行政院國家科學委員會專題研究計畫 成果報告

研究與開發具有文件自動分類與摘要功能之網頁查詢系統 研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2221-E-216-021-
執 行 期 間 ：99 年 08 月 01 日至 100 年 07 月 31 日
執 行 單 位 ：中華大學資訊工程學系

計 畫 主 持 人：周智勳
共 同 主 持 人：陳建宏、石昭玲
計畫參與人員：碩士班研究生-兼任助理人員：李釗旭

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 100 年 10 月 31 日

中文摘要：　本論文針對文件自動摘要進行研究，文中結合潛在語意分析(Latent Semantic Analysis)與文章詞彙關聯度及文章句子關聯度的概念，建構出關聯性矩陣，用來加強與擷取文件中的概念結構，以得到語意層面的分析，進而篩選最適合之句子，作為文件摘要的依據。實驗針對 13 大類(生活、地方、社會、政治、科技、旅遊、財經、健康、國際、教育、運動、戲劇、藝文)共 1300 篇網路新聞進行測試。在效果評估方面，提出兩個評估指標，以達到較客觀的摘要評估結果。實驗結果顯示，我們所提之方法能在較低相似度與重複性下，獲得較具代表性的摘要句。

英文摘要：

目錄

研究與開發具有文件自動分類與摘要功能之網頁查詢系統

主持人：周智勳
執行機構：中華大學資訊工程學系
執行期間：民國99年08月01日至100年07月31日
國科會計畫編號：NSC 99-2221-E-216-021

摘要

本論文針對文件自動摘要進行研究，文中結合潛在語意分析(Latent Semantic Analysis)與文章詞彙關聯度及文章句子關聯度的概念，建構出關聯性矩陣，用來加強與擷取文件中的概念結構，以得到語意層面的分析，進而篩選最適合之句子，作為文件摘要的依據。實驗針對 13 大類(生活、地方、社會、政治、科技、旅遊、財經、健康、國際、教育、運動、戲劇、藝文)共 1300 篇網路新聞進行測試。在效果評估方面，提出兩個評估指標，以達到較客觀的摘要評估結果。實驗結果顯示，我們所提之方法能在較低相似度與重複性下，獲得較具代表性的摘要句。
關鍵詞：中文文件摘要，潛在語意分析，奇異值分解，中文斷詞。

1. 前言

隨著電腦科技與數位資訊化的進步，網際網路的存在已成為現代人們不可或缺的角色。然而要從這些龐大的資訊中取得所需資訊時，要如何有效率地獲得符合個人所需資訊則變成一個很重要的議題。為了要解決上述問題，人們需要藉助額外的工具使其能在短時間之內能得到符合自身需求且有用的資料。

文件自動摘要雖然是自然語言處理(Natural Language Processing)指標之一，但是對於文章所節錄出的摘要句，在可讀性以及前後句的連慣性，卻一直不能有重大的突破。目前一般的文件摘要技術，多半以摘要文章的"某一段落"或使用"統計分析"方法，計算句子的權重分數、位置以擷取重要的句子形成摘要，但其摘要內容之正確率、可續性和整體連慣性確有其不足之處。

自動摘要研究始於1958年，由美國IBM 公司的 Luhn〔7〕開創了自動摘要研究的先河。隨後，學者開始考慮文章的句法特徵和語義特徵，建立起以人工智慧特別是計算語言學為基礎的方法。至此，自動摘要研究分為兩大陣營：基於「統計」的自動摘要和基於「意義」的理解摘要。此外，學者也試圖尋求其他的解決方法。特別是隨著機器學習、認知心理學、語言學等領域不斷湧現出新的成果，自動摘要研究也進入了一個多元化的新時代。美國 Syracuse 大學的 Liddy 提出擬人法，日本 Toshiba 公司的 Kenji Ono 等依據修辭結構研究自動摘要，蘇聯的 Skoroxod'ko〔4〕依據語句關聯網生成摘要，美國的 Kupiec〔8〕基於語料庫的方法來計算每個語句的權值，以色列 Ben Gruion 大學的 Barzilay〔9〕依據詞彙鏈產生摘要，美國多倫多大學的 Marcu〔2〕採用修辭結構樹的方法產生摘要，美國馬塞諸塞州大學採用查詢擴展的方法選取摘要。

