行政院國家科學委員會專題研究計畫　期中進度報告

多語言複合式文件自動摘要之研究(1/3)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-126-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立交通大學資訊科學學系

計畫主持人：楊維邦

計畫參與人員：楊維邦，柯皓仁，葉鎮源，謝佩原，梁哲瑋，鄭佳彬，劉政璋

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 5 月 31 日

中文摘要

自動文件摘要(Automated Text Summarization)研究，探討如何分析、統整文件中重要的資訊，並以簡明的形式呈現，可作為使用者或其他資訊系統的判斷與決策依據。過去關於文件摘要的研究，大多著眼於單文件摘要。近來，多文件摘要愈顯得重要。多文件摘要與單文件摘要最大的差異，在於過濾重複的資訊(Anti-redundancy)，同時須避免重要資訊的流失；此外，摘要內容排序(Content Ordering)，亦是多文件摘要必須探討的議題。

本計畫為三年期研究案『多語言複合式文件自動摘要之研究』之第一年計畫。本年度計畫之目的，在於研究並發展多文件自動摘要的技術，將探討如何導出文件的結構、如何組織文件的內容及如何表示所抽取出文件抽象涵義等相關議題。研究範疇包含：1) 文件模型的建構及表示；2) 文件主題偵測及重要性評估；3) 摘要內容組織及排序。

我們利用潛在語意分析(Latent Semantic Analysis)與主題關係地圖(Text Relationship Map)作為多文件分析模型，提出三種段落重要性評估模型，分別為 1) Global Bushy Path；2) Aggregate Similarity；3) Spreading Activation。根據上述模型，我們基於 Maximal Marginal Relevance 提出新的摘要段落挑選方式。同時，針對摘要內容排序，考慮主題呈現順序，亦提出新的排序方法。

關鍵詞：多文件自動摘要；潛在語意分析；主題關係地圖

英文摘要

Automated text summarization investigates the process of extracting the most important information from a source (or sources), and presenting a summary to the user. In the past, most related work on text summarization focuses on single-document summarization. Multi-document summarization obtains increasing attentions in recent years. The distinction between single and multi-document summarization is that the latter has to handle anti-redundancy as well as to keep salient information meanwhile. Moreover, content ordering, which is to provide a cohesive and coherent summary, is an important issue to be examined.

As the first part of the complete project "The Research on Cross-Language, Composite and Multi-Document Automated Text Summarization", the principal objective is to develop new approaches to address multi-document summarization. Our researches include 1) Conceptual Modeling and Representation for Multiple Documents, 2) Topic Detection and Paragraph Significances Measurement, and 3) Content Ordering for the summary.

We exploit latent semantic analysis and text relationship map to derive conceptual model for multiple documents. Based on the model, we propose three approaches to measure the significance of a paragraph. They are 1) Global Bushy Path, 2) Aggregate Similarity, and 3) Spreading Activation. Besides, we propose a novel paragraph re-ranking approach on the basic foundation of Maximal Marginal Relevance to extract salient paragraphs. Furthermore, a content ordering method is proposed as well to take into account topic relations.

Keywords: Multi-document Summarization; Latent Semantic Analysis; Text Relationship Map

1. 研究背景及目的

今日電腦與資訊技術蓬勃發展的數位時代，網際網路已成為現代生活中不可或缺的重要角色，更帶動人類文明往新的資訊紀元(Information Era)推進。拜科技之賜，各種媒體資料的數位化；透過網際網路管道，大量且豐富的數位內容(Digital Content)得以無遠弗屆地傳播。就現況而言，各式各樣的資訊於網際網路中流通，資訊的傳播不再單純藉由傳統平面媒體，人們亦漸漸習慣經由網路找尋所要的資料。資訊的蒐集變得方便，然而亦衍生相關問題，如「資訊爆炸(Information Explosion)」。

龐大的資訊量，使得搜尋及辨別有用資訊的困難度大幅提昇，如何快速且有效地獲得真正符合自身需求的資訊，亦是目前熱門的研究議題。為解決此類問題，使用者藉由輔助工具的幫助，得以快速獲知資料的意涵，期能正確地判斷是否符合自身的需求。相關的輔助工具有 1) 搜尋引擎(Search Engine)及 2) 自動摘要(Automated Summarization)。其中，搜尋引擎扮演『資訊過濾器(Information Filter)』的角色，其功用乃是分析檢索條件(Query)，搜尋與檢索條件相關的資料；自動摘要系統則扮演『資訊監督者(Information Spotter)』的角色，其功用在於分析、統整相關的資料，以簡明的形式呈現，以幫助使用者在最短時間得知資料內容的意義[20]。

自動摘要乃是從原始資料中精鍊出最重要資訊的過程，其結果即為該原始資料的精簡化版本，且可作為人們或其他資訊系統的判斷與決策依據[34]。

> 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).

自動摘要系統依原始資料之性質可分為：

- 文件摘要(Text Summarization) – 原始資料為純文字；
- 多媒體摘要(Multimedia Summarization) – 原始資料為影像及聲音；
- 複合性摘要(Hybrid Summarization) – 原始資料綜合純文字和多媒體。

文件摘要研究起源於 1950 年代，使用的技術可大致分為下列幾種：

- 1950至1960年代，研究方法著重於寫作格式(Genre)的分析。相關研究，如[16][32]。舉例來說，語句中含有提示片語(Cue Phrase)，如『In Summary』或『In Conclusion』，則該語句可視為摘要語句。此類技術之優點在於簡單容易，然而卻與文件類型相關，技術重複利用性並不高。
- 1970至1980年代初期，研究方法轉而利用人工智慧來建構知識的表示法

(Knowledge Representation)，藉以達到分析文件主題及涵義的目的。相關研究，如[53][14]。此類技術以模板(Template)來辨認人物、地點及時間等基本要素(Entity)，並透過知識模型的推演偵測主題及產生摘要。其缺點在於模板的定義不夠詳盡，導致摘要意義上與原內容可能有所出入。

- 1990年代開始，資訊擷取(Information Retrieval, IR)技術被廣泛應用。相關研究，如[4][8][20][24][27][45][48]。然其分析只著重於字詞層面(Word-level)，並沒有考慮同義詞(Synonymy)、一詞多義(Polysemy)及字詞依屬(Term Dependency)關係等語意層面(Semantic-level)分析。

文件摘要可分為單文件摘要(Single Document Summarization)及多文件摘要(Multi-Document Summarization)。單文件摘要將文件精簡化、重點化，著重於刪減無用資料。相關研究，如[1][2][4][20][24][27][30][40][48][52]。多文件摘要針對多篇探討相似主題的文件(Topically-related Documents)，著重於刪減、過濾無用且重複的資料。相關研究，如[7][19][23][31][33][37][45][46]。一般來說，理想的多文件摘要系統必須滿足以下要求[19]：

- Clustering – 具有將相似的文件或段落群集成相關資訊的能力。
- Coverage – 摘要應涵蓋不同文件中所有主題的能力。
- Anti-redundancy – 刪減摘要中各段落間重複資料的能力。
- Summary Cohesion Criteria – 將各項資料結合成一致性高且適合閱讀的摘要，包含各段落的排序。
- Quality – 摘要結果須具有高可讀性，並具有相關性高且內容豐富的資訊。
- Identification of Source Inconsistence – 辨認不同文件所提供不同資訊、錯誤及不一致性的能力。
- Summary Updates – 追蹤時間性事件發展能力，以提供使用者最新的資訊。
- Effective User Interface – 提供與使用者互動的介面，如個人化摘要及呈現。

由語言的角度來看，文件摘要亦可分為單語言文件摘要(Mono-lingual Summarization)與多語言文件摘要(Multi-lingual Summarization)。多語言文件摘要，如[55][10][57][29]。多語言摘要係指文件來源可能為不同語言或單文件中包含不同語言等，此類摘要著重於克服各語言在型態、結構及用語習慣的差異，並提供各語言間互相轉譯的能力。

本計畫為三年期研究案『多語言複合式文件自動摘要之研究』之第一年計畫。本年度計畫之目的，在於研究並發展多文件自動摘要的技術，將探討如何導出文件的結構、如何組織文件的內容及如何表示所抽取出文件抽象涵義等相關議題。

研究範疇包含：

1. 文件模型的建構及表示
2. 文件主題偵測及重要性評估
3. 摘要內容組織及排序

表格 1-1 針對上述研究項目，整理所對應的採用方法。詳細研究方法之敘述，請參考第 3.3 節中所提之 LSA-based MD-T.R.M.方法。

表格 1-1：本計畫研究項目及採用之方法

研究項目	採用方法
文件模型的建構及表示	■ Latent Semantic Analysis ■ Text Relationship
文件主題偵測及重要性評估	■ Global Bushy Path ■ Average Similarity ■ Spreading Activation
摘要內容組織及排序	■ 考慮主題關聯之排序方式

