庫**
10	**殷琪**	**20**	**選舉機器**

表格 11：利用詞彙相關程度所找到的部分新詞

5.3.3 整體結果比較

表格 12 比較傳統以文件集為訓練基礎的方法與我們所改進的方法，其中 Original **及** Modified **兩種方法** Score Function **中的** *w1*、*w2*、*w3*、*w4* **及** *w5* **大小皆固定為** 1**。由表中可以知道對於每個集合的測試結果而言，大致上二種方法的結果是非常接近的，並沒有特別的好壞之分，且相對於** Original **來說，**Modified **進步的平均幅度大約只是** Original **的** 0.6%**左右。有些集合如** Set 1 **及** Set 4 **甚至結果會比較差。**

	Original	Modified	Improvement
Set 1	0.2746	0.2684	-2.3%
Set 2	0.3700	0.3772	1.9%
Set 3	0.2769	0.2841	2.6%
Set 4	0.2633	0.2574	-2.2%
Set 5	0.2419	0.2478	2.4%
Average	0.2853	0.2870	0.6%

表格 12：Original**與** Modified **的實驗數據比較(考慮所有的特徵)**

前一節討論不同特徵對於摘要系統的影響，由結果可知語句位置(Position)**、正面關鍵詞**(Positive Keyword)**、與標題的相似度**(Resemblance to the Title)**及向心性**(Centrality)**四個特徵較為重要，因此，接下來的實驗我們探討這四個特徵組合的影響。**

表格 13 **中列出** Original **與** Modified **的方法在不考慮負面關鍵詞**(Negative Keyword)**情況下的結果。相對於** Original **來說，**Modified **進步的平均幅度大約是** Original **的** 5.5%**左右，這個結果與前一節中所討論的結果—負面關鍵詞的影響不重要—互相呼應；明顯地，只考慮語句位置、正面關鍵詞、與標題的相似度及向心性是較為適當的組合。**

	Original	Modified	Improvement
Set 1	0.4647	0.4906	5.6%
Set 2	0.3799	0.4028	6.0%
Set 3	0.4191	0.4491	4.7%
Set 4	0.5142	0.5348	4.0%
Set 5	0.5149	0.5410	5.1%
Average	0.4586	0.4837	5.5%

表格 13：Original**與** Modified **的實驗數據比較(不考慮負面關鍵詞)**

表格 14 **列出基因演算法對每個訓練文件集所訓練出來的特徵值權重組，訓練的過程中，我們只考慮重要的四個特徵(亦即不考慮負面關鍵詞)。表格中** T1

代表訓練集 1(亦即以 Set 2~Set 5作為訓練集)，其餘類推，Recall 代表利用原訓練集作為測試集時所得到的召回率。

	Position	Positive Keyword	Resemblance to Title	Centrality	Recall
T1	0.926	0.013	0.359	0.002	0.7841
T2	0.867	0.013	0.689	0.011	0.7875
T3	0.996	0.013	0.401	0.025	0.7674
T4	0.981	0.021	0.527	0.004	0.7782
T5	0.875	0.012	0.581	0.022	0.7746

表格 14：利用基因演算法所得到的特徵權重組(不考慮負面關鍵詞)

表格 15 是經過基因演算法訓練的 Score Function (Modified+GA)與沒有經過訓練的 Score Function (Modified)的比較。從表中可知，Modified+GA 表現比 Modified 來得好，Modified+GA 進步的幅度平均為 Modified 的 7.4%左右；由此驗證了先前每個特徵的重要性皆不同的假設。

將基因演算法應用在訓練 Score Function 上，最大的益處在於得到的特徵權重組合是比較適當的，可以提供研究人員了解整個訓練文件集(Training Corpus)的特性，並當作系統參數調整的參考。當測試文件集(Test Corpus)的特性越接近訓練文件集的特性時，將基因演算法所找出來的 Score Function 套用在測試文件集時，我們認為亦可得到不錯的結果。

	Modified	Modified+GA	Improvement
Set 1	0.4906	0.5556	13.2%
Set 2	0.4028	0.4790	18.9%
Set 3	0.4491	0.4604	2.5%
Set 4	0.5348	0.5376	0.5%
Set 5	0.5410	0.5655	4.5%
Average	0.4837	0.5196	7.4%

表格 15：Modified 與 Modified+GA 的實驗數據比較(不考慮負面關鍵詞)

第四節　潛在語意分析語句摘要之可行性評估

5.4.1 實驗結果

這個實驗評估 LSA 文件摘要系統的可行性。為了與以文件集為基礎的方法(Corpus-based Approach)作比較，同樣地，我們針對前一節中所提到的每個文件集作評估；不同的是，對每個集合而言，LSA 模型建構中的矩陣維度約化(Dimension Reduction)的程度會因為文件集特性的不同而有所差異；比如說 Set 1 的最佳維度為 65%，即矩陣分解後 S 的 Rank 為 n 的話，那麼 S'的 Rank 便是 0.65*n。平均來說，這些文件集合的維度大約 64%左右可以得到不錯的摘要結果。

這一節裡，我們比較以關鍵詞(Keyword)的向量表示法建構主題關係地圖(Text Relationship Map)的方法(Keyword-based Text Relationship Map，簡稱 Keyword-based T.R.M.) [Salton97]及以 LSA 所得到的語意表示法建構主題關係地圖的方法(LSA-based Text Relationship Map，簡稱 LSA-based T.R.M.)間的差異性。由表格 16 中可知，LSA-based T.R.M.所得到的結果皆比 Keyword-based T.R.M. 來得好；平均來看，相對的改進的幅度大約為 12.9%左右，其中幾個集合如 Set 1，相對改進的幅度更高達 34.8%。表格 17 中我們列出不同維度約化對於摘要結果的影響。

	LSA-based T.R.M.		*Keyword-based T.R.M.*	
	Dimension Reduction	**Recall**	**Recall**	**Improvement**
Set 1	0.65	0.4616	0.3425	34.8%
Set 2	0.45	0.4005	0.3817	4.5%
Set 3	0.8	0.4567	0.4469	2.2%
Set 4	0.65	0.4657	0.4276	9.6%
Set 5	0.65	0.4943	0.4201	17.7%
Average	0.64	0.4558	0.4038	12.9%