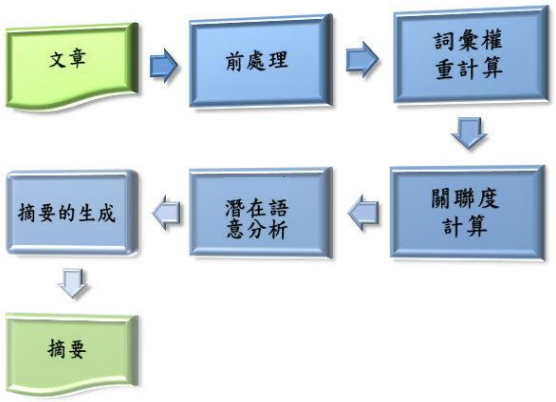
自動摘要的作法，大致上可分為「摘意」(Abstraction)與「摘句」(Extraction)兩大種〔3〕。「摘意」或稱「語意分析」(Text Understanding Analysis)，藉由語言的文法特性萃取出文章的主題以及藉著額外的資訊或是資源來達成摘要句的產生，如辭典、同義詞庫、詞性標記等〔1〕。「摘句」或稱「文字特徵統計分析」(Text Feature Analysis)，計算字詞的頻率，依照字詞的

重要程度決定原文句的重要性進而形成摘要。兩者間的差異在於後者作法較不依賴外在的知識或資源，由於「摘意」所需資源較多，所以目前的研究以「摘句」占多數。

本論文綜合此兩種方法的優點進行研究，提出一個結合統計與語言學的文件摘要方法。本論文架構除第一段的前言，第二段為方法描述，第3段說明本論文實驗結果，最後，第4段為本論文之結論。

2. 方法描述

本研究所提之文件自動摘要方法，在架構上分為五階段，如圖一所示。第一階段為「前處理」：主要的任務是先自網頁新聞文件中選取有用的特徵(關鍵詞)來代表文件，第二階段為「詞彙權重計算」：利用TFSF(Term Frequency Sentence Frequency)權重計算方法，決定每個關鍵詞彙的權重，第三階段為關聯度計算：計算詞彙與詞彙間、句子與句子間的關聯度，第四階段為「結合詞彙關聯度及語句關聯度之潛在語意分析(Combination of terms relation and sentence relation on the Latent Semantic Analysis, CTSRLSA)」：將第二、三階段的權重矩陣與關聯度矩陣以矩陣乘法運算合成出一個新的關係矩陣，最後第五階段為「摘要的生成」：將摘要句依據在原始文件中出現的先後順序依序排列。



圖一 系統架構圖

2.1 前處理

對中文文件自動摘要而言，除了蒐集中文文件之還必須從中文文件中擷取出構成中文文件的各個元素，也就是以詞彙作為特徵詞，但因為中文沒有空白斷開詞彙，故需要借助移除多餘的空白、標點符號以及斷詞的工作來處理。

2.1.1 詞彙斷詞

首先，移除空白與標點符號，此步驟移除「，」、「。」、「、」、「；」、「：」、「！」、「「」」、「『』」、「( )」、「_」以及「？」等標點符號，即可將文章段落切割成較小單位的句子，以利後續的處理。因為中文不像英文可以使用空白斷開詞彙，所以須藉斷詞的工作來處理，目前最常見的自動斷詞方法大約有三種(非人工斷詞)，即長詞優先法、法則式法及機率式方法。除此之外，還有另一種是 N-元詞（N-Gram）取詞方式，本研究即採此方法斷詞。

N-元詞選詞之所以叫選詞而不是斷詞是有原因的，前述三種方法都需要藉由人工辭典當作斷詞的基礎，其斷詞的結果當然也比較接近人工斷詞的水準。而 N-元詞的方法則不然，其不需要辭典，全部只需依靠語言的統計分析決定。故稱為選詞，不稱斷詞。N-元詞為文件中任意連續 n 個字的字串，雖然大部分的 N-元詞沒有意義，但 N-元詞仍能抓住文件的用詞，可以有效的代表該文件，使用方法若以每兩個字為一個單位切開，稱2-元詞，每三個字為一組，稱3-元詞，其餘類推，直到 N-元詞。

本篇研究將前述三種方法與 N-元詞方法結合，即將一段字串使用 N-元詞斷完詞後，不採統計方式選詞，而是採事先建置好的18萬詞庫進行比對，僅留下存在於詞庫裡的詞彙。此優點在於過濾後的詞彙都為有意義的詞彙。

2.1.2 關鍵詞擷取

關鍵詞為有意義且具代表性的片語或

詞彙，也是表示一篇文件特性最直接的方法。然而關鍵詞的認定牽涉到個人的主觀判斷，且相同的詞彙在不同的主題下，也有不同的認定，在此情況下，要比較各種方法的擷取成效，並不容易。目前關鍵詞擷取方法〔1〕大略分為三種：詞庫比對法、文件剖析法及統計分析法。此本篇論文則將前述方法之詞庫比對法與前一節所述之 N-元詞方法結合，即將一段字串使用 N-元詞斷完詞後，不採統計方式選詞，而是採與事先建置好的 18 萬詞庫進行比對，僅留下存在於詞庫裡的詞彙。這樣的好處在於過濾的速度較快且過濾後的詞彙都為有意義的詞彙，即關鍵詞。