2. 國內外相關研究

2.1. 相關研究介紹

網際網路興起後，文件摘要從 80 年代的蟄伏期復甦又蓬勃發展起來。總括而言，大部分文件摘要之研究著重於單文件摘要。目前的研究涵蓋資訊擷取(Information Retrieval)與自然語言處理(Natural Language Processing)等技術，除詞性分析、詞組分析，更包含運用 WordNet [39]等領域知識(Domain Knowledge)輔助，進行資訊萃取(Information Extraction)等較複雜的語意分析。相關的研究單位很多，著名的有卡內基美濃大學(Carnegie-Mellon University)、康乃爾大學(Cornell University)、南加州大學(Southern California University)、密西根大學(Michigan University)、哥倫比亞大學(Columbia University)等；國內大學，如台灣大學、清華大學及交通大學都有相關的研究。美國國防部高等研究計畫機構亦舉辦大型的文件摘要比賽；如，SUMMAC [50]與 DUC [15]，皆有詳細的介紹文件摘要方法及評估的標準。以下分別介紹著名的相關研究成果。

哥倫比亞大學發展 PERSIVAL 系統[43]，將病人就醫紀錄與資料庫中相關的醫學影像和聲音作關聯，並透過適當的版面設計將結果以摘要的形式呈現出來。文件摘要不再只侷限於純文字資訊摘取，更包括內容上下文中影像與聲音的關聯及呈現。該系統亦建置個人化環境(Personalized Environment)，提供不同層面的摘要內容。舉例而言，病人及親屬所看到的摘要內容與醫療人員所看到的摘要內容，其深度及廣度皆有所不同。該系統更提供專業術語(Terminology)的轉換，使摘要的內容與使用詞彙，隨使用者背景知識與興趣的差異而有所不同。同時，提供文件的視覺化(Visualization)摘要，以方便使用者快速了解各文件所提及的主題與各文件的差異性。

哥倫比亞大學發展 Columbia Summarizer [36]。該系統依據不同類型的文件整合不同的摘要技術；文件類型分為 1) 單一事件(Single Event)；2) 相關多事件(Multiple Related Event)；3) 傳記(Biography)；4) 相關事件的討論議題(Discussion Issue)等。圖 2-1 為該系統的系統架構圖，共分為三個部分 – Preprocessing、Routing 及 Summarizer Module。Preprocessing 將輸入的文件轉換成統一的 XML 格式；Router 則依據輸入文件的類型轉送給適當的摘要器；MultiGen [37]處理具有相同事件的文件集；DEMS (Dissimilarity Engine for Multi-document Summarization) [49]則依據輸入文件的特徵分析，處理多事件及傳記等類型的摘要。目前，該系統已整合於線上新聞摘要系統 NewsBlaster [11]，提供每日新聞主題偵測、追蹤及摘要服務。

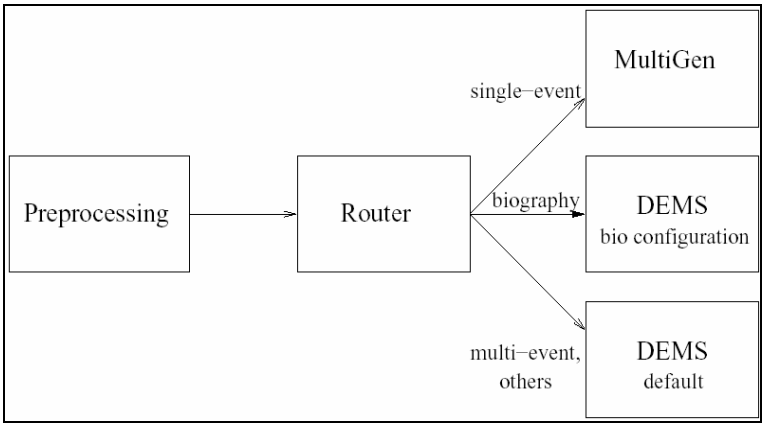

圖 2-1：Columbia Summarizer 系統架構[36]

南加州大學發展一套摘要系統 GLEANS [13]。該系統將文件集中所有文件所描述的物件(Entity)及物件間關聯(Relationship)抽取出來，並以資料庫表格的表示法儲存。同時，利用分類技術將文件集分為四種不同的類別，分別為 1) 單人物(Single Person)；2) 單事件(Single Event)；3) 多事件(Multiple Event)；4) 天然災害(Natural Disaster)。根據不同的類別，依據事先定義好的模板(Template)產生長度較短的內容提要。最後，依照不同的類別，考慮內容一致性(Coherence)的關係以產生最後的摘要內容。GLEANS 之特點在於利用模板產生品質較佳的摘要(Abstract)。圖 2-2 為其系統架構。

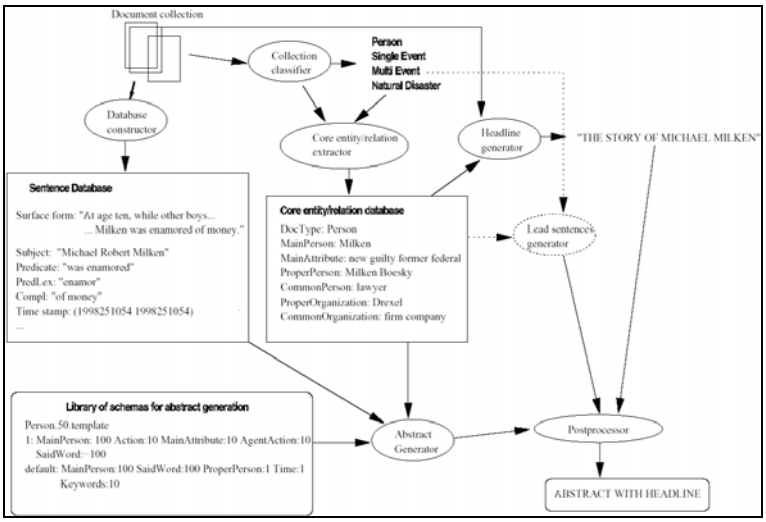

圖 2-2：GLEANS 系統架構圖[13]

南加州大學同時發展 NeATS (Next Generation Automated Summarization) [31]。該系統分析字詞的重要性，包含單字詞(Unigram)、二字詞(Bigram)及三字詞(Trigram)，自動擷取出文件集中與主題相關的部分，並將結果以一致性的順序呈現。NeATS 系統可產生一般性摘要(Generic Summary)，且其結果會依照使用者喜好的主題有所調整。其運作步驟如下：1) 擷取主題特徵(Topic Signature) [24]及主題語句[24]，並依照語句的重要性排序(Ranking)；2) 利用 OPP (Optimal

Position Policy) [24]將不具重要性之語句移除；3) 強化一致性(Cohesion)及連貫性(Coherent)；4) 利用 MMR [8]篩選語句以減少重複性，並保留下固定的摘要語句數目的語句；5) 強化時間順序(Chronological)的一致性；6) 將摘要結果格式化並輸出。

密西根大學對於多文件摘要提出各種不同的摘要技術，包含 Centroid-based Approach [45]、Cross-Document Structure Theory [54]、Revision-based Approach [42]及 Event-based Approach [12]。同時，亦實作三個不同類型之線上摘要系統，分別為 MEAD [38]、NewsInEssence [41]及 WebInEssence [44]。目前，MEAD 已經發展為適用於一般領域的多文件摘要模組，其研究目的為 1) 發展中英文的多文件摘要模組；2) 發展單/多文件摘要系統的評估工具；3) 實驗並評估四種不同的摘要標準，包含 Co-Selection、Content-based、Relative Utility 及 Rank Preservation。NewsInEssence 為應用於新聞領域的摘要系統，提供新聞文章的主題群集(Topic Clustering)、即時搜尋、文章摘要及使用者互動(User Interaction)等功能。WebInEssence 則整合文件摘要技術於搜尋技術中。

德州大學開發 GISTexter 摘要系統[21]，該系統利用資訊萃取(Information Extraction, IE)系統 – CICERO [22]與外在的知識，如 WordNet [39]，抽取文件中所提到的物件與事件，並且建構物件與事件的關聯模型。摘要的產生方式則是套用既有的模板來產生內容連貫性與一致性高的摘要。對單文件來說，主要是透過語句抽取(Sentence Extraction)的方式；對多文件來說，則是將分散於多文件中的相同主題(Shared Topics)抽取出來。圖 2-3 為 GISTexter 的系統架構圖。