表格 16：以 LSA 與 Keyword 向量表示法來實作 Global Bushy Path [Saltoon97]摘要方法的比較

Dimension Reduction	Set 1	Set 2	Set 3	Set 4	Set 5
0.05	0.4519	0.3985	0.3614	0.3044	0.3272
0.10	0.3229	0.2959	0.3690	0.2897	0.3526
0.15	0.3002	0.3581	0.3183	0.3716	0.3609
0.20	0.3144	0.3211	0.3600	0.3811	0.3810
0.25	0.3235	0.3251	0.3294	0.3761	0.3824
0.30	0.3343	0.3178	0.3663	0.4097	0.4043
0.35	0.3164	0.3842	0.3861	0.4165	0.4381
0.40	0.3911	0.3841	0.4122	0.4024	0.4441
0.45	0.4102	0.4005	0.4265	0.3773	0.4648
0.50	0.4029	0.3798	0.3985	0.4173	0.4860
0.55	0.4374	0.3557	0.3889	0.4374	0.4784
0.60	0.4236	0.3443	0.4312	0.4305	0.4737
0.65	0.4616	0.3359	0.4256	0.4657	0.4943
0.70	0.4319	0.3591	0.4512	0.4342	0.4500
0.75	0.4121	0.3541	0.4545	0.3789	0.4792
0.80	0.4273	0.3742	0.4545	0.3789	0.4792
0.85	0.4217	0.3635	0.4515	0.4073	0.4467
0.90	0.3781	0.3087	0.4313	0.3922	0.3803
0.95	0.3614	0.3705	0.4441	0.3849	0.3859
Average	0.3854	0.3543	0.4033	0.3935	0.4253

表格 17：不同的維度約化比例對摘要結果的影響

5.4.2 範例文件討論

我們分析新聞文件的摘要內容發現，LSA-based T.R.M.**所得到的摘要，其所提供的概念比** Keyword-based T.R.M.**的方法所得到的摘要更具說服力，且涵蓋的內容較為完整。我們舉以下的範例作說明，該文件的標題為「前總統夫人曾文惠出現在台北地方法庭」，內容是有關曾文惠「八千五百萬元美金運送風波」與謝啟大、馮滬祥以及戴錡等三人的惡意誹謗官司案。**

其中，每個語句的開始會標示該語句的段落位置，如 P1S1 **代表第一段第一句。如果語句屬於人工摘要的話，語句前面會標示有** *[1]* **的符號；如果語句屬於** Keyword-based T.R.M.**所生成的摘要，語句前面會標示有** *[2]* **的符號；如果語句屬於** LSA-based T.R.M.**所生成的摘要，語句前面會標示有** *[3]* **的符號。**

標題	前總統夫人曾文惠出現在台北地方法庭
內容	[1,2,3]<P1S1>三月四日一大早約九點出頭，前總統夫人曾文惠在女兒李安妮與隨扈的護送下，出現在台北地方法庭。[2,3]<P1S2>在出發之前，前總統李登輝才對曾文惠表示了精神上的完全支持，但是她還是抵擋不住硬吞下眼淚的那種心情。 [1,2,3]<P2S1>台灣有史以來，第一次出現前第一夫人到法院出庭的情況，曾文惠臉上沒有面對群眾時慣有的那種溫暖笑容，而是勉強擠出淺淺的笑，低著頭快速地進入法庭。<P2S2>只有在步出法庭時，看到熱情的支持群眾，她才露出親切溫柔的笑臉。 <P3S1>許多人都還記得，當然，李登輝一家人也都深深地記得。[1,3]<P3S2>兩年前總統大選後的那幾天，許多「國民黨人士」包圍國民黨中央黨部，在民眾情緒激憤，要求李登輝下台的時候，謝啟大在宣傳車上，對著底下的群眾喊著「曾文惠帶了八千五百萬美金逃到美國」。 <P4S1>接下來，前立委馮滬祥以及前僑務委員戴錡更召開記者會，提出洋洋灑灑的「證據」，公開指稱曾文惠搭乘長榮航空，私運八千五百萬美元到美國，被美方拒絕入境，又緊急搭華航班機運回美元，於是引來了所謂的「八千五百萬元美金運送風波」。 [1]<P5S1>小女兒李安妮不甘曾文惠被如此惡意誹謗，建議曾文惠自訴謝啟大等三人涉嫌誹謗，並求償三億元賠償。[1,2,3]<P5S2>但是，法官出身的謝啟大深閨司法，第一次出庭就採取反擊，反控曾文惠誣告，也要求三億元賠償，並且要求曾文惠出庭，也使得曾文惠必須在三月四日出庭應訊。 <P6S1>當天，曾文惠進入台北地院的北大門時，離開庭時間還有約半個小時，她快速地走上樓梯進入休息室，並準時出現在位於二樓的第七法庭。[2]<P6S2>經過冗長的庭訊過程，從上午九點四十分開庭到中午一點休息，曾文惠完全沒有發言。<P6S3>經過短暫的休息之後，曾文惠才站在法庭前接受法官的詢問，否認運美金赴美。 [1,2]<P7S1>在經過身體與精神的雙重煎熬之下，下午三點多，曾文惠終於承受不住心裡的委屈，趴在桌上偷偷地落淚，並在李安妮的攙扶下暫時離開法庭。<P7S2>在庭訊的過程中，曾文惠也不禁用紙張寫下她的心情，「上帝創造人的眼淚是流下來的，我的眼淚卻是吞進去的」。 <P8S1>實際上，基於對司法的尊重，曾文惠與家人也完全不願意對這件官司發表談話。<P8S2>而儘管曾文惠的高中校友鄭玉麗，曾經在二〇〇〇年三月二十二日下午打了通電話給她，並聊了將近半個小時，但基於自己沒有舉證責任的原則之下，曾文惠也不願鄭玉麗出面作證。 [1]<P9S1>對曾文惠而言，這場官司是一種捍衛自己尊嚴的官司。<P9S2>看著老妻受到這麼大的委屈，李登輝心底絕對是相當心疼的。