2.2 詞彙權重計算

經過斷詞後，雖然可以取得每篇文章所包含的詞彙，仍不足以選出具較代表性的關鍵詞，主要原因是每個詞彙在文件中的重要性都不相同。所以，為了要選出較具代表性的關鍵詞，我們可利用計算詞彙位於文件中的重要性來達成，即詞彙位於文章中的權重。

在此我們提出一個用來決定關鍵詞權重的方法，「TFSF(Term Frequency Sentence Frequency)」：TF 稱「詞頻」表示一字詞在某篇文件中出現頻率次數；SF 稱「句頻」表示文章中包含有該字詞的句子次數。所以，當 SF 值越高時表示該字詞出現在較多的句子中；相反的，當 SF 值越低時表示該字詞出現在較少句子中。此篇研究在關鍵詞擷取處以「詞庫比對」的方式(18 萬詞庫)過濾出有意義的詞彙，每個有意義的詞彙即「關鍵詞彙」，為了賦予每個關鍵詞權重，我們考慮其在文章中出現的次數與其在出現在文中句子數，其計算方程式如下所示：

$$w_{ik} = tf_{ik} \cdot sf_i \tag{1}$$

其中 w_{ik} 代表關鍵詞 k_i 在文句 S_k 的權重，tf_{ik} 代表關鍵詞 k_i 在文句 S_k 中出現的次數，而 sf_i 代表關鍵詞 k_i 在文章所有句子中共出現在幾句句子中，即為文句頻率，其定義如下：

$$sf_i = \frac{n}{N} \tag{2}$$

其中 N 代表文章中句子總數，n 代表文章中含有關鍵詞 k_i 的句子數。經由此權重計算方法，我們可以計算出文章中關鍵詞的權重，最後並得到關鍵詞對應語句的權重向量矩陣。

2.3 關聯度計算

由前一節所得之關鍵詞對應語句的權重關係矩陣，我們可以對其分別計算，關鍵詞與關鍵詞間的關聯度以及語句與語句間的關聯度。目前常見用來計算關聯度的方法，大約有「Inner product」、「Dices coefficient」、「Cosine coefficient」與「Jaccard coefficient」這四種〔5〕。本研究採Jaccard coefficient計算詞彙間以及語句間的關聯程度。Jaccard coefficient視句子的重要性，決定於各個句子與其他句子連結的多寡而定。而句子間的連結關係，為彼此間有出現相同的關鍵詞彙，當句子中使用相同的關鍵詞愈多，表示所討論的主題應該是近似的，其重複性也越高。其方程式如下所示：

$$Sim(X,Y) = \frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i} \tag{3}$$

其中若計算關鍵詞間相似度則 X、Y 為文章的隨意兩個關鍵詞，x_i、y_i 為隨意兩個關鍵詞向量的第 i 個元素，t 為語句總數；若計算語句間相似度則 X、Y 為文章的隨意兩個語句，x_i、y_i 為隨意兩個語句向量的第 i 個元素，t 為關鍵詞總數。

2.4 潛在語意分析

潛在語意分析 (Latent Semantic Analysis, LSA)〔10〕不僅可用在文件知識表示以外，還可用在近於人腦用來理解文件知識的推導與認知模型。LSA其主要

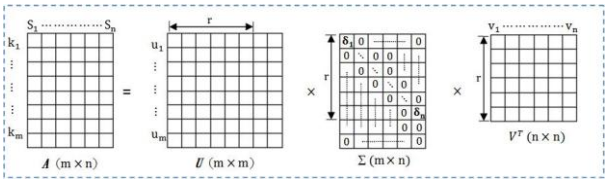
應用在一字多義或多字一義的問題上，也常用在特徵為度約化以及因子分析上。LSA主要是以奇異值分解(Singular Value Decomposition, SVD)與維度縮減化為其邏輯推導核心。LSA的基本概念是以較低維度(維度縮減)的共同語意因子(Semantic Factors)呈現原文章字詞與原文章語句間的關連。其作法是利用奇異值分解找出文章字詞對應語句的語意結構，對於字詞對應語句關係，矩陣A利用奇異值分解將會分解成三個子矩陣乘積。