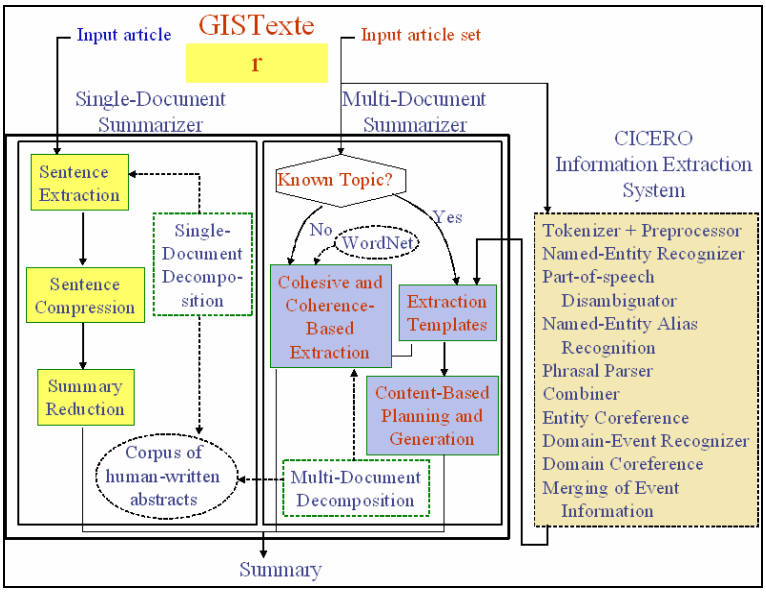

圖 2-3：GISTexter 系統架構[21]

馬里蘭大學發展適用於大型文件集(Large Corpus)的摘要系統，名為 XDoX [23]，其處理的文件集大小約為 50-500 篇文章。該系統利用分群技術(Clustering)將文件集分為幾個有意義的主題，接著以段落為單位，依據段落與主題群集的相關程度作分類，最後依據不同的群集產生摘要，其流程如圖 2-4 所示。另外，XDoX 提供使用者兩種不同的摘要結果，一為詳細的摘要，提供較豐富的資訊；另一個則依據使用者的需求，如壓縮比等等，提供資訊量較少的摘要。

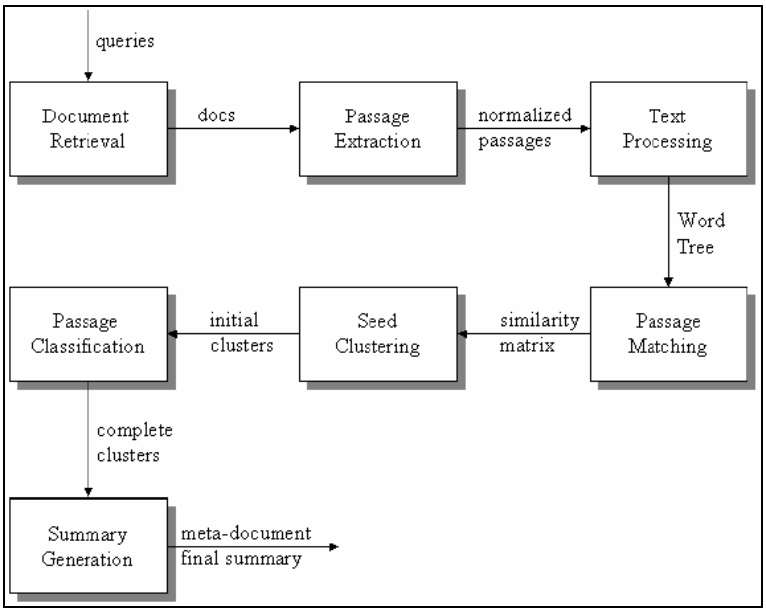

圖 2-4：XDoX 系統架構[23]

微軟劍橋研究中心利用語彙鏈結(Lexical Bonds)技術產生摘要[25]。其方法共分為三個步驟，分別為 1) 分析(Analysis)；2) 轉換(Transformation)；3) 合成(Synthesis)。首先，將文件的特徵抽取出來；共考慮 12 種文件的特徵，如語句在文件中的位置等。接著，利用 SVM (Support Vector Machine)將文件中的重要語句挑選出來。最後，依照語句出現在文件中的順序排序以產生摘要。該方法的好處在於利用統計與機器學習的技術，並且透過語彙鏈結將文字的語意納入考量，可產生連貫性較佳的摘要。

其他相關研究，如[46]針對同一事件的新聞文件作摘要。他們利用專有名詞識別技術(Named Entity Identification)擷取人名、地名及組織名等資訊，並由新聞摘要的語料庫中學習摘要的產生方式；同時，利用事先定義好的摘要模板來產生摘要。McKeown et al. [37]將機器學習與統計的技術整合應用於多文件摘要的研究。他們的方法可分為三個部分：1) 主題辨識(Theme Identification) [18] – 透過分群技術將文件中的主題(Theme)抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [5] – 將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) [37] – 利用 FUF/SURGE [17][47]將所摘錄出來的重要字詞重新組合以產生流暢的摘要。國內相關研究，如台灣大

學[56]，將中文新聞文件拆解成小句(Sub-sentence)，同時萃取名詞與動詞關鍵詞。接著，利用關鍵詞計算任兩小句間關聯強度，將關聯度大於門檻值之小句作成連結，最後評估小句連結並將重要的小句取出作為摘要。清華大學[55]，提出一個可調式中文文件摘要系統，該系統包含三個部分，分別為 1) 文件分群；2) 文件內容分析；3) 摘要呈現。其概念乃是將語句分群，以抽取出代表某事件的重要語句；接著，去除重複的資訊，同時標示重點後將新聞摘要呈現給使用者。

本計畫研究團隊亦對於單文件摘要進行相關研究，提出兩種新的文件摘要方法，以摘錄文件中重要語句，分別為 Modified Corpus-based Approach (MCBA) [52] 及 LSA-based T.R.M. Approach (LSA+T.R.M.) [52]。MCBA 基於統計模型與特徵分析，以評估語句的重要性。考慮的特徵，分別為語句位置(Position)、正面關鍵詞(Positive Keyword)、負面關鍵詞(Negative Keyword)、向心性(Centrality)及與標題相關度(Resemblance to the Title)。我們提出三個新的想法：1) 利用語句位置重要性分級提高不同語句位置的重要性；2) 利用詞彙關聯(Word Co-occurrence)分析文件中新詞，並將新詞加入關鍵詞的重要性計算；3) 利用基因演算法(Genetic Algorithm)找出適合之語句權重計算方式(Score Function)。LSA+T.R.M.利用潛在語意分析(Latent Semantic Analysis)技術，以擷取文件概念結構(Conceptual Structure)，即語意矩陣(Semantic Matrix)，可達到進行語意層面分析之目的。同時，利用語意矩陣導出語句表示式(Sentence Representation)，以建構主題相關地圖(Text Relationship Map)。最後，透過主題相關地圖，篩選重要的語句成為摘要。針對語意矩陣的建構，我們考量單文件層面(Single-document Level)及文件集層面(Corpus Level)，並比較兩種模式之適用性。實驗收集 100 篇關於政治類的中文文件。實驗結果顯示，我們所提的方法有良善的表現。當壓縮比(Compression Ratio)為 30%時，平均而言，F-measure [3]分別為 0.5151 及 0.4242。

2.2. 相關文獻探討

2.2.1. MEAD

MEAD [38]接受分群過後的文件集[1]，以語句(Sentence)為單位，針對每個文件群(Document Cluster)抽取出具有代表性的語句為摘要。方法如圖 2-5 所示。MEAD 考慮文件群中每個語句與群中心(Centroid)的相關度、語句的位置及該語句與所屬文件中首句的相似度，以評估每個語句的重要性。同時對每個文件群抽取出 n_i*r 個語句，以組成摘要；其中，n_i 代表 $Cluster_i$ 中語句的總數，r 代表壓縮比。

