International Journal of Science and Engineering
http://140.133.4.61/IJSE/index.html



# 以詞彙關聯為基礎的主題導向式中文文件摘要

Shu-Wei Chang 張書瑋
國立東華大學資訊管理研究所

**摘要**

自動化中文文件摘要技術是資訊探索中重要的輔助工具。本研究藉著人工摘要範例的觀察，發現在這些真實範例中存在相同的摘要特徵，亦即摘要中皆包含：(1)許多與主題相關的主題字；(2)許多彼此存在語意關係或詞彙組合的詞彙。因此，本論文應用網路結構圖形化的方式與網路向心性分析演算法，並以主題字與詞彙關聯兩項摘要特徵為重點來設計新的自動摘要技術方法。

**關鍵詞：**自動化中文文件摘要技術、主題字、詞彙關聯、網路結構圖形化的方式、網路向心性分析演算法

## 壹、前言

自動化摘要技術，簡單來說即是電腦對資訊進行精簡並萃取重要部分的處理過程，種類包含文件摘要(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；混合式摘要(Hybrid Summarization)：原始資料綜合文字與多媒體。本研究的範疇屬於文件摘要，定義如下 [1]：

> 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).
>
> 針對特定使用者或特定目標，將原始資料內容進行濃縮精簡的過程，目標乃萃取重要內容並以精簡版呈現。

文件摘要處理流程，包括：分析(Analysis)、轉換(Transformation)與合成(Synthesis)等步驟，如圖 1 所示：

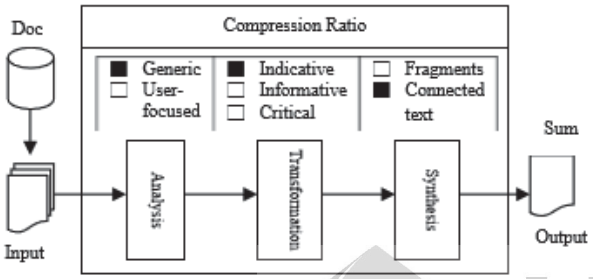



圖 1：文件摘要技術流程[1]

- 輸入(Input)/輸出(Output)：輸入文件數量為單一文件或多文件，輸出為一份濃縮精簡的摘要。
- 文件分析(Analysis)：分析文章說明原文字詞意涵。
- 內容轉換(Transformation)：將分析完的結果轉換成系統的表示方式。
- 內容合成(Synthesis)：最後利用轉換好的表示法進行摘要評估，依照重要性挑選出摘要內容，挑選完畢後再合成輸出最後摘要結果。

這些過程中有些因素必須考慮，如使用者的需求、摘要內容的形式、摘要的連貫性與可讀性或文件與摘要間的壓縮比(Compression Ratio)等等，都會直接或間接影響系統所產生摘要結果的好壞[1]。有鑑於文件摘要考慮因素很多，導致產生出來的文件摘要技術有所不同，以下針對文件摘要技術分成不同形式來作介紹，舉凡從輸入對象、摘要的目的、到輸出的形式。

本研究討論以中文為主的文件摘要方法，由於中文相較於英文在寫作文法上較為複雜，中文的字跟字之間沒有空格，所以根據斷詞的位置不同，產生出來的詞彙涵義也有所不同。另外，語言處理上討論的同義詞(Synonymy)、一詞多義(Polysemy)、片語(Phrase)等問題，雖然在英文中也有同樣問題，但由於斷詞的複雜性與產生的多變結果不同，其相關議題也是本論文需要面對的問題之一。

本論文藉由觀察網路新聞文章的內容架構，以及針對現存手寫摘要做內容分析，從中找出設計自動化摘要的方法，分析出以下特徵：

(1) 摘要中含許多與主題相關的詞彙—即主題字，而主題字能夠引導出這個文章的主題。

(2) 主題詞彙彼此間有語意上的關聯存在—觀察中發現，兩兩詞彙間存在著一些關聯，可以是詞彙的組合關係，也可以是語意上相似的關係。

依據前述的兩項特徵，本研究提出符合指示性(Indicative)、單文件(Single Document)、單語系(Monolingual)、一般性(Generic)以及萃取式(Extract)的中文文件摘要方法，並期望達到以下兩個目的：(1) 內容精簡，濃縮原文；(2) 內容符合讀者期待，方便閱讀。

## 貳、相關研究

由[10][11]文獻整理中知道自動化文件摘要技術最早起源於 1950 年，由於過去硬體設備的不足，電腦所能計算處理的資料有限，因此必須依賴固定的寫作形式及文章架構分析，例如：文章內的分佈位置(Position in text)、詞