奇異值分解具有將高維度的矩陣資料降低為低維度之特性，應用此方法是將文章字詞及語句投影到一個空間，此空間可以表達字詞與字詞間的關係，字詞與語句的關係以及語句與語句間的關係。假設 A 為一個 m×n 的矩陣，rank(A) = r，SVD具有以下形式:

$$A_{m\times n} = U_{m\times m} \sum_{m\times n} V^T_{n\times n} \quad (4)$$

其中 U 與 V^T 皆為正交矩陣，$\Sigma = [\Sigma_r \; O]^{\mathrm{T}}$，$\sum_r$ 為n×n的 $diag(\delta_1 \cdots \delta_n)$，$O$ 為(m-n)×n個0的矩陣，其中 $\delta_n > 0$，當n = 1, 2, …, r，而 $\delta_n = 0$，當 n > r 時，稱為奇異值 (Singular Values)，如圖二所示。U的行向量為 u_j，V 的列向量為 v_j，$A = U\Sigma V^T$ 可以表示為r個rank-one矩陣之和：

$$A = u_1\delta_1 v_1^T + u_2\delta_2 v_2^T + \cdots + u_r\delta_r v_{\mathrm{r}}^{\mathrm{T}} \quad (5)$$



圖二 奇異值分解示意圖

由此可知 A 由 U 的前r個行向量(u_r)，V^T 的前r個列向量(v_r^T)，以及 $\sum$ 的左上r×n區塊決定($\sum_r$)，由此可知 A 由 U 的前r個行向量(u_r)，V^T 的前r個列向量(v_r^T)，以及 $\sum$ 的左上r×n區塊決定($\sum_r$)，其重組有重大的意義。重組後的矩陣 A' 總共有 m×n 個元素，係由三個子矩陣 U'、V'、$\sum'$ 所組成，其中 U' 有 m×r 個元素，$V^{T'}$ 有 r×r 個元素，$\sum'$ 則只需儲存 $\sum$ 主對角的r個非零元素，當r遠小於m和n時，利用矩陣的 SVD 可以大幅減少儲存量。

利用奇異值分解，並將奇異值做排序，其中奇異值(δ)越大的，表示該數值越重要，因此可以透過奇異值分解，找出前 r 個奇異值所對應的句子，這些句子即文章摘要句。

2.5 結合關聯度之潛在語意分析

我們提出結合潛在語意分析與文章詞彙關聯度及文章句子關聯度的概念，建構出關聯性矩陣，用來加強與擷取文件中的概念結構。由前 2.2 與 2.3 小節所得之關鍵詞對應語句的權重關係矩陣、關鍵詞間關聯度矩陣以及語句間關聯度矩陣，以矩陣乘法的方式合成此三個矩陣，建構出新個關聯性矩陣，並以潛在語意分析得到語意層面的分析，進而篩選最適合之句子，作為文件摘要的依據。

利用前述之Jaccard coefficient關聯度計算以及TFSF詞彙權重計算方法，我們可以分別計算出關鍵詞間以及語句間的相似度矩陣，以及關鍵詞對應語句關係的權重矩陣。在此我們提出一個方法，藉由矩陣乘法將此三個關係矩陣合成出一個新的關鍵詞對應語句關係矩陣，再將此新的關係矩陣以潛在語意分析為基礎節錄出適當摘要句。步驟如下：

a. 詞彙權重計算(使用TFSF計算關鍵詞位於語句中的權重值)。
b. 關聯度計算(使用Jaccard coefficient分別計算詞彙間以及語句間關聯度)。
c. 合成運算(將詞彙對應語句權重矩陣、詞彙關係矩陣及語句關係矩陣合成)。
d. 合成關係矩陣(經由前一步驟得到的新關係矩陣)。
e. 計算交互乘積矩陣。

f. 從 AA^T 矩陣求出 U、$\sum$、V。
g. 排序奇異值矩陣 $\sum$，挑選適當的前r個奇異值 δ。
h. 前r個 δ 奇異值 δ 所對應的句子 S。
i. 文章摘要候選句。

最後，根據每個文章摘要候選句在原始文章出現之先後順序依序排列，則此精簡語句所構成之簡短文章即文章摘要。

3 實驗結果

實驗裡，我們從台灣 Yahoo!奇摩網路新聞蒐集 13 大類(生活、地方、社會、政治、科技、旅遊、財經、健康、國際、教育、運動、戲劇、藝文)每一大類共 100 篇文章，合計 13 大類共 1300 篇台灣 Yahoo!奇摩網路新聞，並將本研究所提之方法應用在中文文件摘要的實驗上。