[1] MEAD 接受相關的文件集，以產生摘要。然此處所提及之相關文件集，實為考慮 loosely-related documents。

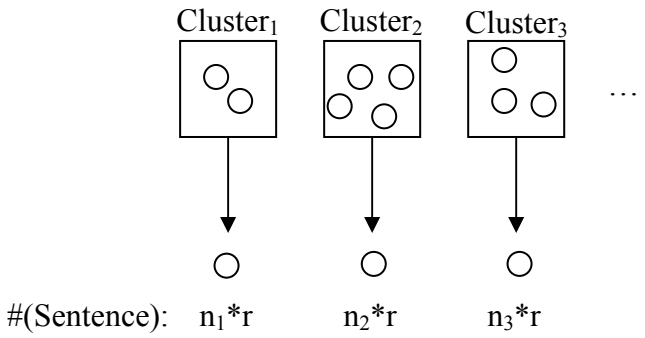

圖 2-5：MEAD 運作原理示意圖

評估語句的重要性，MEAD 考慮以下三個特徵，分別為 1) 語句所在的文件群；2) 語句於文件中的位置，通常出現在文件中的首句可視為代表整篇文章，因此加重這些語句的重要性；3) 語句如果與首句有關的話，亦加重該語句的重要性。最後，MEAD 以線性組合(Linear Combination)綜合地評估語句的重要性，如方程式 2-1：

$$S_i = w_1 \times C_i + w_2 \times F_i + w_3 \times L_i$$

方程式 2-1：MEAD 語句重要性評估[38]

其中，C_i 代表語句所在文件群的群中心(Centroid)的相關度，F_i 代表跟所屬文件之首句的相似度，L_i 代表該語句是否為所屬文件的首句。一般而言，MEAD 使用的首句加重計分法，比較適用於藝術類的文章，或是新聞文章[2]；如果文件集是為其他領域，例如技術類的文件，則首句加重計分法要再調整才合適。

2.2.2. Theme Recognition

McKeown et al. [37]認為主題相關的文件集中，存在有許多不同的主題(Theme)；依著此假設將機器學習與統計的技術整合應用於多文件摘要的研究。他們的方法，分為三個部分：1) 主題辨識(Theme Identification) [18] – 透過分群技術將文件中的主題(Theme)抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [5] – 將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) [37] – 利用 FUF/SURGE [17][47] 將所摘錄出來的重要字詞重新組合以產生流暢的摘要。

首先，考慮以下特徵以決定兩段落的相似度，進而利用分群法將找出主題，即相似段落的集合。

- Word co-occurrence：假如段落中有許多相似的字，則兩個段落可視為相似。

[2] 此類文章通常於第一段第一句說明整篇文章的重點。因此，首句之重要性必須加重考慮。

- Matching noun phrases：利用 LinkIt [51]判斷是否互相關聯的名詞片語群組。
- WordNet synonyms：使用 WordNet [39]找出同義詞組。
- Common semantic classes for verb：判斷具有同一語意的動詞詞組。

接著，利用 Information Fusion 的技術，從主題中萃取出具有代表性的詞組或片語。接著，依照出現在文章中的次序，對片語排序。最後，藉由 FUF/SURGE 自然語言產生器生成完整語句。FUF(Functional Unification Formalism)利用 SURGE 產出句法樹(Parsing Tree)，接著，藉由句法樹的轉換，以產生新的語句。

2.2.3. MMR 及 MMR-MD

MMR (Maximal Marginal Relevance) [8]適用於單文件摘要，可用於降低摘要中具有相同涵義的語句，即減少重複性資訊。其概念乃是對所挑選出與 Query 相關的語句重新排序，以符合具有最大相關度及最大差異度的特性。排序方式如方程式 2-2：

$$MMR \overset{def}{=} Arg \max_{S_i \in R \setminus S} [\lambda Sim_1(S_i, Q) - (1-\lambda) \max_{S_j \in S} Sim_2(S_i, S_j)]$$

方程式 2-2：Maximal Marginal Relevance [1]

其中，S 代表以挑選出的語句集合，S_i 代表某個語句，Q 代表 Query，$Sim_1(S_i,Q)$ 計算 S_i 與 Q 的相似度，$Sim_2(S_i,S_j)$計算 S_i 與 S_j 的相似度。

MMR-MD [19]延伸 MMR 的概念，針對多文件摘要提出適合的排序方式。主要目標是使摘要語句對文件的主題和 Query 有極高的相似度，同時能夠降低摘要中具有重覆意思的段落數目。MMR-MD 同時考慮到時間順序、專有名詞、對主題的相似度以及代名詞的 Penalty。其挑選段落的方式，如方程式 2-3：

$$MMR-MD \overset{def}{=} Arg \max_{P_{ij} \in R/S} [\lambda Sim1(P_{ij}, Q, C_{ij}) - (1-\lambda) \max_{P_{ij} \in S} Sim_2(P_{ij}, P_{nm}, C, S)]$$

方程式 2-3：Maximal Marginal Relevance – Multiple Document [19]

其中，$Sim_1(P_{ij},Q,C_{ij})$計算 P_{ij} 與 Q 的相似度，同時衡量與段落所在的文件群的相關度；$Sim_2(P_{ij},P_{nm},C,S)$計算 P_{ij} 與 P_{nm} 的相似度，其中 P_{nm} 為一以挑選出之段落。上述兩相似度的計算方式，整理於表格 2-1。

MMR-MD 希望能使的摘要中的段落儘可能的相似於 Query，但其所選到的段落間要儘可能的不相似。λ則是用來控制要取與 Query 相似度高的段落，但彼此之間的重複性可能也高，或是要與段落相似度稍低的段落，但彼此之間的重複性也低。適當的λ值可以找到兼具主題但又不會有過多重複性段落為摘要。

表格 2-1：MMR-MD 中 Sim_1 及 Sim_2 的計算方式[19]

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_1*(P_{ij}\cdot Q)+w_2*coverage(P_{ij}, C_{ij})+w_3*content(P_{ij})+w_4*time\_sequence(D_i, D)$$

$$Sim_2(P_{ij}, P_{nm}, C, S, D_i) = w_a*(P_{ij}\cdot P_{nm})+w_b*clusters\_selected(C_{ij}, S)+w_c*documents\_selected(D_i, S)$$

$$coverage(P_{ij}, C) = \sum_{k\in C_{ij}} w_k * |k|$$

$$content(P_{ij}) = \sum_{W\in P_{ij}} w_{type}(W)$$

$$time\_sequence(D_i, D) = \frac{timestamp(D_{maxtime}) - timestamp(D_i)}{timestamp(D_{maxtime}) - timestamp(D_{mintime})}$$

$$clusters\_selected(C_{ij}, S) = |C_{ij} \cap \bigcup_{v,w:P_{vw}\in S} C_{vw}|$$

$$documents\_selected(D_i, S) = \frac{1}{|D_i|} * \sum_{w} [P_{iw} \in S]$$

where
Sim_1 is the similarity metric for relevance ranking
Sim_2 is the anti-redundancy metric
D is a document collection
P is the passages from the documents in that collection (e.g., P_{ij} is passage j from document D_i)
Q is a query or user profile
$R = IR(D, P, Q, \theta)$, i.e., the ranked list of passages from documents retrieved by an IR system, given D, P, Q and a relevance threshold θ, below which it will not retrieve passages (θ can be degree of match or number of passages)
S is the subset of passages in R already selected
$R\backslash S$ is the set difference, i.e., the set of as yet unselected passages in R
C is the set of passage clusters for the set of documents
C_{vw} is the subset of clusters of C that contains passage P_{vw}
C_v is the subset of clusters that contain passages from document D_v
$|k|$ is the number of passages in the individual cluster k
$|C_{vw} \cap C_{ij}|$ is the number of clusters in the intersection of $C_{vw} and C_{ij}$
w_k are weights for the terms, which can be optimized
W is a word in the passage P_{ij}
type is a particular type of word, e.g., city name
$|D_i|$ is the length of document i.

2.2.4. Graph Matching

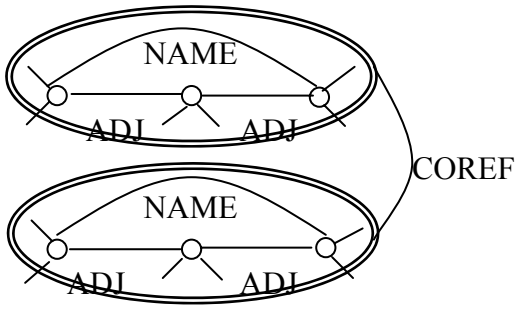

圖 2-6：文件圖形模型範例

Mani et al. [33]將文件表示成圖形(Graph)，其中，每個節點代表一個關鍵詞(Term)，節點與節點間用不同的關係連接起來，包含 1) 片語關係(PHRASE)；2) 形容詞關係(ADJ)；3) 同義關係(SAME)；4) 關聯關係(COREF)。文件圖形模型，

如圖 2-6 所示。

首先，賦予每個節點一權重(Weight)，權重值初始為該關鍵詞的 TF-IDF [3]值。接著，利用 Spreading Activation [9]演算法，透過節點間相連的連結權重變更節點的權重值，以找出與 Query 相關的節點。接著，比較兩兩文件圖形模型的相似度(Commonality)及差異性(Difference)。他們提出 FSD (Find Similarities and Differences)演算法，以找出兩圖形中相似或差異的節點。透過以下演算法，將節點分成兩群，一群為相似的節點，另外一群則為具有差異的節點。

1) Common = {c | concept_match(c, G1) & concept_match(c, G2)}
2) Differences = (G1 ∪ G2) – Common
concept_match(c, G) is true iff c1 ∈ G such that c1 is a topic term or c and c1 are synonyms.