* Corresponding author:610239009@ems.ndhu.edu.tw
DOI：10.6159/IJSE.2016.(6-1).03

彙關聯提示(Lexical cues)。以論文為例，論文文章有固定的寫作格式，從開始的前言至最後結論，所以就能利用這格式判斷語句內容是否可能為摘要內容。

接著來到 1970 年，由於人工智慧發展，開始使用框架(Frames)或模板(Templates)方式做摘要，利用事先定義好的模板或框架輔助來擷取重要部分，例如：將人事時地物等內容萃取出來，再分析其涵義，舉例來說，此方法可以定義好事件語句的架構為名詞+動詞+受詞，藉此萃取文章事件中的人物，但這種方法最大的缺點在於這些模板和框架必須事先由專家事前進行，且必須嚴格且詳細定義，否則若這些模板框架定義廣泛度不夠，摘要出來的結果就不具正確性，也導致結果可能扭曲錯誤。

從 1990 年開始，資訊擷取(Information Retrieval, IR)技術開始大量運用在摘要技術上面。資訊檢索利用關鍵字在資料庫中找尋相關的資料，同理移轉到文件摘要中，以單文件摘要來說，就是找尋與其主題相關的語句，並重組產生摘要結果。資訊擷取技術的缺點是著重在字層級上的分析，未考慮到同義詞(Synonymy)、一詞多義(Polysemy)、詞彙依屬關係(Term Dependency)以及片語(Phrase)的語意層級的分析，導致摘要內容可能不適當。

除了上面三種比較重要的方法類型外，期間也包含許多其他的摘要技術，像語言學或認知心理學的方法。目前最常見的方法像是利用語句特徵摘要[7] [12][13]，以及用機器學習法(Machine Learning)的摘要方法[3][4]，或是結合網路結構模型的摘要方式[5][9][15]。

一、 以語句特徵為主的摘要技術

以語句特徵為主的摘要技術[7][12][13]，主要是分析語句的特徵，藉著這些特徵評估文章內語句的代表性，接著進行摘要語句挑選。舉[8]來說，其所使用的語句特徵包含如下：(1) 語句長度(Sentence Length)、(2) 提示片語(Fixed-phrase)、(3) 段落位置(Paragraph)、(4) 主題字詞(Thematic Words)、(5) 大寫字詞(Uppercase Words)。

此篇文獻雖然只利用此五種常見特徵作為挑選依據，但經過實驗後，該文獻認為並不是全部特徵都考慮就會有較好結果，其實驗結果也說明只考慮上面前三點的效果較為出色。

二、 結合網路結構模型分析的摘要技術

討論到網路結構分析演算法，其想法從近年來知名的搜尋引擎 Google 而來，利用網路結構演算法 PageRank Algorithm[14]將網頁進行重要性分析，其核心在於利用網頁引用程度，將網頁進行重要性的重新定義。應用到文件摘要，過去研究也有許多利用這種概念演算法進行文件摘要方法的技術，例如：[2][5]。

LexRank [15]使用 PageRank 的方法作為多文件摘要的核心模型，賦予語句重要性進而能夠分辨摘要語句與否。以下公式中 *PageRank* 皆表示演算法運算函式，代入參數為頁面 $p$，定義如 Eq.(1)：

$$PageRank(p_i)=\frac{1-d}{N}+d\sum_{p_j\in M(p_i)}\frac{PageRank(p_j)}{L(p_j)} \quad (1)$$

$p_1,p_2,...p_n$ 是被引用有關聯的頁面，$d$ 為阻尼係數（damping factor）設定為 0.85，$M(p_i)$是 $p_i$ 頁面鏈結的集合，$L(p_j)$是 $p_i$鏈出頁面的數量，而 $N$ 是所有頁面的數量，同樣套用到文件摘要上面，根據分析對象不同，可以將頁面 $p$ 替換成語句 $s$，如 Eq.(2)：

$$PageRank(s_i)=\frac{1-d}{N}+d\sum_{s_j\in M(s_i)}\frac{PageRank(s_j)}{L(s_j)} \quad (2)$$

Eq.(2)同於 Eq.(1)，分析對象變成語句 $s_i$、$s_j$，$M(t_i)$表示與 $s_i$ 鏈結的語句集合，$L(s_j)$表示 $s_i$ 鏈出的語句數量，$N$ 為總句數，$d$ 為阻尼係數。

三、 利用詞彙關聯性的摘要選取方法

文獻中也有討論到有關詞彙關聯的摘要方法，這些方法利用詞彙關聯作為挑選摘要語句的依據。例如：[16]提出一個對於選取摘要句的新觀點。他們認為整個文件是一個全集合，其中一定包含摘要這個子集，但要如何挑選出這個子集合是該文獻所要討論的議題。過去很多方法都以語句為主體，對語句評分後排序，接著挑選語句。但此篇文章是以更細的觀點來看摘要，此篇認為摘要應以字詞為考慮對象而不是句子，定義如下 Eq.(3)：

$$w(t_i,t_j)=2\sum_{a\in\{"ij","i-j","-ij","-i-j"\}}n_a\left(\log p_a^D-\log p_a^{B+D-ind}\right) \quad (3)$$