3.1 評估方法

評估摘要是一件困難的工作，很難達到所謂的客觀。對於自動摘要的評估，學者一般多從系統研發成本與成果效益雙方面進行評估分析。在成果效益上，多半針對自動摘要的可讀性，要求使用者或是領域專家提供意見，此方法也比較主觀。Salton〔6〕所進行的自動摘要評估方式，認為使用者的反應也都是評估的重要指標之一。以使用者或是領域專家進行評估，無法避免人為主觀因素於其中，評估的結果會依使用者背景及需求的不同，而有所不同，故結果未必正確。然而，至今似乎仍無法有一個較能正確且客觀的自動摘要評估法。在此，我們提出二個指標評估文件摘要。指標 1，「摘要句平均相似度」：由於語句間相似度越高，表示其之間所牽扯的主題與內容越相近，重複性也越高，所以我們希望彼此間相似度越低越好，表示摘要句間的重複性越低。指標 2，「文件分類正確率」：當摘要所包含的類別關鍵詞越多且屬該文章之類別，則表示此文件自動摘要方法能保留原文之文章關鍵詞，進而保留原文之主要內容。指標 2 所採用的評估方法，是在資訊檢索領域中常見的衡量指標之一-*Precision*。*Precision* 表示自動摘要能正確分到類別之準確度。如方程(6)所示：

$$precision = \frac{|A|}{|B|} \qquad (6)$$

其中 A 為辨識為正確類別之文件數，B 為測試文件。

3.2 實驗參數設定

由於本論文主要是以奇異值來決定自動摘要之語句數，所以我們提出以壓縮率(Compress Ratio)的方式挑選奇異值：設定固定壓縮率，保留固定壓縮率的奇異值及其所對應的文章語句。奇異值挑選方法之參數設定如下表一所示：

表一 奇異值挑選之設定

奇異值挑選之設定							
壓縮率	20%	30%	40%	50%	60%	70%	80%

由於指標二為文件分類正確率，所以我們定義了 12 種類別關鍵詞庫，以比較在不同數量的類別關鍵詞庫下，對於自動摘要方法的影響。類別關鍵詞主要由文件集中隨機挑選 650 篇文件當訓練文件並以 2.2 小節所介紹的 TF-IDF 關鍵詞權重計算法，計算並挑選出類別關鍵詞；並以剩下的 650 篇文件當做測試文件，在此我們只比較代表性的實驗結果，即所有實驗的前中後(每類別關鍵詞數為 5、每類別關鍵詞數為 35 以及每類別關鍵詞數為 60)如表二所示。

表二 類別關鍵詞庫詞數設定

每類別關鍵詞數	類別數	類別關鍵詞庫關鍵詞總數	備註
5	13	65	前
35	13	455	中
60	13	780	後

3.3 實驗結果分析

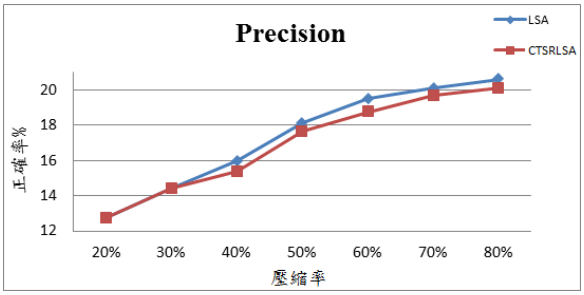
本實驗首先將比較類別關鍵詞數各為5個、35個以及60個的情況下，我們所提之CTSRLSA與LSA在文件分類正確率的比較；其次比較分別以測試集文件650篇所產生之摘要的摘要句間平均相似度情況，最後並說明實驗的結果。文件分類正確率的計算方法，主要是以摘要所包含的類別關鍵詞數為依據，當包含某類別關鍵詞越多，則將該摘要分到該某類別中。最後再計算有多少篇摘要是有被分到其該所屬之類別，如方程(6)。摘要句間的平均相似度計算方法，主要是所節錄出之摘要句藉由方程(3)所計算之語句間相似度，統計摘要句彼此間相似度之值並求得相似度總和，最後再將相似度總和除以摘要句數以求得摘要句間平均相似度之值。如方程(7)所示。

$$Avgsim = \frac{\sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} sim_{ij}}{N} \tag{7}$$

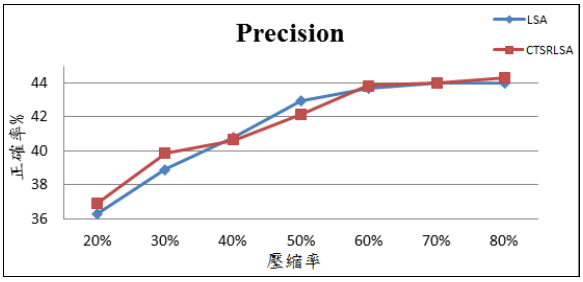
其中 sim_{ij} 為某篇文章中，語句 i 與語句 j 之相似度；N 為某篇文章之語句總數；n 為某篇文章之摘要句數。