最後，透過分析 Common 及 Difference 中的關鍵詞，計算語句的重要性，並挑選出重要的語句當成摘要結果。語句重要性的計算方式，如方程式 2-4：

$$score(s)=\frac{1}{|c(s)|}\sum_{i=1}^{|c(s)|} weight(w_i), \quad where\ c(s)=\{w \mid w \in Common \cap s\}$$

方程式 2-4：語句重要性計算[33]

3. 研究方法

我們以先前研究單文件摘要所提出的方法 – LSA-based T.R.M. Approach [52]為基礎，加以改良以適用於多文件摘要的研究，同時提出段落重要性評估的三種模型。本節中，首先介紹潛在語意分析(Latent Semantic Analysis) [28]與主題關係地圖(Text Relationship Map) [48]，最後說明我們所提出的多文件摘要技術模型 – LSA-based MD-T.R.M. Approach。

3.1. 潛在語意分析 (Latent Semantic Analysis)

潛在語意分析 (Latent Semantic Analysis) [28]為以數學統計為基礎的知識模型，其運作方式與類神經網路(Neural Net)相似。不同的是類神經網路以權重的傳遞(Propagation)與回饋(Feedback)修正本身的學習；潛在語意分析則以奇異值分解(Singular Value Decomposition, SVD)與維度約化(Dimension Reduction)為核心作為邏輯推演的方式，其原理如圖 3-1 所示。潛在語意分析將文件或文件集表示為矩陣，透過 SVD 將文件所隱含的知識模型，抽象轉換到語意空間(Semantic Space)，再利用維度約化萃取文件知識於語意空間中重要的意涵。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。

潛在語意分析的應用非常廣泛，包含資訊擷取、同義詞建構、字詞與文句相關性判斷標準、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。

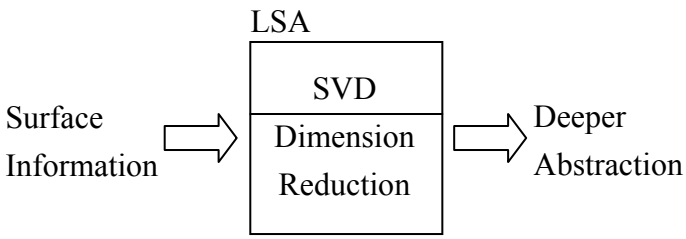

圖 3-1：LSA 工作原理

做法上，首先將文件集(Corpus)中所有文件的 Context[3] 建構為 Word-by-Context 矩陣(A)。矩陣中的每個元素($a_{i,j}$)，即某關鍵詞(W_i)在某 Context (C_j)中的權重或出現頻率。接著，透過奇異值分解將 A 分解轉換成三個矩陣乘積，即 $A=USV^T$。其中，S 代表語意空間(Semantic Space)，U 代表關鍵詞於此語意空

[3] Context 可視需求定義為語句(Sentence)，段落(Paragraph)，或文件(Document)的層面來考量。

間中的表示法，V^T 則代表 Context 於此語意空間中的表示法。再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 $A'=U'S'V'^T$，可更進一步導出 Word-Word、Word-Context 或 Context-Context 的關聯強度。值得一提的是，潛在語意分析具有知識推演的能力；如果將原始矩陣中的任一數值改變，其結果會影響到最後重建的矩陣，且影響的範圍不單為原先經過改變的數值，更會影響到矩陣中的其他數值。

3.2. 主題關係地圖 (Text Relationship Map)

主題關係地圖(Text Relationship Map) [48]將文件集中文件間關聯度表示成關係地圖。作法上將每篇文件以關鍵詞的向量表示法(Vector)表示，計算兩兩文件的相似度(Similarity)；當相似度大於臨界值時，表示此兩篇文件存在連結關係(Semantic Related Link)。依此原則可以建構出所有文件間的關係地圖。舉例來說，圖 3-2 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 的相似度則低於臨界值，因此於主題關係地圖中並不存在連結。一般來說，具有連結的文章，可說它們之間具有關聯性。

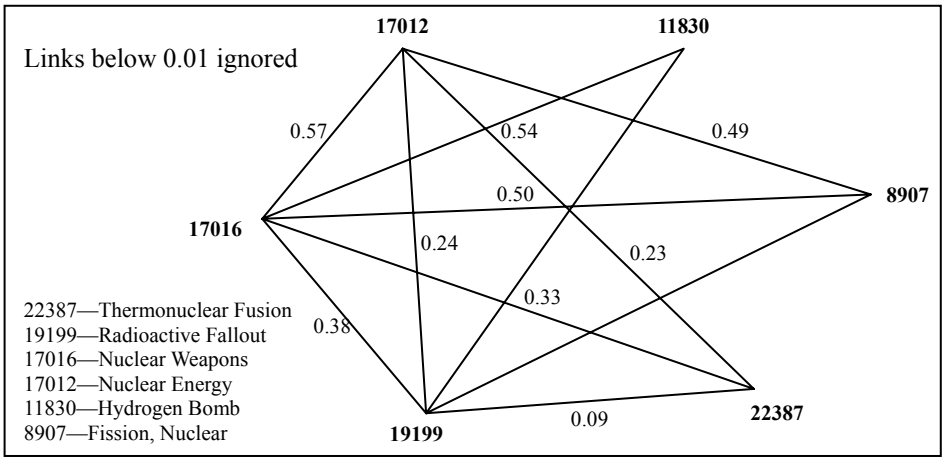

圖 3-2：主題關係地圖的範例[48]

[48]將主題關係地圖的概念應用於單文件摘要研究。以每個段落(Paragraph)為單位計算兩兩段落的相似度，建構主題關係地圖[4]。當某個節點具有的連結數愈多，則代表該節點所對應的段落和整篇文件中主題的相關度愈高。[48]依據連結數目的多寡來決定摘錄段落順序，並提出以下三種方法以產生單文件摘要：

1. Global Bushy Path

[4] 地圖上每個節點為文件中的某個段落；兩節點的連結，則表示兩節點的相似度大於臨界值。

首先定義任一節點的 *Bushiness* 為該節點與其他節點的連結數目；擁有越多關聯連結的節點，表示該節點所對應的段落與其他段落所討論的主題相似，因此，該段落可視為討論文件主題的段落。Global Bushy Path 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。接著，挑選排名前 *K* 個段落(Top-*K*)，即為該文件的摘要。

2. Depth-first Path

Depth-first Path 選取某個節點 – 可能為第一個節點或是具有最多連結的節點，接著每次選取於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。

3. Segmented Bushy Path

Segmented Bushy Path 分為兩個步驟，首先分析文件結構進行文件結構切割(Text Segmentation)。接著針對每個 Segmentation 個別利用 Global Bushy Path 來選取重要的段落。為了保留所有 Segmentation 的內容，每個 Segmentation 至少要挑選出一個段落納入最後的摘要。

3.3. LSA-based MD-T.R.M. Approach

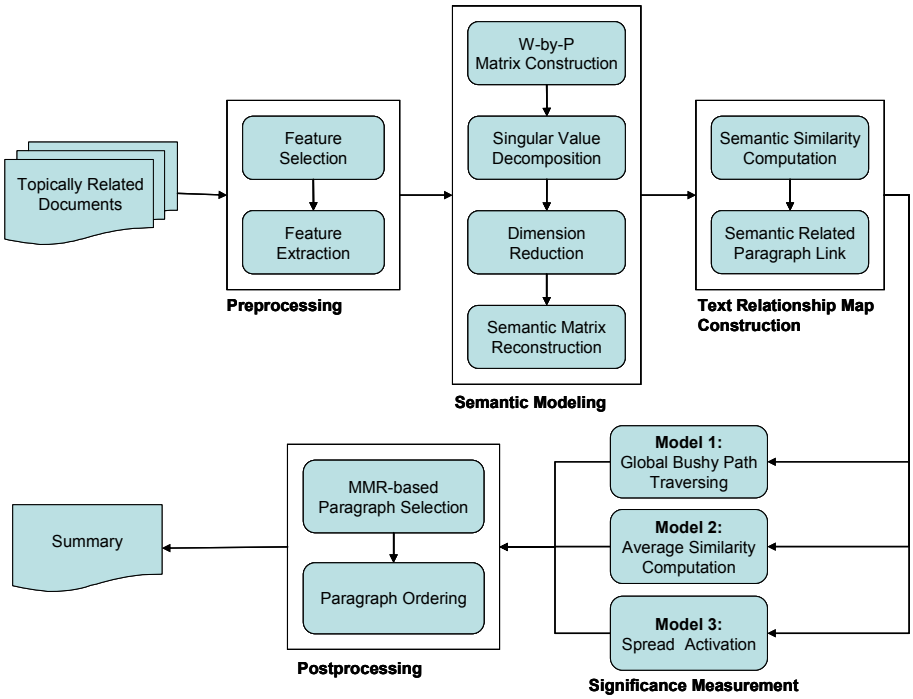

圖 3-3：多文件摘要架構

本節以我們先前對於單文件摘要所提出的方法 – LSA-based T.R.M. Approach [52]為基礎，加以改進以適用於多文件摘要，並提出段落重要性評估的三種模型。系統架構如圖 3-3 所示[5]，共包含五個模組，分別為前處理(Preprocessing)、語意模型建立(Semantic Modeling)、主題關係地圖建構(Text Relationship Map Construction)、段落重要性評估(Significance Measurement)及後處理(Post-processing)。以下分別說明各個模組之功用。