$t_i,t_j$ 分別為兩個字詞，$P$ 為考慮在 $B$(*Corpus*)與 $D$(*Document*)內兩字的出現機率，$a$ 表示兩字的集合組合 $a\in\{ij,i-j,-ij,-i-j\}$，$n$ 表示詞彙出現數量，即透過詞彙共現規則(Co-occurrence)的概念導出詞彙的關係式，摘要( $Sum'$ )內的兩字詞權重和訂為摘要挑選的方法，以下 Eq.(4)為摘要集合挑選的方法定義：

$$w(Sum')=\sum_{\{t_i,t_j\}s.t.t_i\neq t_j\wedge\exists s\in S:\{t_i,t_j\}\subset s}w(t_i,t_j) \quad (4)$$

此篇利用「摘要內詞彙關聯性要最大」的評估方法進行挑選，利用此特徵對文件內的字詞關聯性挑選，以此挑選出語句組合，如下 Eq.(5)：

$$S^*=\arg\max_{|Sum|\le L}w(Sum') \quad (5)$$

$S^*$為摘要候選，Sum'為摘要的篇幅，$L$ 為壓縮比設定下的篇幅。

## 參、以詞彙關聯為基礎的主題導向式文件摘要

# 以詞彙關聯為基礎的主題導向式中文文件摘要

本章介紹本研究所提出的文件摘要，整個系統功能模組如圖 2 說明。單文件進入系統後，共透過四個主要模組進行摘要處理，其中包含：(1) 前置處理(Preprocessing)：對文章分析即語言分析處理；(2) 詞彙分析(Term Analysis)：此處進行詞彙權重計算，利用網路結構化及分析演算法進行運算；(3)語句篩選(Sentence Selection)：依據觀察到的摘要規則，以主題字與詞彙關聯為特徵評估語句集合重要性；(4)摘要生成(Summary Generation)：將挑選完的摘要結果根據壓縮比設定篇幅和摘要優化處理，生成符合使用者期待的摘要型式。

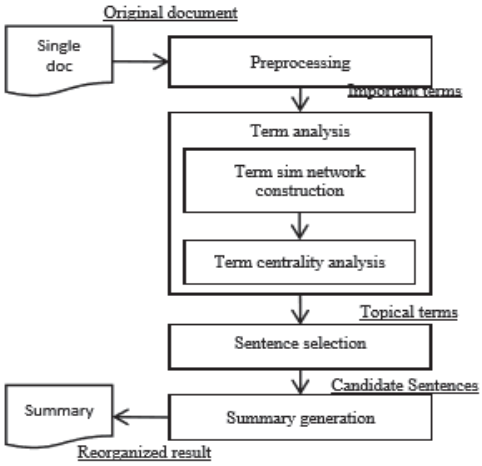



圖 2：完整系統架構

(一)前置處理

一個文件進入系統後，首先必須經過一連串分析，將文章內容拆解分析，動作包含分段、斷句、斷詞、標註詞性、過濾贅詞和停止詞(stop-word)，接著將有用的部分保留下來。一般來說，名詞與動詞的重要性要比冠詞、介系詞或感嘆詞等來得重要，且大多數語句都是“主詞-述詞-受詞”的結構，這裡的主詞及受詞常以名詞出現，而述詞也同樣以動詞表示[6]。因此，在前置處理步驟中，本研究就會保留如名詞或動詞這類較為重要的詞彙以利後續進行摘要分析。另外因為是中文系統，此處借助中研院中文斷詞系統 (http://ckipsvr.iis.sinica.edu.tw/)進行斷詞切字步驟。

圖 3 的斷詞結果範例，分別顯示成“詞彙(詞性)”的型式。如此一來就能根據表 1 的詞性進行詞彙篩選，過濾掉標點符號，並保留名詞及動詞作為摘要處理過程的關鍵詞候選。

```
東華(Nb)  大學(Nc)  資訊(Na)  管理(VC)  碩士(Na)  論文(Na)  ，(COMMACATEGORY)
------------------------------------------------------------
中研院(Nc)  中文(Na)  斷詞(VA)  系統(Na)  範例(Na)  測試(Na)
------------------------------------------------------------
```

圖 3：斷詞結果及標註詞性

表 1：CKIP 中研院中文斷詞系統定義之部分重要標記

| 符號 | 意義 |
|---|---|
| Na | 普通名詞 |
| Nb | 專有名詞 |
| Nc | 地方詞 |
| VA | 動作不及物動詞 |
| VC | 動作及物動詞 |