3.3.1 文件分類正確率之比較

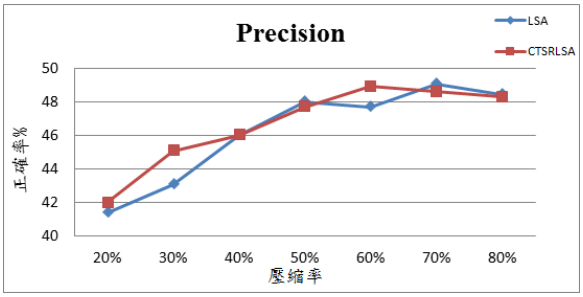
圖三至圖五的曲線可看出我們所提CTSRLSA方法在低壓縮率的情況下(20%~40%)，正確率與LSA一樣甚至更高；由表四可得知在類別關鍵詞數各為35個且在壓縮率為30%的情況下，CTSRLSA的正確率比LSA高出0.92%(約6篇文件)；當類別關鍵詞數各為60個(圖五與表五)的時候在壓縮率為30%的情況下，可以看得出來CTSRLSA的正確率比LSA高出2%(約12篇文件)，由以上結果得知我們所提之方法能在原文句保留較低的情況下，仍能有效保留原始文章的主題。



圖三 類別關鍵詞數各5個時，文件分類正確率比較



圖四 類別關鍵詞數各35個時，文件分類正確率比較。



圖五 類別關鍵詞數各60個時，文件分類正確率比較。

表三 類別關鍵詞庫詞數各5個時，文件分類正確率。

方法\壓縮率	20%	30%	40%	50%	60%	70%	80%
LSA	12.76	14.46	**16.00**	**18.15**	**19.53**	**20.15**	**20.61**
CTSRLSA	12.76	14.46	15.38	17.69	18.76	19.69	20.15

表四 類別關鍵詞庫詞數各35個時，文件分類正確率。

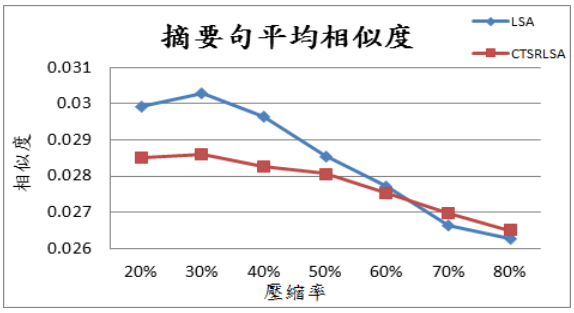
方法\壓縮率	20%	30%	40%	50%	60%	70%	80%
LSA	36.30	38.92	40.76	**42.92**	43.69	44.00	44.00
CTSRLSA	**36.92**	**39.84**	40.61	42.15	**43.84**	44.00	44.30

表五 類別關鍵詞庫詞數各 60 個時，文件分類正確率。

方法\壓縮率	20%	30%	40%	50%	60%	70%	80%
LSA	41.38	43.07	46.00	**48.00**	47.69	**49.07**	48.46
MMLSA	**42.00**	**45.07**	46.00	47.69	**48.92**	48.61	48.30

3.3.2 摘要句間平均相似度之比較

由圖六的曲線趨勢圖可以看出我們所提之 CTSRLSA 的方法在 650 篇文件中，整體的平均相似度都比 LSA 來的低，由表六可看出在 650 篇文件中 CTSRLSA 能在較低壓縮率(文章語句保留較少)之下所節錄出的摘要，能有效地降低摘要句間的平均相似度。由此反映出我們所提之 CTSRLSA 方法能有效地降低摘要句間的相似性即所謂的文句重複性。



圖六 650篇文件摘要句平均相似度比較圖。

表六 650 篇文件摘要句平均相似度。

方法\壓縮率	20%	30%	40%	50%	60%	70%	80%
LSA	0.029	0.030	0.029	0.028	0.027	0.026	0.026
MMLSA	**0.028**	**0.028**	**0.028**	0.028	0.027	0.026	0.026

4. 結論

由實驗結果得知我們所提之方法 CTSRLSA 在低壓縮率(20%~40%)的情況下文件分類正確率都比 LSA 高，最好可以高出 2%(約 12 篇文件)，且摘要句間平均相似度也都比 LSA 低。所以我們提之方法能有效地找出原文內相似度較低(重複性較低)的語句，但又不失其原文意與我們預期的一樣。在未來展望中，將對此客觀評估方法做進一步的研究，試著將一些演算法套入，並試著找出更加適合評估摘要句語意即可讀性的評估方法。並在未來對增強型語意分析為基礎的文件摘要上，進一步的修改其演算法或是結合其他演算法以致達到更加擬人的文件摘要成效。

致謝

感謝國科會計畫補助，計畫編號：NSC 99-2221-E-216-021。

參考文獻

[1] 曾元顯，(1997)，「關鍵詞自動擷取技術之探討」，*中國圖書館學會會訊*，5 卷，3 期(106)。

[2] D. Marcu, (1997). “From discourse structures to text summaries”, In Proc. *ACL Workshop Intell. Scal. Text Summar.*, pp. 82-88.