這個例子中，LSA-based T.R.M.**所得到的召回率是** 0.67**，**Keyword-based T.R.M.**所得到的召回率為** 0.50**。造成這個差異的原因乃是** LSA-based T.R.M.**選擇** P3S2**，而** Keyword-based T.R.M.**選擇** P6S2**；直覺來看，**P3S2 **的重要性比** P6S2 **的重要性來得高，由此可以看出** LSA-based T.R.M.**比** Keyword-based T.R.M.**更強大的地方在於** LSA-based T.R.M.**可以將原本隱含在語句間的語意顯現出來。**

此外，我們分析 Keyword-based T.R.M.**及** LSA-based T.R.M.**這兩種方法所建構出來的主題相關地圖，如表格** 18**；表格中第一行**(Column)**表示語句** S_i**，第二行表示與** S_i **具有連結的語句，第三行表示** S_i **中的連結數目**(**即先前所提過的** *Bushiness* [Salton97])**。**

	LSA-based T.R.M.		*Keyword-based T.R.M.*	
	相關聯語句	**連結數**	**相關聯語句**	**連結數**
P1S1	P2S1, P5S2, P6S2, P7S1	4	P1S1, P2S1, P4S1, P5S1, P6S2, P7S1	6
P1S2	P2S1, P3S2, P7S1, P7S2	4	P1S1, P2S1, P2S2, P7S1, P7S2	5
P2S1	P1S1, P1S2, P2S2, P3S2, P4S1, P5S2, P6S1	7	P1S1, P1S2, P2S2, P5S2, P6S1	5
P2S2	P2S1	1	P1S2, P2S1, P6S1	3
P3S1	P3S2	1	P3S2	1
P3S2	P1S1, P2S1, P3S1, P4S1, P5S1, P5S2, P9S2	7	P3S1, P4S1, P5S2	3
P4S1	P2S1, P3S2, P5S2, P6S3	4	P3S2, P6S3	2
P5S1	P3S2, P5S2, P7S1	3	P1S1, P5S2, P7S1	3
P5S2	P1S1, P2S1, P3S2, P4S1, P5S1	5	P1S1, P2S1, P3S2, P5S1	4
P6S1	P2S1, P8S2	2	P2S1, P2S2, P8S2	3
P6S2	P1S1	1	P1S1, P6S3, P7S1, P7S2	4
P6S3	P4S1, P7S1	2	P4S1, P6S2	2
P7S1	P1S1, P1S2, P5S1, P6S3, P8S2, P9S2	6	P1S1, P1S2, P5S1, P6S2, P8S2, P9S2	6
P7S2	P1S2	1	P1S2, P6S2	2
P8S1	P9S1	1	P9S1	1
P8S2	P6S1, P7S1	2	P6S1, P7S1	2
P9S1	P8S1	1	P8S1	1
P9S2	P3S2, P7S1	2	P7S1	1

表格 18：LSA-based T.R.M.**及** Keyword-based T.R.M.**得到的主題相關地圖**

以 P3S2 **與** P4S1 **來討論，**Keyword-based T.R.M.**計算兩者間的關聯程度為** 0.0831**，而** LSA-based T.R.M.**得到的關聯程度則提高為** 0.8604**。對** Keyword-based T.R.M.**而言，兩個語句間的關鍵詞重複性並不大，只有『八千五百萬』、『美金』、『美國』與『曾文惠』等關鍵詞重複, 因此算出來的關聯性並不大, 但是,** LSA-based T.R.M.**卻可以將兩句間的關聯性提高。**

綜合上述，LSA-based T.R.M.的確可得到更高的文件摘要正確率。對於每個測試集而言，只要能夠找得到最佳的維度約化比，語句間的關聯強度分析便能夠越正確，且可將文件內容所隱含的詞意表現出來。

實驗的結果證實，以 LSA 為核心的摘要技術不失為一種可行的方法。最後，我們總結上述的幾個方法，將各種方法的比較結果整理於表格 19。由表中可知，LSA-based T.R.M.的結果與 Original Corpus-based 相近，由此更可驗證 LSA 適用於文件摘要的可行性。LSA-based T.R.M.的方法的好處在於其不像以文件集為基礎的摘要方法(Corpus-based Approach)一樣需要經過大量文件集的學習，因此不受限於領域(Domain)的相關性。最後，附錄一展示我們的所實作的系統，附錄二中列出幾篇範例文章及其摘要作為參考。

	Set 1	Set 2	Set 3	Set 4	Set 5	Average
Original Corpus-based	0.4647	0.3799	0.4191	0.5142	0.5149	0.4586
Our Modified Corpus-based	0.4906	0.4028	0.4491	0.5348	0.5410	0.4837
Our Modified Corpus-based+GA	0.5556	0.4790	0.4604	0.5376	0.5655	0.5196
Keyword-based T.R.M.	0.3425	0.3817	0.4469	0.4276	0.4201	0.4038
Our LSA-based T.R.M.	0.4616	0.4005	0.4567	0.4657	0.4943	0.4558

表格 19：各種摘要方法的綜合比較

第六章　結論與未來研究方向

本章總結本論文，以及說明未來的研究方向。第一節討論本論文所提的兩種方法應用在摘要系統上的效益與可行性，並且說明這兩種方法的特性，包括其優點及限制；第二節則說明未來可能的研究發展方向。

第一節　結論與討論

本論文討論指示性(Indicative)、一般性(Generic)與單文件(Single-Document)摘錄(Extract)的自動摘要方法，並將研究結果應用於中文文件上。首先，針對過去以文件集為基礎的摘要技術(Corpus-based Text Summarization)作修改，提出改良型的語句權重摘要方法；第二，採用潛在語意分析(Latent Semantic Analysis)作為文件的知識模型，並結合主題關係地圖(Text Relationship Map)的概念提出一套新的摘要技術。實驗的結果證實以上兩種方法結果皆比其他的方法來得好。此外，我們所提出的方法並不侷限於中文文件，亦可應用於英文文件上。