(二)詞彙分析

資訊檢索中最常使用距離來求得兩點之間的距離作為評量兩者相似度的關係標準。本研究採用 NGD (Normalized Google distance)[17] ，計算詞彙間的距離關係，並將其結果供後續摘要處理步驟使用，如下 Eq.(6)：

$$NGD(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log M-\min\{\log f(x),\log f(y)\}} \qquad (6)$$

$f(x)$, $f(y)$, $f(x, y)$分別表示兩個詞彙單獨與共同出現於網頁的結果。這裡的 $M$ 為一個假想的資料庫資料量，根據近幾年研究統計估計，Google 資料量至少超過 $10^{12}$，所以此處 $M$ 定義為 $10^{12}$。

(1) 建立詞彙關係網路模型

經過繁雜計算詞彙關係後，此處必須將這些詞彙統整，所以採用網路模型的方式如圖 4，之後才能更進一步進行關係分析。圖形為無向圖表示為 $G = (V, E)$，$V$ 是節點，每個節點代表一個詞彙 $t$，$E$ 是邊，若有連線則代表彼此存在詞彙關係，本研究採用 $NGD$ 來計算詞彙關係。

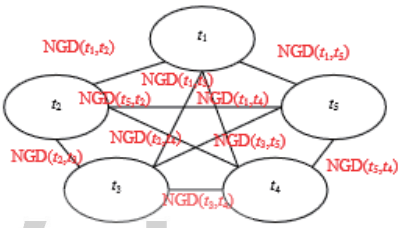



圖 4：詞彙關係網路圖型

(2) 網路模型的向心性分析

完成詞彙網路模型後，套入網路結構分析演算法進行更進一步結果分析。此處參考 Google 採用的網頁向心性

分析演算法 PageRank，詞彙等同於網頁為節點，詞彙關係等同於網頁引用程度為邊，運算後將權重重新更謹慎的分配一次，最後產生新的字詞權重。

圖 5 為向心性分析前的網路圖型範例，由於一開始權重彼此間未定義，所以每個詞彙初始權重皆相同，此處設定為 1，透過 Eq.(7)，如下：

$$PR(t_i)=\frac{1-d}{N}+d\sum_{t_j\in M(t_i)}\frac{PR(t_j)}{L(t_j)} \qquad (7)$$

$PR$ 表示經 Pagerank 計算後的詞彙權重，$t_i$、$t_j$ 表示詞彙，$d$ 為阻尼係數設為 0.85，$N$ 表示所有節點數即總字詞數，$M(t_i)$表示所有與 $t_i$ 鏈結的節點集合，$L(t_j)$表示所有 $t_i$ 鏈出的數量，反覆計算，直到當全部節點的變動量達到平穩，也就是收斂狀態。停止的判斷方式如 Eq.(8)所示：

$$\sum\left|\left(PR'(t_i)-PR(t_i)\right)\right|\le\varepsilon \qquad (8)$$

正常只要未達條件就會繼續運算，直到滿足上列公式，即代表狀態平穩，此處 $PR'(t_i)$代表 $t_i$ 經 PageRank 演算法運算一遍後的結果，$PR(t_i)$則是未運算前的結果，$\varepsilon$ 表示一個極小到可以忽略的值，本研究設定為 0.0001。

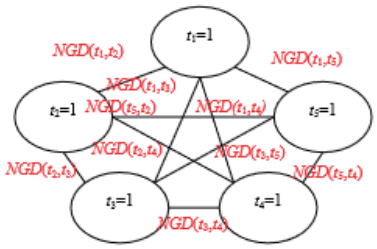



圖 5：分析前網路圖

經過 PageRank 重新分配權重以後，$t_1$~$t_5$ 的權重重新的調整分配，更謹慎地計算出各個詞彙彼此的重要程度，字詞間權重不再一樣，$t_1$=0.4 變為最大，$t_4$,$t_5$ 相等同為最小 0.1，如下圖 6：

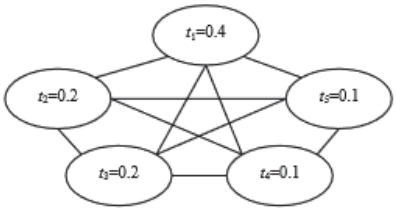



圖 6：向心性分析後的結果

透過網路模型分析後，產生新的詞彙權重結果。本論文利用這些新結果進行篩選主題字的處理。將算好的詞彙權重對應給予文章內詞彙一個權重值，並且設定主題字權重門檻藉此篩選出主題字，透過主題字的運用，就能找出更多對摘要有幫助的詞彙。