[3] D. McDonald, H. C. Chen, (2002). “Using sentence-selection heuristics to rank text segment in TXTRACTOR”, in Proc. *second ACM/IEEE-CS joint conf. Digital libraries*, Portland, Oregon, USA, pp. 28-35.

[4] E. F. Skorochod'ko, (1972).” Adaptive method of automatic abstracting and indexing”, in *Proc. IFIP Cong. 71*, Amsterdam, North-Holland Publishing Company.

[5] G. Salton, (1988). “Automatic Text Processing”, *Addison-Wesley Longman Publishing Co., Inc.* Boston, MA, USA.

[6] G. Salton, et. al., (1997). “Automatic Text Structuring and Summarization”, *Inform. Proc. & Manag.*, Vol. 33, No. 2, pp. 193-207.

[7] H. P. Luhn, (1958). “The Automatic creation of literature abstracts”, *IBM Journ. Res. and Devel.*, Vol. 2, No. 2, pp. 159-165.

[8] J. Kupiec, J. Pedersen, F. Chen, (1995). “A Trainable Document Summarizer”, *In SIGIR*, ACM, Seattle WA, USA.

[9] R. Barzilay, M. Elhadad, (1997). “Using Lexical Chains for Text Summarization”, in Proc. Workshop on *Intelligent Scalable Text Summarization*, Madrid, Spain, Augus.

[10] S. Deerwester, et. al., (1990). “Indexing by latent semantic analysis”, *Journ. Americ. Soc. Inform. Science*, Vol. 41, No. 6, pp. 391-407.

計畫成果自評：

1. 研究內容與原計畫相符程度

自評：如實驗所示，我們提出的摘要方法比原本之 LSA 法，在文件內容較高壓縮率情況下，亦能有較高的文件分類正確性，以及句子的代表性。

2. 達成預期目標情況

計畫書中，對於自動摘要欲完成之目標：

1. 計算欲摘要的文章詞與詞之間的關聯強度。。
2. 應用Latent Semantic Analysis計算詞與句子之間的關聯度。
3. 設計指標函數評估句子的重要程度。
4. 結合1.，2.及3.的結果，計算每個句子的重要程度。
5. 篩選句子完成文件摘要。

自評：如結案報告所示，章節2.3為詞間關聯度計算，章節2.4及2.5應用及改良潛在語意分析(Latent Semantic Analysis, LSA)，章節2.5同時以上述原理完成句子選取之摘要動作。

3. 研究成果之學術或應用價值

自評：對於專業資料庫的建立，以及提供數位內容服務十分有幫助。

4. 是否適合在學術期刊發表或申請專利

自評：本計劃應用了不少理論並改良先前的方法，部分內容已發表於 2010 交通大學舉辦之 AI 研討會，目前正撰寫成期刊投稿型式，準備投稿 SCI 論文。

5. 主要發現或其他有關價值

自評：摘句的主要難題在於三項步驟：句子重要性計算、句子相似度計算以及摘句的形成。這三項過程，目前並沒有一個標準作法。句子重要性計算通常以包含的關鍵詞來評估。句子相似度計算需加入文法觀念較能有明顯改良。摘句的形成目前則以選取重要句子的方式進行，而無法重新創造語意濃縮之後的句子，以做到真正的摘要。這些都是從事此研究所需面臨的挑戰。

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

99 年 10 月 23 日

附件三

報告人姓名	周智勳	服務機構及職稱	中華大學資訊工程系
時間 會議 地點	2010/10/15~2010/10/17 德國-達姆斯塔特	本會核定 補助文號	計劃編號： NSC99-2221-E-216-021
會議名稱	（中文） （英文）The Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2010)		
發表論文題目	（中文） （英文）Fast Forgery Detection with the Intrinsic Resampling Properties		