3.3.1. 前處理(Preprocessing)

前處理包含兩個步驟，分別為特徵選取(Feature Selection)及特徵擷取(Feature Extraction)。

■ 特徵選取

我們以段落(Paragraph)為單位，考慮所有的單字詞(Unigram)、二字詞(Bigram)及三字詞(Trigram)。針對二字詞及三字詞，利用 Mutual Information [35]計算其代表性，以篩選不具代表性之特徵，計算方式如方程式 3-1[6]：

$$MI(x,y) = \frac{P(x,y)}{P(x)p(y)}$$

方程式 3-1：x 與 y 組合的 Mutual Information [35]

其中，x 與 y 為相鄰之兩個單字詞[7]，$P(x)$為 x 出現於文件集的個數，$P(y)$為 y 出現於文件集的個數，$P(x, y)$則為 x 與 y 共同出現的個數。為了更進一步篩選出具有代表性的特徵，針對每個特徵計算其 IDF (Inverse Document Frequency) [3]，其計算如方程式 3-2 所示：

$$IDF(w_j) = \log\frac{N}{n}$$

方程式 3-2：w_j 之 IDF 值計算公式

其中，w_j 為一特徵關鍵詞，N 為文件集中段落的總數，n 為 w_j 出現的段落總數。當 IDF 值大於預設的臨界值，則表示該特徵具有代表性。

■ 特徵擷取

---

[5] 本計畫所提之多文件摘要架構，乃延伸先前研究所提出適用於單文件摘要之 LSA-based T.R.M. Approach [52]，利用潛在語意分析(LSA) [28]與主題相關地圖(Text Relationship Map) [48]作為文件分析模型。

[6] 方程式 3-1 為計算二字詞之 MI 值。計算三字詞之 MI 值則為 MI(x, y, z)。

[7] 考慮二字詞或三字詞時，以段落為 window。

每個特徵的重要性，除了考慮每個特徵關鍵詞於段落出現的頻率外，亦考慮每個特徵關鍵詞於文件集中的重要程度。定義其權重為 K_{ij}，其計算如方程式 3-3：

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

方程式 3-3：K_{ij} 的計算公式

假設文件集中段落的集合為 $P = \{P_j \mid P_j$ 代表某一 $P_{k,l}$，即文件 D_l 中 P_k 段落$\}$，G_i 代表特徵關鍵詞 W_i 於 P 集合中的分佈權重，L_{ij} 代表 W_i 在 P_j 中的分佈權重。假設 c_{ij} 為 W_i 出現在 P_j 中的次數，t_j 為 W_i 出現在 P 集合中的次數，則 W_i 在 P_j 中的相對頻率計算方式如方程式 3-4：

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

方程式 3-4：W_i 於 P_j 中的相對頻率 f_{ij} [6]

接著，考慮 P 集合中 W_i 的資訊分佈量(Entropy)，計算方式如方程式 3-5：

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

方程式 3-5：W_i 於 P 集合中的資訊分佈值 [6]

方程式 3-5 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 P 集合中的分佈越平均，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 只出現在某些段落， W_i 的重要性便比平均分布在 P 集合中的特徵關鍵詞來得高。最後，定義 W_i 於 P 中的總體權重 G_i，如方程式 3-6：

$$G_i = 1 - E_i$$

方程式 3-6：W_i 於 P 集合中的總體權重 G_i [6]

此外，定義 W_i 於 P_j 中的權重 L_{ij}，如方程式 3-7，其中 n_j 代表 P_j 中所含的特徵關鍵詞總數。

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

方程式 3-7：W_i 於 P_j 中的權重 L_{ij} [6]

3.3.2. 語意模型建立(Semantic Modeling)

我們以建構 Word-by-Paragraph 的矩陣作為代表文件集之語意模型。假設該

矩陣為 A，其中 a_{ij} 代表 W_i 於 P_j 的權重值[8]。接著，將矩陣 A 作奇異值分解(SVD)，使得 $A=USV^T$。對於 S 進行維度約化(Dimension Reduction)，同時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-Paragraph 矩陣表示法，其中，每個列向量(Row-Vector)代表該關鍵詞在每個段落中的權重，而每個行向量(Column-Vector)代表該段落由各個關鍵字所組成的意義。

先前提及潛在語意分析(LSA) [28]能將文章中的隱性語意(Latent Semantic)表現出來。若以潛在語意分析所導出之段落表示式計算任兩段落的相似度，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，我們以潛在語意分析所得到的段落表示式 – 行向量(Column Vector)套用在主題相關地圖(Text Relationship Map) [48]，並衡量潛在語意分析對於摘要結果的影響。

3.3.3. 主題關係地圖建構(Text Relationship Map Construction)

以潛在語意分析重建之後得到的行向量當作段落的表示法，並計算任兩向量的 Cosine 值來衡量計算任兩段落的相似度。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，若有 n 個段落的話，那麼總共的連結數目 C(n, 2)個，而最後只保留相似度高的前 1.5*n 個連結。

3.3.4. 重要性評估(Significance Measurement)

我們提出三種評估方式，以評估主題相關地圖上節點(即段落)的重要性。分別敘述如下：

- Model 1: Global Bushy Value

  Global Bushy Value (GBV)[9]為主題相關地圖上任一節點與其他節點間的連結數目；定義如方程式 3-8 所示，其中，P_i 為主題地圖上一節點。由此可知，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，該段落視為討論主題的段落。

$$GBV(P_i) = \sum_{\forall P_j, P_j \text{ has a link with } P_i} 1$$

方程式 3-8：節點 P_i 的 GBV 值

- Model 2: Average Similarity

  相較於 Model 1 只考慮到主題相關地圖上每個節點的連結個數，我們參

[8] a_{ij} 的值可透過方程式 3-3 的公式計算。
[9] 即[48]中定義之 Bushiness 值。

考[26]，並考慮每個連結權重的方式，以 Aggregate Similarity 計算每個節點的重要性，Aggregate Similarity 的示意圖如圖 3-4：

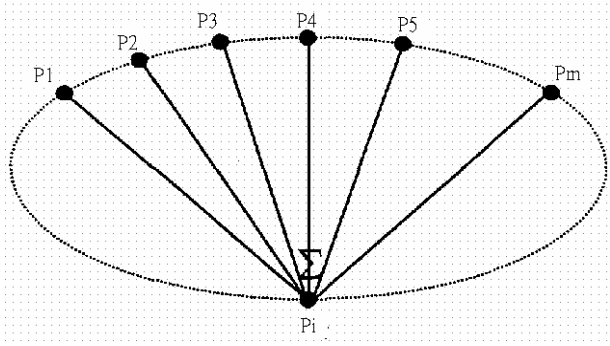

圖 3-4：計算 Aggregate Similarity 的概念圖示[26]

圖中的每個節點代表某個段落的向量表示法，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。Aggregate Similari 之計算如方程式 3-9：

$$AvgSim(P_i) = \sum_{\substack{\forall P_j \neq P_i \\ P_j \text{ has a link with } P_i}} sim(P_i, P_j)$$

方程式 3-9：P_i 的 Aggregate Similarity 的計算方式

其中，$sim(P_i, P_j)$為兩個節點間的相似度，即是計算相對應向量間 Cosine 值。計算每個節點的 Aggregate Similarity，其好處在於除了考慮到每個節點的連結個數，同時亦考慮到每個連結的權重值。