(三)語句篩選

此節說明摘要系統挑選摘要語句的方法。圖 7 為 Sentence Selection 演算法，其中 $D$ 表示文件，$s$ 代表只含主題字的語句，$n$ 代表文件總句數，$Sum'$代表摘要，$|Sum'|$代表摘要句數量，其初始為空集合 $\varphi$，並且最大限度為 $n \times CR$，$CR$ 代表壓縮比。首先輸入一篇已經完成分析處理的文件，文件內只包含主題詞，過程可視為兩個集合 $D$ 與 $Sum'$，將所有非摘要集合中的句子$s \in D-Sum$逐一加入到摘要集合內，並對摘要進行評估 $SCORE(Sum')$，並且挑選其中最高結果 $arg\max SCORE(Sum')$，將配對的語句選入 $Sum'$內，接著反覆挑選動作直到摘要集合內的語句數量超過所設定壓縮比下的摘要語句數$|Sum'|$ > ceiling($n \times CR$)，挑選動作隨即停止，並將目前 $Sum'$內挑選的句子送往下一個步驟進行摘要句生成處理。

**Input:**
$D = \{s_1, s_2 \ldots s_n\}$: a document without nontopical terms
$CR$: compression ratio
**Output:**
$Sum$: a summary candidate where $|Sum'| \le$ ceiling($n \times CR$)
**Procedure:**
**STEP 1:** Initialize $Sum' = \varphi$
**STEP 2:** Pick up a sentences, $s \in D - Sum'$, such that
$arg\max SCORE(Sum')$ where $Sum' = Sum \cup \{s\}$
**STEP 3:** Check $|Sum'|$
if $|Sum'| >$ ceiling($n \times CR$) end
else add $s$ into $Sum'$ and go to STEP 2

圖 7：Sentence Selection 演算法

$SCORE$ 為此研究設計方法，以下方法圍繞著本研究最初的動機進行方法設計。根據研究動機的觀察，摘要特徵有：(a) 主題字較多；(b) 摘要內彼此詞彙關聯緊密。本論文設計三種評估(*SCORE)*的方法，方法中 $PR$ 代表透過 PageRank 演算法計算後的值，$NGD$ 代表透過 Normalized Google distance 公式計算兩點關係的值，方法分別說明如下：

(1) 考慮語句中主題字數量

方法一探討主題字對於摘要語句的影響，但不考慮主題字權重，採考慮語句內所含主題字個數，討論主題字的多寡是否影響摘要語句挑選，如 Eq.(9)：

$$SCORE(Sum')=Number\ of\ terms(Sum') \qquad (9)$$

(2) 考慮語句中主題字關聯性

方法二是從文獻[16]提出的關聯摘要的衍生方法，由於摘要內的內容應該是緊密有關聯性，藉著語句內的詞彙關係來挑選摘要語句，可得到詞彙關係最緊密的摘要結果。此處每組配對只算一次，如 Eq.(10)：

$$SCORE(Sum')=\sum_{\{t_i,t_j\mid t_i\neq t_j\,\&\,t_i,t_j\in Sum'\}}NGD(t_i,t_j) \qquad (10)$$

(3) 同時考慮主題字與詞彙關聯性 (考慮主題字個數與詞

彙關係)

方法三為此篇論文研究提出的方法。透過實際摘要範例觀察出摘要特性，結合摘要富含主題字以及摘要主題字彼此存在關聯兩種特徵，提出此種摘要方法，本研究用線性方式來結合兩項因素，如 Eq. (11)：

$$SCORE(Sum') = Number\ of\ terms(Sum') + \sum_{\{t_i,t_j \mid t_i \neq t_j\ \&t_i,t_j \in Sum'\}} NGD(t_i,t_j) \quad (11)$$

(四)摘要生成

上述所產生的結果，可以用來進行文件摘要。本論文依照評估摘要方法依序挑選語句至摘要內，直到摘要達到 $k$ 個語句。參數 $k$ 一般是依照使用者所設摘要比例篇幅而訂，即壓縮比(Compression Ratio)。本論文研究討論壓縮比共分為 10%、20%、30%三種型式。

最後的結果還要進行摘要重組(Sumary Reorganization)，顧名思義即將摘要結果重組排序達到符合讀者期待。由於系統挑選後的摘要沒有按照原始文件中的語句順序排列，導致系統所產生的結果並沒有達到內容的一致性與連貫性，所以本研究必須將系統摘要結果依照原始文件的語句順序，將其還原排序，藉此達到語意的連貫性，提高讀者對摘要結果的滿意度。

## 肆、實驗分析與評估

此章節驗證本研究第三章所提之單文件摘要方法的成效。首先介紹本研究實驗是如何驗證摘要的結果，接著描述相關實驗設計，其包含資料的來源、實驗的目的與實驗的結果評估方式，最後透過實驗數據分析系統的表現與適用性。