本論文所提的第一個方法具有以下的特性：

1. 除了計算語句位置出現在摘要的條件機率外，針對每個語句位置的不同，我們賦予不同的權重以加強其重要性；亦即，對於每個語句而言，所考慮的是該語句位置落於摘要的期望分數。
2. 考慮到中文斷詞好壞會深深影響摘要的正確性，因此，我們利用詞彙相關程度(Word Co-occurrence)以找出文件集中的新字詞，並將這些新詞加入關鍵詞權重的計算。
3. 利用基因演算法得知 Score Function 中每個特徵值的權重，以增進研究人員對於訓練集特性的了解，並提供系統設計者對於系統效能的參數調整參考。

本論文所提的第二個方法具有以下的特性：

1. 利用潛在語意分析(LSA)可以直接建構文件內容的知識模型；相較於過去利用外在資源分析的方法，LSA 更能將文件內容的涵義分析提昇至語意的層面，並可以避免語意混淆的問題。
2. LSA 可將文件中潛在的語意表現出來，因此，以 LSA 分析的結果作為語句分群時能夠更精確地定義語句意義，以得到更好的語句分群。
3. 當找到最佳的維度約化時，LSA 語意模型中的語意推演及擷取的結果會更精準，更能表現出文件中的潛在語意與字詞間的關聯。

第二節　未來研究方向

本論文利用 LSA 辨認並推演文件隱含知識的模型，將傳統文件摘要中採用的資訊擷取技巧提昇至知識概念的層面，以達到更精確的語意分析及自動化文件摘要的目的，用以幫助閱讀者在短時間之內正確地得知文件所要表達的概念，並做為決策參考。

未來我們將針對以下四點進行更深入的研究：

1. 實作上，我們保留名詞及動詞來建構 LSA 的模型，但是受限於斷詞切字系統能力的不足，無法將正確的關鍵詞切割出來且會導致關鍵詞詞性的錯誤標示，而這個錯誤可能使最後的語意分析混淆。未來我們將進一步分析中文語句的結構樹，並加入專有名詞的判斷，以期能更正確地分辨詞意。
2. 我們以單篇文件為單位建構 LSA 模型，但是，受限於文件大小的問題，LSA 的結果只表現出字詞與字詞間共同出現的關聯性，隱性語意的表現比較缺乏。未來，希望將 LSA 套用到較大的文件集上，將 Context 由語句提升到以文件為單位來建構 LSA 的矩陣模型(即建構

Word-by-Document)，並將 LSA 的結果套用到單文件的摘要上。如此一來，除了考慮到單一文件的涵義之外，亦考慮到單一字詞在文件集中所有文件的意義，以期能夠更精準的將隱性的語意顯現出來。

3. LSA 分析的維度約化過程中，實驗的時候我們嘗試各種不同的可能性，並保留下較佳的結果。未來，希望藉由整個文件集的分析與使用者回饋機制(User Feedback)的學習能力，期望自動調適得到最適當的矩陣維度。

4. 我們預期 LSA 理論探究的成果除可應用在文件自動摘要技術之外，尚可應用於文件分類(Document Classification)、主題偵測及追蹤(Topic Detection & Tracking)、語意搜尋引擎(Semantic Search Engine)等知識管理(Knowledge Management)系統的核心技術，這也是我們未來擬繼續探究的方向。