- Model 3: Spreading Activation

相較於 Model 1 及 Model 2 設計產生一般性摘要(Generic Summary)，Model 3 利用 Spreading Activation [9]產生與查詢相關之摘要(Query-oriented Summary)。作法上，首先將原始查詢($\vec{q}$)轉換為潛在語意分析所導出之語意空間表示法，轉換方式如方程式 3-10：

$$\vec{q}' = \vec{q} \times U' \times S'$$

方程式 3-10：轉換原始 $\vec{q}$ 向量為語意空間向量 $\vec{q}'$

其中，$\vec{q}$ 為原始向量表示法，U'及 S'為利用潛在語意分析所導出之語意矩陣[10]。接著，挑選出與 $\vec{q}$ 最相似的 k 個節點當作 Spreading Activation 的輸入，

[10] 假設 w-by-p(word-by-paragraph)矩陣為 A，將 A 經過奇異值分解(SVD)後可得，$A=USV^T$。經過維度約化(Dimension Reduction)，則 $A'=U'S'V'^T$。其中，U'可視為 w 於 S'的表示法，S'為一語

並透過下列演算法變更主題相關地圖中相關節點的權重。當節點的權重不再變動時，此時擁有越大權重值的節點，其與查詢($\vec{q}$)的相關度便越強，亦即越具代表性。

```
Algorithm Spread (TRM, Relevant):
    Input := Relevant = {Pi|Pi is relvant to  q'};
    sort(Input); //根據 sim(Pi,  q')作排序
    while (Continue?(Output)) {
        Node := first(Input);
        insert(Output, Node);
        Succs := ActivateSuccs(Node, TRM);
        While (Succs) {
            insert(Input, pop(Succs));
       }

Algorithm ActivateSuccs(Node, TRM):
//利用相連結點間的 weight (即 similarity)  調整 Node 之 Neighboring Nodes 的 weight
    While (<Node1, Link> := links(Node, TRM)) {
        Node1.wt = max(Node1.wt, (Node.wt * Link.wt));
    }
```

上述演算法中，每個 Node(即 P_i)的權重初始值($P_i.wt_{initial}$)計算，如方程式 3-11。假設，P_i=<$w_{1,i}$,…,$w_{M,i}$>為潛在語意分析所導出之段落表示法，$w_{j,i}$ 表示 $word_j$ 於 $paragraph_i$ 中的權重。計算 $P_i.wt_{initial}$ 所考慮的是段落中特徵關鍵詞的平均權重，以避免過長的句子因為其所擁有的關鍵字比較多，導致累計權重時造成偏差而影響真正的重要性。

$$P_i.wt_{initial} = \frac{1}{M}\sum_{i=1}^{M} w_i$$

方程式 3-11：主題相關地圖上節點的初始權重值

每個 Iteration 中，上述演算法利用相連結點間的相似度，調整相鄰節點的權重。假設目前考慮之節點為 P_i，與 P_i 有連結的節點(即 Neighboring Node)為 P_j，其權重變更如方程式 3-12。其中，$Link.wt$ 即為 P_i 與 P_j 的相似度，亦即 $sim(P_i, P_j)$。

$$P_j.wt = \max(P_j.wt, P_i.wt * Link.wt)$$

方程式 3-12：Neighboring Node 的權重變更計算

Spreading Activation 的概念與 Best-first Search 相似，運作時先將與 $\vec{q}$' 相關的節點放到 Priority Queue 中，接著依序調整 Priority Queue 中每個節點之

意空間，V'可視為 p 於 S'的表示法。

Neighboring Node 的權重。調整的方式，考慮節點上所有連結，當連結權重值越大時，則其重要性越高；由於每個節點所擁有之連結權重並不相同，因而可以達到調整節點權重的目的。當所有節點的權重皆不再變更，或是達到既定的臨界值時，則 Spreading Activation 便停止。考慮 Spreading Activation 終止的條件，可以考慮第 *i* 次 iteration 做完時，此時所有節點的總變化量與前面第 *i-k* 次 iteration 時的總變化量差值，當兩次 iteration 間總變化量差異性不大時，便可以終止。

3.3.5. 後處理(Post-processing)

後處理包含兩個步驟，分別為段落選取(Paragraph Selection)及段落排序(Paragraph Ordering)。

■ 段落選取

我們參考 Maximal Marginal Relevance (MMR)[11] [8]的概念，提出段落選取的方法，如方程式 3-13 所示：

$$PS \overset{def}{=} Arg \max_{P_i \in R \setminus S} [\lambda SIG(P_i) - (1-\lambda) \max_{P_j \in S} REL(P_i, P_j)]$$

方程式 3-13：段落選取方法

其中，S 代表已被選到之段落的集合，$SIG(P_i)$ 代表 P_i 的重要性，$REL(P_i, P_j)$ 代表 P_i 與 P_n 的關聯強度。

做法上，*PS* 依序選取出組成摘要的段落，同時評估目前衡量的段落與先前取出的段落間相關程度；此機制乃是為了去除重複性(Anti-redundancy)，以提供使用者更多的資訊。表格 3-1 整理針對前一步驟所提出的三個模型中 *SIG* 與 *REL* 的計算方式。

表格 3-1：Model 1-3 之 *SIG* 與 *REL* 的計算方式

	$SIG(P_i)$	$REL(P_i, P_j)$
Model 1	$GBV(P_i)$	$\begin{cases} \alpha & \text{if } P_i \text{ has a link with } P_j \\ 0 & \text{otherwise} \end{cases}$
Model 2	$ASim(P_i)$	$sim(P_i, P_j)$
Model 3	P_i.wt	$sim(P_i, P_j)$

■ 段落排序

[11] 請參考相關研究工作中所提及之 MMR [8]及 MMR-MD [19]。

段落排序之目的，將挑選出來的段落依據內容一致性(Cohesion)及連貫性(Coherent)重新排序，以提供使用者適合閱讀的摘要。[31]考慮 Paired Sentences 將挑選出來的語句重新排序。他們的方法如下所示，*x.y* 代表 *x* 文件中 *y* 語句：

原始順序：	考慮 Paired Sentences 順序：
4.3, 6.6, 2.5, 5.2, …	→ 4.1, 4.3, 6.1, 6.6, 2.1, 2.5, 5.1, 5.2, …

Paired Sentences 考慮每個語句與所屬文件首句的關係。舉例來說，4.1 與 4.3 有關聯，因此 4.1 會被排序到 4.3 前面。依此原則，可以將所有挑選出來的語句重新排序。然而，Paired Sentences 的方法並沒有考量到主題關聯。舉例來說，假設 4.3 與 2.1 有關聯，則上述排序方式會造成使用者閱讀的障礙。亦即，主題的轉變順序錯亂。

我們克服上述問題，提出新的段落排序方法。首先，計算選出之段落集合中，每篇文件所選出之段落的個數，以找出當作主要排序順序的文件。接著，將其他的段落與主要文件的段落作關聯，當相似度大於預設之臨界值時，則將段落歸於同一群。同時，收集無法歸於主要文件的段落集合。如圖 3-5 所示，找出 x_i 文件當作主要的排序；接著，將其他段落指定給任一 $Bucket_{ij}$，同時收集無關連的段落於 $Bucket_{other}$。最後，針對每個 *Bucket* 利用原先的順序(即方程式 3-13 的順序)排序，產生最後的摘要呈現順序。

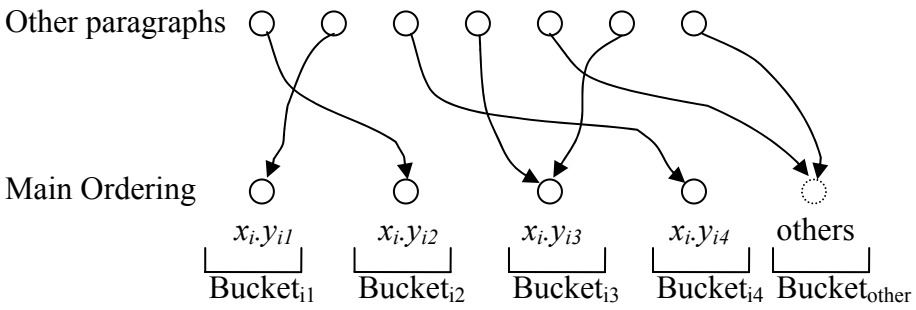

圖 3-5：我們提出之段落排序方法示意圖

4. 研究進度及成果自評

本計畫為三年期研究案之第一年計畫，我們考量相關文獻所提多文件摘要的方法之優缺點，並以先前發展之單文件摘要方法[52]為基礎，加以改進以適用於多文件摘要，同時對於段落重要性評估提出三種模型。另外，我們亦對於段落排序方式進行研究，提出新的演算法以提高摘要內容主題的連貫性。根據圖 3-3 所提之多文件摘要架構，目前已經完成下列各模組設計及實作：