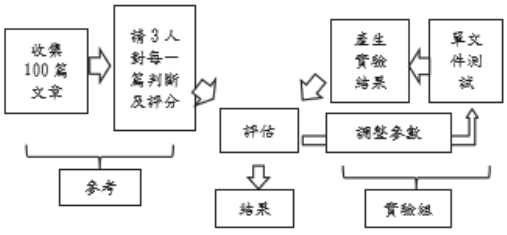



圖 8：實驗架構

一、摘要評估方法

本論文採以下方法來評估，作法如下：(1) 採每篇請三位讀者進行摘要撰寫的方法，利用多重觀點解決單一專家能力受限問題；(2) 評估順序是先判斷是否為摘要句，接著給予一個分數代表其重要性，分數為 10 分制，愈高代表越重要，藉此解決主觀以及偏誤問題。此外本論文研究根據判斷是否摘要句的方式將人工摘要分成嚴謹人工摘要以及寬鬆人工摘要，嚴謹人工摘要是指三人皆同意的語句才能視為摘要句；寬鬆人工摘要則是指至少兩人同意的才視為摘要句。將語句分類成摘要與非摘要語句後，利用專家給予的分數採平均後作為摘要語句的權重分數，再排序後挑選出 10%、20%、30%比例的摘要內容，作為摘要的標準解答。

嚴謹摘要語句計算方式如 Eq.(12)：

$$S_{sum} = \frac{(p_1 + p_2 + p_3)}{3} \quad (12)$$

$S_{sum}$ 表示摘要語句，$p_1$、$p_2$、$p_3$ 代表三位專家各自給予的分數。

寬鬆摘要語句計算方式如 Eq.(13)：

$$S_{sum} = \frac{\sum_{i=\{1,2,3\}} (p_i \mid S_{pi} \in S_{sum})}{N(p)} \quad (13)$$

$p_i$ 為第 $i$ 位專家，$S_{pi}$ 指的是由 $p_i$ 這位專家所評估的摘要語句，分子為判斷是摘要語句的分數和，分母 $N(p)$代表所有判斷是摘要語句的專家數量。

二、實驗設計

本論文測試資料是從教育部人權資源庫內收集文章如表 2，總共收集一百篇文章，主題為討論人權議題，時間從 2007 年開始至 2008 年。

表 2：來源資料介紹

| 類別 | 平均句子 | 平均字數 | 最多句子 | 最多字數 | 最少句數 | 最少字數 |
|---|---|---|---|---|---|---|
| 數量 | 36 句 | 1492 字 | 106 句 | 2375 字 | 17 句 | 1187 字 |

表 3：摘要結果說明

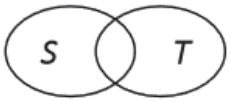
| 類別 | 平均句子 | 平均字數 | 最多句子 | 最多字數 | 最少句數 | 最少字數 |
|---|---|---|---|---|---|---|
| 數量 | 12 句 | 535 字 | 16 句 | 786 字 | 6 句 | 325 字 |

本研究設計三個實驗來驗證提出的方法，主要目的與作法如表 4：

表 4：實驗目的與作法

| 項目 / 類別 | 目的 | 作法 |
|---|---|---|
| 實驗一 | 評估主題字門檻的設定對摘要影響 | 設定不同門檻，分別為 100%(全選)、90%、80% |
| 實驗二 | 評估主題字與詞彙關聯兩種特徵的權重影響 | 設定 $\alpha$、$\beta$ 分別為主題字與詞彙關聯的比重，其 $\alpha+\beta=1$ |
| 實驗三 | 摘要方法根據不同壓縮比的適用性 | 所有實驗方法皆產生10%、20%、30%三種結果 |

本研究採取資訊擷取技術中常用的指標—精確度(Precision)、回傳率(Recall)以 F 檢定(F-measure)。Precision 代表對系統摘要的準確評估；Recall 代表對人工解答的正確回傳率；F-measure 則是代表兩者的調和平均數。本實驗評估的對象分別討論以句子為主以及以字詞為主的衡量指標。示意圖如圖 9：




*T*: manual summary of *D*, *S*: machine-generated summary of *D*

圖 9：集合圖示

圖中 $S$ 代表系統摘要結果，$T$ 代表人工摘要結果，中間交疊的部分即是兩集合相同的部分。計算方法如 Eq.(14)、(15)：

$$Precision = \frac{|S \cap T|}{|S|} \tag{14}$$

$$Recall = \frac{|S \cap T|}{|T|} \tag{15}$$

最後計算 F-measure 得以討論 Precision 與 Recall 調和平均值如 Eq.(16)：

$$F-measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{16}$$

三、 實驗結果

本研究討論壓縮比對於本研究所提之摘要方法的影響與適用度，下面將本研究所產生的所有結果作一表格進行整理。見表 5、表 6、表 7、表 8：

表 5：不同壓縮比下的寬鬆摘要比較(語句)