附錄一：實作系統展示

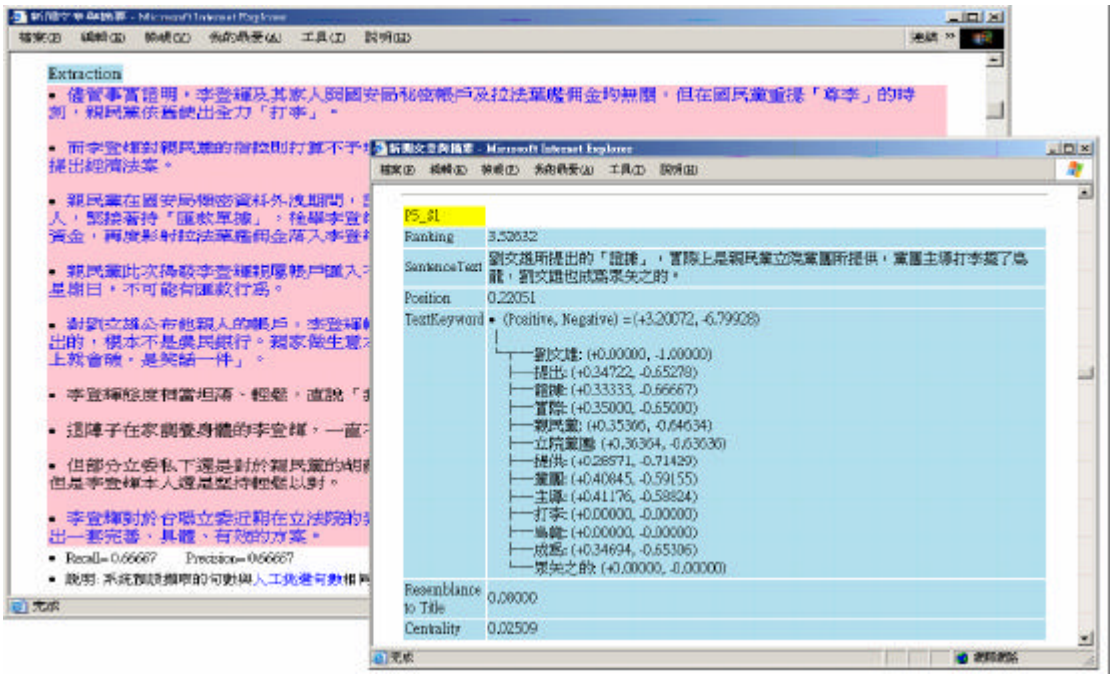

圖 18：Modified Corpus-based Approach

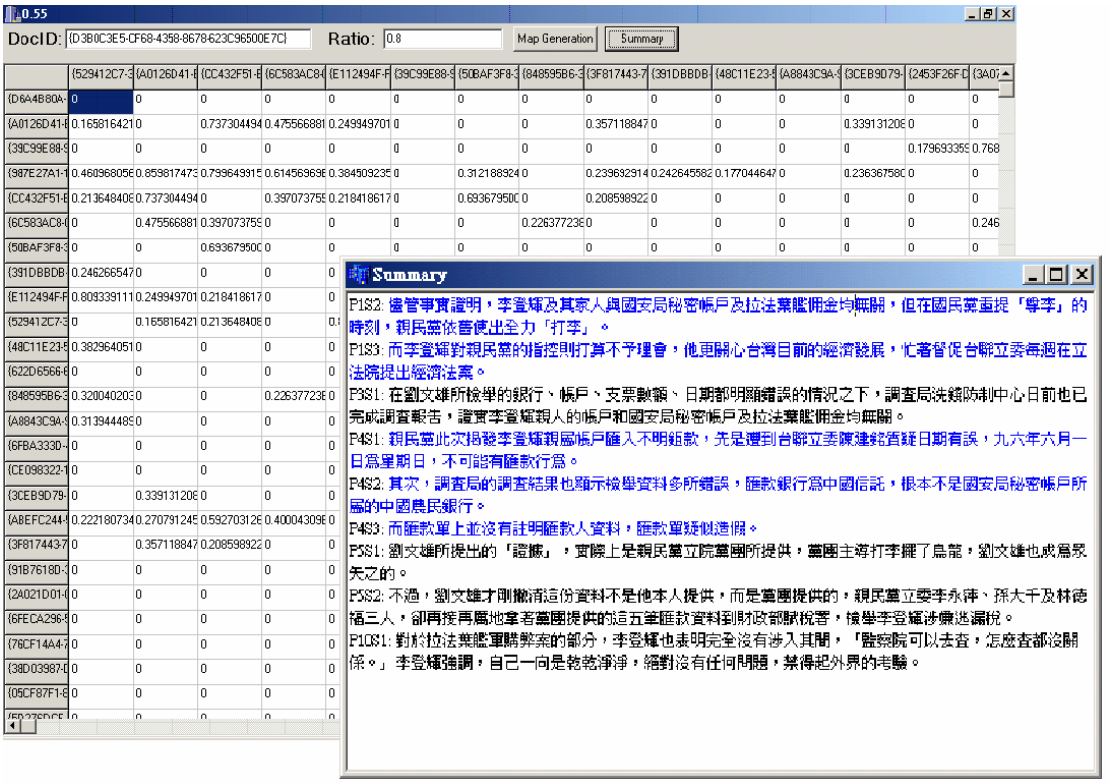

圖 19：LSA-based T.R.M. Approach

附錄二：範例文件

我們列出三篇範例文件及其相對應的摘要結果，其中 [1] 代表人工摘要，[2] 代表 Original Corpus-based Approach**，[3] 代表** Our Modified Corpus-based Approach**，[4] 代表** Keyword-based T.R.M.**，[5] 代表** Our LSA-based T.R.M.**所得到的摘要。**

(1)標題：平衡派系，洪奇昌可望出線

[1,2,3,4,5] 立法院副院長到底花落誰家，在陳水扁總統於二十二日分別接見民進黨內五位有意參選正、副院長的人選之後，已經逐漸明朗，在陳水扁所提出的「資望、倫理、平衡」的三原則之下，屬於新潮流系的洪奇昌，目前是民進黨最資深的五連任立委，不僅符合前兩項要素，而且在「平衡」派系的考量之下，洪奇昌反而有可能因為新潮流的背景而出線，可望在二月一日和王金平搭檔競選正、副院長。

*[1,2,3,4,5] 在民進黨取得國會第一大黨之後，黨內想要角逐立法院正副院長的人選就開始浮現，不過除了蔡同榮把第一志願放在立法院院長之外，其餘有興趣的人，都把目標放在副院長，因為民進黨雖然是國會最大黨，但是還沒有過半，而台聯黨已經表態會支持王金平連任院長，所以沈富雄、柯建銘及洪奇昌三人都很有默契只爭取立法院副院長。*而且民進黨不強求拿下正、副院長，而讓王金平連任院長，就是不讓泛藍軍因院長選舉而再度結盟，才符合民進黨的最大利益。

經過一個多月的爭逐，蔡同榮雖然透過相關人士不斷向總統府方面建議，民進黨已經是第一大黨，應該要提出自己的正、副院長人選，不過總統府方面始終沒有正面回應。*[5] 反而是王金平的態度一次比一次明確，從強調正、副院長的人選應該透過黨對黨協調，到後來表示正、副院長應該脫鉤競選，並說「這次要自私一點，要自己選」。[1,2,3] 民進黨人士解讀王金平這些話，認為王金平已經擺平黨內和民進黨合作的雜音，並且要求連戰表態只提名立法院院長，不要提名副院長人選，這些動作，再加上連戰一句「副院長人選要考量政治現實」，這位民進黨人士強調，王金平算是已經對民進黨表達善意，而且總統府也從來沒有說過要和王金平搶立法院院長，所以立法院正副院長的「國、民配」應該會是水到渠成。*

所以整個副院長的局勢，在陳水扁總統於二十二日接見五位候選人之後，已經大勢底定。原本有意競選立法院長的蔡同榮，在離開總統府之後，馬上和立委鄭寶清及尤清闢室密談，在研判陳水扁將支持王金平競選院長後，隨即作出改選副院長的決定。*[4] 而沈富雄在離開總統府之後，也轉述兩人在一個多小時內的談話。[4] 沈富雄表示兩人討論的議題，主要是談到立法院的運作問題，包括黨團三長、黨內派系，但是並沒有提到要把他放在那一個位置。*雖然席間兩人曾針對副院長、黨團總召及黨主席人選的條件有一番討論，但是並沒有提出那一個人比較適合放在那個位置的問題。不過沈富雄表示，他自己倒是和陳總統談過他和洪奇昌兩人特質的「十大比較」，但是並無關兩人孰優孰劣，純粹供陳總統參考。

[1,2,3,4,5] 二十二日晚間洪奇昌和柯建銘兩人也陸續到總統府，洪奇昌於七點多離開總統府之後表示，他與陳總統都不希望看到副院長人事問題造成黨團內部緊張，洪奇昌也強調，陳總統知道黨團運作狀況，也相信黨團會在適當時機，協調出兼具「資望、倫理、平衡」的適當人選，對於副院長人選，洪奇昌也表示陳總統並沒有明確表示。[1,4,5] 緊接在洪奇昌之後，福利國立委柯建銘也隨即進入總統府，也在會後轉述他和陳總統的談話表示，陳總統衡量副院長人選最重要的因素就是「平衡」，在立法院副院長、黨主席、黨團總召集人的佈局上將會維持派系平衡。