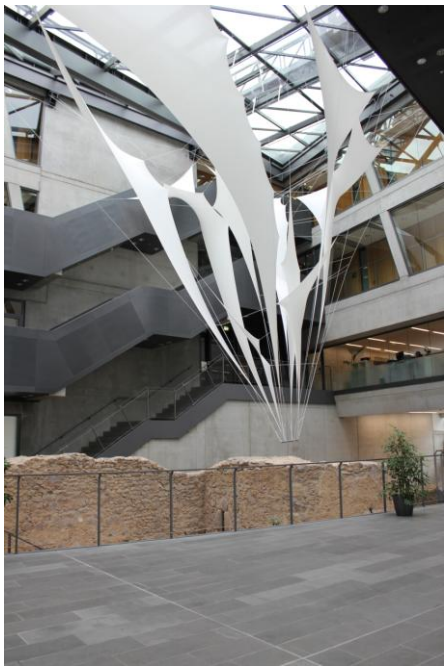
報告內容應包括下列各項：

一、參加會議經過

這是個跟智慧型計算及多媒體相關的會議，會議時間在 10/15~10/17，地點位於法蘭克福南方二、三十公里的小鎮。由於直航班機已客滿，改搭須於曼谷轉機的班機，行程上多花了些時間。

會議場地設於達姆斯塔特的一個會議中心，特殊的建築結構(圖一)，搭配室內藝術採光設計(圖二)，有種後現代建築風。國外的研討會不像國內一般，會有明顯的動線指標與工作人員引導，因此花了不少時間找會議地點。

圖一 會議現場外



圖二 會議現場內

圖三 會議電子看板

圖四 會議註冊處

此次研討會總計三天，由於是在德國舉辦，歐洲人士相對較多，不像在亞洲之研討會，台灣及大陸學者佔了一大半。我的論文排在第一天發表，由於同一個時段並行的 section 不多，因此每個 section 聆聽的人也相對較多。

會議晚宴安排在第二天晚上，地點在會議廳後方附屬餐廳，以自助式形式進行，由於與會人士不少，晚宴會場顯得有點擁擠。過程中除介紹工作人員，並有頒獎及下屆主辦單位的宣傳說明。

二、與會心得

1. 會場的佈置似乎不像國內辦研討會的熱鬧，國內辦研討會，會花不少心思在會場佈置上，感覺比較有那麼個氣氛。
2. 國內也常舉辦國際性研討會，可以於議程中安排半日遊，讓外國學者增加認識台灣的機會，或於晚宴時安排有代表性的表演，如此對推展觀光也許有一些幫助。
3. 近年來大陸方面參加研討會的學者漸多，國內方面出國留學的學生日漸減少，因此應該鼓勵國內的研究生，參加國際研討會。

三、考察參觀活動(無是項活動者省略)

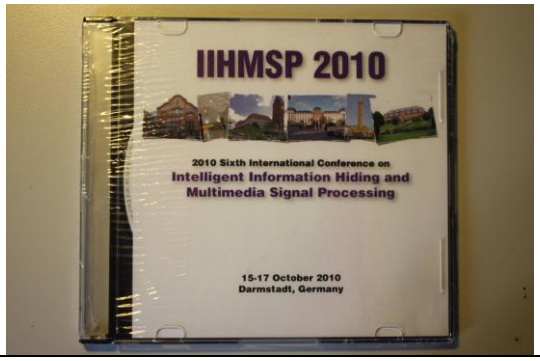
無。

四、建議

1. 一個研討會的晚宴，算是一個重要的流程，可藉以相互交流，並可由節目的安排，介紹一個國家的特色。此次晚宴場地顯得有點擁擠，過程亦有點吵雜，感覺只是填飽肚子之用，有點可惜。建議國內辦研討會時，多加注意。

五、攜回資料名稱及內容

Proceedings of the Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing.



六、其他

1. 研討會網站首頁，應可提供天氣訊息讓與會者參考，方便準備所攜衣物。
2. 國內航空公司直飛德國的班次不多，多花了好幾個小時於轉機過程。
3. 工作人員安排得不多，過程中遇到問題，比較不容易解決。
4. 感謝國科會工程處的補助。

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：周智勳	計畫編號：99-2221-E-216-021-
計畫名稱：研究與開發具有文件自動分類與摘要功能之網頁查詢系統	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	2	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	1	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	1	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 ■撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

部分成果已發表於 2010AI 研討會(交通大學舉辦)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

如實驗所示，我們提出的摘要方法比原本之 LSA 法，在文件內容較高壓縮率情況下，亦能有較高的文件分類正確性，以及句子的代表性。此研究對於專業資料庫的建立，以及提供數位內容服務十分有幫助。本計劃應用了不少理論並改良先前的方法，部分內容已發表於 2010 交通大學舉辦之 AI 研討會，目前正撰寫成期刊投稿型式，準備投稿 SCI 論文。然而摘句的主要難題在於三項步驟：句子重要性計算、句子相似度計算以及摘句的形成。這三項過程，目前並沒有一個標準作法。句子重要性計算通常以包含的關鍵詞來評估。句子相似度計算需加入文法觀念較能有明顯改良。摘句的形成目前則以選取重要句子的方式進行，而無法重新創造語意濃縮之後的句子，以做到真正的摘要。這些都是從事此研究所需面臨的挑戰。