| $(\alpha,\beta)$ / 壓縮比 | (0,1) (單一詞彙關聯) | 在$(\alpha,\beta)$下的最佳結果 (主題字+詞彙關聯) | (1,0) (單一主題字) |
|---|---|---|---|
| 10% avg Recall | 0.2608 | 0.2862 (0.9,0.1) | 0.2920 |
| 10% avg Precision | 0.2732 | 0.2966 (0.9,0.1) | 0.3020 |
| 10% avg F-measure | 0.2528 | 0.2907 (0.9,0.1) | 0.2964 |
| 20% avg Recall | 0.3950 | 0.4033 (0.8,0.2) | 0.4004 |
| 20% avg Precision | 0.4105 | 0.4200(0.8,0.2) | 0.4167 |
| 20% avg F-measure | 0.4021 | 0.4109 (0.8,0.2) | 0.4078 |
| 30% avg Recall | 0.4493 | 0.4836 (0.6,0.4) | 0.4738 |
| 30% avg Precision | 0.4726 | 0.5031 (0.6,0.4) | 0.4931 |
| 30% avg F-measure | 0.4600 | 0.4925 (0.6,0.4) | 0.4826 |

表 6：不同壓縮比下的寬鬆摘要比較(字詞)

| $(\alpha,\beta)$ / 壓縮比 | (0,1) (單一詞彙關聯) | 在$(\alpha,\beta)$下的最佳結果 (主題字+詞彙關聯) | (1,0) (單一主題字) |
|---|---|---|---|
| 10% avg Recall | 0.4624 | 0.4926(0.9,0.1) | 0.4943 |
| 10% avg Precision | 0.5510 | 0.5767(0.9,0.1) | 0.5800 |
| 10% avg F-measure | 0.4953 | 0.5252(0.9,0.1) | 0.5256 |
| 20% avg Recall | 0.6298 | 0.6314(0.8,0.2) | 0.6287 |
| 20% avg Precision | 0.7100 | 0.7233(0.8,0.2) | 0.7231 |
| 20% avg F-measure | 0.6638 | 0.7233(0.8,0.2) | 0.6692 |
| 30% avg Recall | 0.6832 | 0.7043(0.6,0.4) | 0.6976 |
| 30% avg Precision | 0.7579 | 0.7861(0.6,0.4) | 0.7808 |
| 30% avg F-measure | 0.7160 | 0.7408(0.6,0.4) | 0.7346 |

表 7：不同壓縮比下的嚴謹摘要比較(語句)

| (α,β) / 壓縮比 | (0,1) (單一詞彙關聯) | 在(α,β)下的最佳結果 (主題字+詞彙關聯) | (1,0) (單一主題字) |
|---|---|---|---|
| 10% avg Recall | 0.2296 | 0.2645(0.9,0.1) | 0.2703 |
| 10% avg Precision | 0.2384 | 0.2716(0.9,0.1) | 0.2783 |
| 10% avg F-measure | 0.2329 | 0.2671(0.9,0.1) | 0.2733 |
| 20% avg Recall | 0.3374 | 0.3492(0.8,0.2) | 0.3476 |
| 20% avg Precision | 0.3895 | 0.4057(0.8,0.2) | 0.4028 |
| 20% avg F-measure | 0.3555 | 0.3688(0.8,0.2) | 0.3666 |
| 30% avg Recall | 0.3583 | 0.3723(0.6,0.4) | 0.3646 |
| 30% avg Precision | 0.4814 | 0.5032(0.6,0.4) | 0.4946 |
| 30% avg F-measure | 0.3995 | 0.4157(0.6,0.4) | 0.4073 |

表 8：不同壓縮比下的嚴謹摘要比較(字詞)

| (α,β) / 壓縮比 | (0,1) (單一詞彙關聯) | 在(α,β)下的最佳結果 (主題字+詞彙關聯) | (1,0) (單一主題字) |
|---|---|---|---|
| 10% avg Recall | 0.4329 | 0.4631(0.9,0.1) | 0.4654 |
| 10% avg Precision | 0.5131 | 0.5417(0.9,0.1) | 0.5461 |
| 10% avg F-measure | 0.4620 | 0.4934(0.9,0.1) | 0.4962 |
| 20% avg Recall | 0.5612 | 0.5663(0.8,0.2) | 0.5636 |
| 20% avg Precision | 0.6761 | 0.6740(0.8,0.2) | 0.6928 |
| 20% avg F-measure | 0.6053 | 0.6153(0.8,0.2) | 0.6135 |
| 30% avg Recall | 0.5822 | 0.5910(0.6,0.4) | 0.5853 |
| 30% avg Precision | 0.7487 | 0.7711(0.6,0.4) | 0.7673 |
| 30% avg F-measure | 0.6454 | 0.6593(0.6,0.4) | 0.6538 |