也就是柯建銘一句「平衡」是陳水扁考量副院長最重要的因素，所以洪奇昌轉而成為副院長的大熱門人選，不過洪奇昌仍然保持低調，對於外界的詢問皆以黨團會在二十四日運作帶過。不過據了解，洪奇昌新潮流的背景，不但沒有在這場副院長競爭中帶來負面的影響，反而具有加分的作用。*[1,2,3] 因為在這次內閣改組中，新潮流的龍頭指標之一邱義仁，離開秘書長職務，轉任政務委員，指標之二的民進黨秘書長吳乃仁，也明確表示不再擔任黨秘書長一職，新系兩大指標遠離權力核心，讓原本保持派系平衡的民進黨，一下子傾斜成「扁系」獨大。[1,4,5] 為了要維持民進黨內的派系平衡，所以在總統府的思考當中，立法院副院長、黨團總召及民進黨黨主席三個位置，成為維繫派系平衡的三大指標。[2,3] 而如果以功能性來說的話，其實立法院副院長一職雖然「位*

高」，但是不見得「權重」，所以副院長一職，就成為總統府方面考慮可以讓給新潮流的洪奇昌的主要原因。

[2,3,5] 不論陳水扁向沈富雄所說的，「副院長和黨主席比起來，重要性比不過黨主席」，是屬於「安慰」性質，或是實話，據了解，在高層心中，副院長的位置的確沒有外界所想像的這麼重要，因為從民進黨只主攻副院長來看，未來三年在立法院，王金平仍然是陳水扁在立法院對話的主要窗口，所以副院長的位置雖然稱不上是聊備一格，但是也沒有想要爭取的人所「膨脹」的這麼大，只要不要讓泛藍軍藉著立法院長改選而再度集結，民進黨在立法院長一役中，就算成功。

至於曾經在派系會議中決議要支持沈富雄的正義連線，也在情勢比較清楚之後，改以「全力配合黨團對正副院長運作」、及「全力配合派系對黨團三長運作」為主要目標，而沈富雄也暗示自己可能已經在這場副院長之爭出局。至於柯建銘也僅低調的表示支持黨的候選人，也不願再對副院長人選發表意見。而對親民黨一直傳出不放棄「國親配」的相關談話，立委陳其邁也表示，情勢都已經這麼明朗了，還有需要去計較這些話嗎？

(2)標題：綠色和平電台

*[1]「台灣、台灣，尚愛你啦！台灣、台灣，用感情作伴，FM97 點 3」。*從收音機聽到這段音樂，代表你現在收聽的電台正是綠色和平台灣文化廣播電台沒錯。這也是目前北台灣地區合法的中功率電台當中最具規模的本土電台。

十年前，隨著台灣社會力量的爆發，政治、社會運動此起彼落，地下民主電台也如雨後春筍般設立。*[1,3,5] 抄台的惡夢，卻是每一家地下電台最大的負擔，根據統計，從開放申請後的一九九四年到二??一年，政府總共抄過五十多家非法電台，但在九四年之前的數字，大概是這些的好幾倍。*

[1,2,3,4,5] 因此，在九三年台灣開放第一波電台申請之後，當時已具有相當聽眾基礎的綠色和平電台便集合民間力量，成功地募集到五千多萬的申請資金，成為合法的本土電台。

[4] 當時，綠色和平並不是唯一合法的本土電台，如今卻似乎成為唯一。這是許多反對陣營相繼棄守後的結果，像張俊宏的台灣全民廣播電台，如今納入趙少康的飛碟聯播網。[3,5] 許多當年在政治理想下申請的電台，如今不但易手，還交由完全不同政黨色彩的人在經營。

在目前的廣播天空當中，綠色和平顯得獨樹一格。*[4,5] 然而，從十年前還是地下電台時期，到了九四年合法之後，綠色和平的節目內容還是有所調整。*過去許多民主電台的易手，無非是因為經營不善，綠色和平因此必然要在商業與理想之中，取得一個平衡。總經理周明鳳就直接說了「要賺錢才能生存啊！」

*[1,2,4] 相較於過去以政治、本土文化、環保為主的內容，目前綠色和平的節目性質可說是相當地多元化，台長陳德利認為，這也是反映了目前台灣社會的本質。*他利表示，台灣的政治激情時期已經過去了，不能再停留於過去地下電台時期的心態。

*[2,3,4,5] 目前的節目當中，除了維持了相當多過去政治短評類的談話節目之外，也有醫療、法律服務的節目、台灣文學節目、本土音樂節目，以及相當多吸引年輕聽眾的世界音樂節目、哈日流行節目、同性戀議題節目。*此外，也將部分時段外包，賺取時段外包的收入。

[2,3,4] 與過去總是單調的政治評論節目相較，綠色和平為了迎合多元社會的多元聽眾，做了相當大的改變。[2,4,5] 當然，外界的批評也不是沒有，但周明鳳認為，當初聽眾以熱情資助綠色和平五千萬的創台基金，電台必須有基本的收入來維持運作，不能一直向人伸手。

[2] 既然不能繼續過去的路線，多少需要引進商業性的節目。[2,3,4] 周明鳳說，部分外製的節目，剛開始時，也有聽眾反對，認為這是「賣藥ㄟ」，但是綠色和平並不是隨便就將時段賣人，而是考慮對方的屬性是否符合電台的風格，並且規定節目當中賣的東西必須有衛生署檢驗合格，並且不能直接提到任何療效。

[1,3,5] 漸漸地，由於電台挑選的主持人也相當具有本土意識，並且有一定水準，不但原本的聽眾喜歡收聽，也為綠色和平吸引到相當多原本對政治較不關心的聽眾。[1] 周明鳳認為，吸引新加入的聽眾，更能為本土意識的扎根盡一份力量。