1. 前處理(Preprocessing)
   - Feature Selection
   - Feature Extraction
2. 語意模型建立(Semantic Modeling)
   - Latent Semantic Analysis
   - Semantic Matrix
3. 主題關係地圖建構(Text Relationship Map Construction)
   - Text Relationship Map
   - Similarity Matrix
4. 重要性評估(Significance Measurement)
   - Model 1：Global Bushy Path
   - Model 2：Average Similarity

現階段，我們正進行重要性評估模組中 Model 3 所提之 Spreading Activation 演算法實作。預計六月底可完成所有的模組，並進行實驗評估工作。實驗部分擬以 DUC [15]提供的資料進行實驗，並以過去 DUC 結果評估我們所提之多文件摘要方法的優劣。整體而言，目前所進行之研究，依據所提計畫依序進行中，進度之掌握尚稱得宜。預計實驗完成後，將整理資料並將結果發表於國際會議論文。

5. 參考文獻

[1] Aone, C., Okurowski, M. E., Gorlinsky, J., & Larsen, B. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques. Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[2] Azzam, S., Humphreys, K., & Gaizauskas, R. (1999). Using coreference chains for text summarization. In *Proceedings of the ACL'99 Workshop on Coreference and Its Application*, Baltimore.

[3] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, UK: Addison-Wesley Longman Co Inc.

[4] Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain (pp. 10-17).

[5] Barzilay, R., McKeown, K. R., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA (pp. 550-557).

[6] Bellegarda, J. R., Butzberger, J. W., & Chow, Y. L. (1996). A novel word clustering algorithm based on latent semantic analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 172-175).

[7] Boros, E., Kantor, P. B., & Neu, D. J. (2001). A clustering based approach to create multi-document summaries. In *Proceedings of the Document Understanding Conference* (*DUC-2001*), New Orleans, LSA, USA.

[8] Carbonell, J., & Goldstein, J. (1999). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), Melbourne, Australia (pp. 335-336).

[9] Chen, C. H., Basu, K., & Ng, T. (1994). An algorithmic approach to concept exploration in a large knowledge network. *Technical report*, MIS Department, University of Arizona, Tucson, AZ, USA.

[10] Chen, H. H., & Lin, C. J. (2000). A multilingual news summarizer. In *Proceedings of the 17th Conference on Computational Linguistics*, Saarbrücken, Germany (pp. 159-165).

[11] Columbia Newsblaster: summarizing all the news on the web. Available at http://www1.cs.columbia.edu/nlp/newsblaster.

[12] Daniel, N., Radev, D. R., & Allison, T. (2003). Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL-03 Text Summarization Workshop*, Edmonton, Alberta, Canada (pp. 9-16).

[13] Daumé III, H., Echihabi, A., Marcu, D., Munteanu, D. S. & Soricut, R. (2002). GLEANS : a generator of logical extracts and abstracts for nice summaries. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[14] DeJong, G. F. (1979). Skimming stories in real time: an experiment in integrated understanding. *Doctoral dissertation*, Computer Science Department, Yale University, New Haven, CT, USA.

[15] Document Understanding Conference (DUC). Available at http://duc.nist.gov.

[16] Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM* (*JACM*), 16(2), 264-285.

[17] Elhadad, M. (1993). Using argumentation to control lexical choice: a functional unification implementation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.

[18] Eskin, E., Klavans, J., & Hatzivassiloglou, V. (1999). Detecting similarity by applying learning over indicators. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA.

[19] Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 40-48).

[20] Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'01*), New Orleans, LA, USA (pp. 19-25).

[21] Harabagiu, S. M., & Lacatusu, F. (2002). Generating single and multi-document summaries with GISTEXTER. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[22] Harabagiu, S. M., & Maiorano, S. (2000). Acquisition of linguistic patterns for knowledge-based information extraction. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (*LREC-2000*), Athens, Greece.

[23] Hardy, H., Shimizu, N., Strzaliowski, T., Ting, L., & Zhang, X. (2001). Cross-document summarization by concept classification. In *Proceedings of the Document Understanding Conference* (*DUC-2001*), New Orleans, LSA, USA.

[24] Hovy, E., & Lin, C. Y. (1999). Automated text summarization in SUMMARIST.

Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[25] Karamuftuoglu, M. (2002). An approach to summarization based on lexical bonds. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[26] Kim, J. H., Kim, J. H., & Hwang, D. (2000). Korean text summarization using an aggregate similarity. In *Proceedings of the 5th International ACM Workshop on Information Retrieval with Asian Languages,* Hong Kong, China (pp. 111-118).

[27] Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'95*), Seattle, WA, USA (pp. 68-73).

[28] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

[29] Lenci, A., Bartolini, R., Calzolari, N., Agua, A., Busemann, S., Cartier, E., Chevreau, K., & Coch, J. (2002). Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (*LREC'02*), Las Palmas, Canary Island, Spain.

[30] Lin, C. Y. (1999). Training a selection function for extraction. In *Proceedings of the 8th International Conference on Information and Knowledge Management* (*CIKM'99*), Kansas City, MO, USA (pp. 55-62).

[31] Lin, C. Y., & Hovy, E. (2002). NeATS in DUC 2002. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[32] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159-165.

[33] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.

[34] Mani, I., & Maybury, M. (eds.) (1999). *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[35] Maosong, S., Dayang, S., & Tsou, B. K. (1998). Chinese word segmentation without using lexicon and hardcrafted training data. In *Proceedings of the 17th International Conference on Computational Linguistics* (*COLING'98*) *and the 36th Annual Meeting of the Association for Computational Linguistics* (*ACL'98*) (*COLING-ACL'98*), Montreal, Quebec, Canada (pp. 1265-1271).

[36] McKeown, K. R., Evans, D., Nekova, A., Barzilay, R., Hatzivassiloglou, V., Schiffman, B., Blair-Goldensohn, S., Klavans, J., & Sigelman, S. (2002). The Columbia Multi-document Summarizer for DUC 2002. In *Proceedings of the*

Document Understanding Conference (*DUC-2002*), Philadelphia, PA, USA.

[37] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence* (*AAAI'99*), Orlando, FA, USA (pp. 453-460).

[38] MEAD. Available at http://tangra.si.umich.edu/clair/mead.

[39] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: a on-line lexical database. *Lexicography*, 3(4), 235-312.

[40] Myaeng, S. H., & Jang, D. (1999). Development and evaluation of a statistically based document system. Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[41] NewsInEssence: Interactive Multi-source News Summarization. Available at. http://www.newsinessence.com/nie.cgi.

[42] Otterbacher, J. C., Radev, D. R., & Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Philadelphia, PA, USA (pp. 27-36).

[43] PERSIVAL: Personalized Search and Summarization over Multimedia Information. Available at http://persival.cs.columbia.edu.

[44] Radev, D. R., Fan, W., & Zhang, Z. (2001). WebInEssence: a personalized web-based multi-document summarization and recommendation system. In *Proceedings of the NAACL Workshop on Automatic Summarization*, Pittsburgh, PA.

[45] Radev, D. R., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 21-30).

[46] Radev, D. R., & McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistic*, 24(3), 469-500.

[47] Robin, J. (1994). Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.

[48] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, 33(2), 193-207.

[49] Schiffman, B., Mani, I., & Concepcion, K. J. (2001). Producing biographical summaries: combing linguistic knowledge with corpus statistics. In *Proceedings of European Association for Computational Linguistics*.

[50] TIPSTER Text Summarization Evaluation Conference (SUMMAC). Available at http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac.

[51] Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics in a document. In *Proceedings of Workshop on the Computational Treatment of Nominals*, *COLING-ACL*, Montreal, Canada (pp. 70-79).

[52] Yeh, J. Y., Ke, H. R., & Yang, W. P. (2004). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management: Special Issue on ICADL 2002*. Accepted and to be appeared. (Recommended by ICADL 2002).

[53] Young, S. R., & Hayes, P. J. (1985). Automatic classification and summarization of banking telexes. In *Proceedings of the 2nd Conference on Artificial Intelligence Applications* (pp. 402-408).

[54] Zhang, Z., Blair-Goldensohn, S., & Radev, D. R. (2002). Towards CST-enhanced summarization. In *Proceedings of the 18th National Conference on Artificial Intelligence* (*AAAI'02*), Edmonton, Alberta, Canada (pp. 439-445).

[55] 陳鈺瑾 (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, 台灣.

[56] 黃聖傑 (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, 台灣.

[57] 蘇哲君 (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, 台灣.