[1,2,3] 綠色和平想吸引更多各階層、各年齡的聽眾，就必須推出更多元的節目，因此除了基本的本土性節目之外，不但也針對上班族的股票節目「股市操盤手」，還找了許多年輕的主持人加入，針對年輕人推出許多活潑、另類的節目。

這些選在晚間學生下課後播出的節目，諸如豬頭皮的「萬國大仔標」、哈日話題的「都會物語」、同性戀議題的「真情酷兒」、網路話題的「電腦五四三」……，為綠色和平培養了許多年輕一輩的忠實聽眾。

[2,3,5] 身為老一輩的廣播人，陳德利一開始其實也對部分新的節目內容有所抗拒，但是最後他體認到，應該順著時代的潮流來堅持最終的理想。[5] 只要節目沒有違背當初捐助基金的民眾賦予綠色和平的使命，這也是電台發展的一種必然。

[4,5] 在眾多節目當中，陳德利也提到，令他印象相當深刻的節目是蔡振南的「南歌人生」，這個節目已經推出許久，一直相當受各種聽眾的歡迎。[2,3] 節目除了介紹許多台語歌之外，也因為蔡振南曾經成為受刑人的一段經歷，固定與台北監獄合作，讓不同的受刑人上節目談話，十分有教育意義。

*[1,4] 四月十九日，距離一九九四年綠色和平在熱情民眾的捐助當中正式合法創台，已經八年了。[1] 這八年來，綠色和平的確有許多改變，努力在理想與現實之中拔河，但它的基本精神並沒有改變。*正如統派人士打電話到電台罵「你們是民進黨的電台嗎？」周明鳳的回答是，「我們不是民進黨的電台，但是我們比較支持民進黨」。

(3)標題：親民黨打李不休，李登輝只管拼經濟--檢舉的銀行、帳戶、支票數額、日期都明顯錯誤

在國安局祕密帳戶曝光後，親民黨藉著揭發弊案之名，一直緊咬前總統李登輝不放。*[1,2,3,5] 儘管事實證明，李登輝及其家人與國安局秘密帳戶及拉法葉艦佣金均無關，但在國民黨重提「尊李」的時刻，親民黨依舊使出全力「打李」。[1,3,4,5] 而李登輝對親民黨的指控則打算不予理會，他更關心台灣目前的經濟發展，忙著督促台聯立委每週在立法院提出經濟法案。*

[1,,2,3,4] 親民黨在國安局機密資料外洩期間，試圖影射李登輝將秘密帳戶納入自己口袋不成，親民黨立委劉文雄等人，緊接著持「匯款單據」，檢舉李登輝之女李安娜、媳婦張月雲之姐張桂芬的帳戶，在九七年間有五筆不明資金，再度影射拉法葉艦佣金落入李登輝口袋。

[4,5] 在劉文雄所檢舉的銀行、帳戶、支票數額、日期都明顯錯誤的情況之下，調查局洗錢防制中心日前也已完成調查報告，證實李登輝親人的帳戶和國安局秘密帳戶及拉法葉艦佣金均無關。

[1,3,5] 親民黨此次揭發李登輝親屬帳戶匯入不明鉅款，先是遭到台聯立委陳建銘質疑日期有誤，九六年六月一日為星期日，不可能有匯款行為。[1,4,5] 其次，調查局的調查結果也顯示檢舉資料多所錯誤，匯款銀行為中國信託，根本不是國安局秘密帳戶所屬的中國農民銀行。[1,5] 而匯款單上並沒有註明匯款人資料，匯款單疑似造假。

[5] 劉文雄所提出的「證據」，實際上是親民黨立院黨團所提供，黨團主導打李擺了烏龍，劉文雄也成為眾矢之的。[4,5] 不過，劉文雄才剛撇清這份資料不是他本人提供，而是黨團提供的，親民黨立委李永萍、孫大千及林德福三人，卻再接再厲地拿著黨團提供的這五筆匯款資料到財政部賦稅署，檢舉李登輝涉嫌逃漏稅。

對於李永萍等人的檢舉，台聯立委不禁紛紛搖頭表示，「這分明是牽連九族嘛！」羅志明就表示，「李登輝自己沒有錢，他的親家就不能有錢嗎？像我老婆娘家本來就很有錢，那就代表我都Ａ錢嗎？」

[2,5] 據了解，親民黨主席宋楚瑜相當關切這項李登輝家族帳戶案，對於黨團公布的資料與實際狀況有出入，還曾嚴厲質疑幕僚「這麼大的事，怎會有這種誤差？」[4] 台聯立委錢林慧君就認為，宋楚瑜自己才是逃漏稅的人。事實上，宋楚瑜因為興票案而必須補繳的贈與稅，目前宋楚瑜向財政部提了行政救濟，並不打算繳納九千多萬的稅款與罰款。

而李登輝本人則於四月八日晚間，趁著在鴻禧山莊設宴款待台聯立委之際表示他對國安祕密帳戶案的看法。*[4] 對於近日親民黨的「打李」行動，李登輝笑著表示，有一些人想製造「李登輝貪污」的印象，這完全是黑白講、沒這個事情。*

[1,2,3,4] 對劉文雄公布他親人的帳戶，李登輝輕鬆地表示，「日期不對、金額也不對，全都不對，錢是從中國信託匯出的，根本不是農民銀行。[1] 親家做生意本來就有很多錢進進出出，這些都是亂拼湊出來的，就像一個氣球，馬上就會破，是笑話一件」。

*[5] 對於拉法葉艦軍購弊案的部分，李登輝也表明完全沒有涉入其間，「監察院可以去查，怎麼查都沒關係。」*李登輝強調，自己一向是乾乾淨淨，絕對沒有任何問題，禁得起外界的考驗。因此，對於有心人士對他的汙衊，他完全不會加以理會。

[2,3,4] 李登輝態度相當坦蕩、輕鬆，直說「我乾乾淨淨、清清白白，大家可以去查。」[2,3] 這陣子在家調養身體的李登輝，一直不願對此表示意見，就是不願隨這些人起舞。[4] 對於親民黨的再三指控，他強調正好趁這個機會讓司法單位調查清楚，還他清白。

[2] 由於李登輝傾向採取不回應的態度，台聯黨團近日來也決定盡量不隨親民黨起舞。[2,3] 但部分立委私下還是對於親民黨的胡亂指控，感到相當氣憤，認為應該予以反擊，不能讓李登輝白白被罵，但是李登輝本人還是堅持輕鬆以對。

[1] 對李登輝而言，目前台灣人民的價值觀喪失，產生信心危機，國家的領導沒有一個正確的方向，才是他最關切的問題。[1,2,3] 李登輝對於台聯立委近期在立法院的表現相當滿意，也要求台聯立委再接再厲，全力拚經濟，對經濟政策提出一套完善、具體、有效的方案。

參考文獻

1. [Aone99] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen (1999), "A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp. 71-80, 1999.
2. [Azzam99] S. Azzam, K. Humphreys, and R. Gaizauskas (1999), "Using Coreference Chains for Text Summarization," In *Processings of the ACL'99 Workshop on Coreference and its Applications*, Baltimore, June, 1999.
3. [Barzilay97] R. Barzilay, and M. Elhadad (1997), "Using Lexical Chains for Text Summarization," In *Processings of the Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, August, 1997.
4. [Bellegarda96] J. R. Bellegarda, J. W. Butzberger, and Y. L. Chow (1996), "A Novel Word Clustering Algorithm Based on Latent Semantic Analysis," In *Conference on Acoustics, Speech, and Signal Processing*, IEEE, Vol. 1, pp. 172-175, 1996.
5. [Edmundson68] H. P. Edmundson (1968), "New Methods in Automatic Extracting," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp. 23-42, 1999.
6. [Goldstein99] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell (1999), "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," In *SIGIR*, ACM, Berkley, CA, USA, 1999.
7. [Gong01] Y. Gong, and X. Liu (2001), "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," In *SIGIR*, ACM, New Orleans, Louisiana, USA, September 9-12, 2001.
8. [Habn00] U. Habn, and I. Mani (2000), "The Challenges of Automatic

Summarization," In *Computer*, IEEE, Vol. 33, No. 2000, pp. 29-36, 2000.

9. [Halliday76] M. A. K. Halliday, and R. Hasan (1976), "Cohesion in English," Longman, London, 1976.
10. [Han01] J. Han, and M. Kember (2001), "Classifier Accuracy," In *Data Mining: Concepts and Techniques*, pp. 323-324, 2001.
11. [Hovy99] E. Hovy, and C. Y. Lin (1999), "Automated Text Summarization in SUMMARIST," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp. 81-94, 1999.
12. [Kim00] J. H. Kim, J. H. Kim, and D. Hwang (2000), "Korean Text Summarization Using an Aggregate Similarity," In *Processings of the 5th International Workshop Information Retrieval with Asian Languages*, ACM, 2000.
13. [Kowalski97] G. Kowalski (1997), "Information Retrieval Systems: Theory and Implementation," *Kluwer Academic Publishers*, 1997.
14. [Kupiec95] J. Kupiec, J. Pedersen, and F. Chen (1995), "A Trainable Document Summarizer," In *SIGIR*, ACM, Seattle WA, USA, 1995.
15. [Lam01] W. Lam, H. M. L. Meng, K. L. Wong, J. C. H. Yen (2001), "Using Contextual Analysis for News Event Detection," In *International Journal of Intelligent Systems*, Vol. 16, pp. 525-546.
16. [Landauer98] T. K. Landauer, P. W. Foltz, and D. Laham (1998), "An Introduction to Latent Semantic Analysis," In *Discourse Processes*, Vol. 25, 1998, pp. 259-284.
17. [Lin99] C. Y. Lin (1999), "Training a Selection Function for Extraction," In *CIKM*, ACM, Kansas City, MO, USA, 1999.
18. [Mani99] I. Mani, and M. Maybury (1999), "Introduction," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp. x-xv,

1999.

19. [McKeown95] K. R. McKeown, D. R. Radev (1995), "Generating Summaries of Multiple News Articles," In *SIGIR*, ACM, Seattle Washington, USA, 1995.
20. [Myaeng99] S. H. Myaeng, and D. Jang (1999), "Development and Evaluation of a Statistically Based Document System," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp.61-70, 1999.
21. [Salton97] G. Salton, A. Singhal, M. Mitra, and C. Buckley (1997), "Automatic Text Structuring and Summarization," In *Information Processing & Management*, Elsevier, Vol. 33, No. 2, pp. 193-207, 1997.
22. [Silber00] H. G. Silber, and K. F. McCoy (2000), "Efficient Text Summarization Using Lexical Chains," In *IUI*, ACM, New Orleans, LA, USA, 2000.
23. CKIP AutoTag, available at http://godel.iis.sinica.edu.tw/CKIP/.
24. WordNet (a lexical database for the English language). Available at http://www.cogsci.princeton.edu/~wn/.
25. [**陳光華** 98] **陳光華** (1998), "**新資訊時代的啟發性資訊服務**," 21 **世紀資訊科學與技術的展望學術研討會**, **桃園**, 1998.
26. [**陳鈺瑾** 00] **陳鈺瑾**, **與張俊盛** (2000), "**可調式之中文文件自動摘要**," ***碩士論文***, **國立清華大學資訊工程研究所**, **新竹**, 2000.
27. [**黃聖傑** 99] **黃聖傑**, **與陳信希** (1999), "**多文件自動摘要方法研究**," ***碩士論文***, **國立台灣大學資訊工程研究所**, **台北**, 1999.
28. [**翁鴻加** 01] **翁鴻加**, **與陳信希** (2001), "**多文件摘要一些新技術及評估模型之建立**," ***碩士論文***, **國立台灣大學資訊研究所**, **台北**, 2001.
29. [**蘇哲君** 01] **蘇哲君**, **與陳信希** (2001), "**中英雙語多文件自動摘要系統研究**," ***碩士論文***, **國立台灣大學資訊工程研究所**, **台北**, 2001.
30. [**蘇謖** 96] **蘇謖** (1996), "**自動摘要法**," ***中國圖書館學會會報***, **第** 56 **期**, **頁** 41-47, 1996.
31. **新台灣新聞週刊**. Available at http://www.newtaiwan.com.tw.