結果說明，本論文研究所提出的挑選摘要語句方法「結合主題字個數與詞彙關聯」較適用於篇幅較長的摘要形式，在壓縮比 20%以及 30%的結果中能夠說明詞彙關聯這項特徵需要有足夠篇幅才能比較具有影響力；相反壓縮比 10%中，篇幅較少的情況下，詞彙關聯影響程度甚小，比單純用主題字個數挑選的方法還要低。這與壓縮比的特性有著很大的關係，壓縮比越小代表需求越精準的語句，勢必就需要越重要的詞彙來幫助挑選語句；反之，壓縮比越大，內容包含越廣闊代表語句越容易內容散亂，需要詞彙關聯的幫助來抓住語句的主題，不過從各方面表現來看，本論文所提出的方法「結合主題字個數與詞彙關聯」確實都有較好的表現，不管任何壓縮比或特徵比例下皆優於其他方法。

## 伍、結論與未來研究方向

一、結論

本研究針對中文單語系文件，提出以產生指示性(Indicative)、單文件(Single Document)、一般性(Generic)以及萃取式(Extract)的自動化摘要方法。從觀察實際的摘要範例得到摘要的生成規則，包含：(1) 摘要中富含主題字；(2) 摘要內的詞彙彼此間存在關聯，藉此本研究提出一個結合主題字與詞彙關聯的摘要方法，期望摘要結果能夠有效提升至符合使用者期待。在主題字萃取的部分，本文應用網路結構模型與向心性演算法 PageRank，藉此篩選出符合文章主題的主題字詞。接著，利用摘要產生規則設計五種摘要句挑選方法，分別為：(1) 考慮主題字個數；(2) 考慮主題字間關聯性；(3) 考慮主題字個數與主題字間關聯性。透過以上方法系統再根據設定的摘要篇幅，排序候選摘要語句後，依序挑選並組合成摘要，最後再進行還原及潤飾語句即可生成最終摘要。

實驗結果發現，主題字的定義門檻定為八成較能有效篩選出正確的主題字，並且在方法三中，主題字個數與詞彙關聯性的方法在 20%以及 30%的壓縮比下有較佳的表現，最佳語句及字詞的平均 F 檢定分別有 49%和 74%的表現。因為有考慮到詞彙關聯，本研究認為考慮詞彙關聯主要的貢獻為拉近語句內的內容，所以在低壓縮比下較不能看出其表現效果。反倒是主題字個數這項特徵在低壓縮比下影響較大，由於主題字能夠反映文章的主題內容，主題字數量越多代表語句越接近摘要語句，剛好符合低壓縮比下摘要語句較少，挑選語句時必須考慮更精確。

二、未來研究方向

首先就本研究所提摘要規則與方法而言，本研究利用觀察到的摘要特性提出可行的摘要方法，藉著圖形化模組與演算法應用來達到有效篩選文章內重要字詞的方法，即主題字。主題字的應用使本論文能夠快速分析文章主題，藉著這些主題字詞產生一篇完整摘要。

摘要特性目前可能不只有研究所提的兩種特徵(1) 富含主題字；(2) 彼此詞彙存在關聯，未來研究方向可以以這個架構再納入更多特徵進行摘要的評估便是此研究的未來思考方向。其次，根據實驗結果發現，本研究所提的五種摘要方式皆有共同特性，優先挑選較長的語句作為摘要語句，其原因為長句含較多的內容資訊，所以較容易挑選到。但實際比對人工摘要發現，使用者挑選摘要雖然也是傾向挑選長句，但也有很大的機會挑選到短句，所以之後是否針對這個特性進行討論，這也是未來研究思考的重要方向之一。

## 參考文獻


[1] I. Mani, and M. Maybury, "Introduction", *Advances in Automated Text Summarization*, MIT Press, pp. x-xv.

[2] P. Goyal, L. behera, and T. M. McGinnity, "A Context-Based Word Indexing Model for Document Summarization", *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, 2013.

[3] N. Chatterjee, A. Mittal, and S. Goyal, "Single Document Extractive Text Summarization Using Genetic Algorithms", *Proc. Third International Conference on Emerging Applications of Information Technology (EAIT),* pp.19-23, Kolkata, 2012.

[4] Y. X. He, D. X. Liu, D. H. Ji, H. Yang, and C. Teng, "MSBGA: A Multi-Document Summarization System Based On Genetic Algorithm", *Proc. 5th International Conference on Machine Learning and Cybernetics,* pp. 2659 – 2664, Dalian, 2006.

[5] J.Y. Yeh, H. R. Ke, and W. P. Yang, "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network", *Expert Systems with Applications,* vol. 35, no. 3, pp. 1451-1462, 2008.

[6] 葉鎮源, "文件自動化摘要方法之研究及其在中文文件的應用", *碩士論文*, 國立交通大學資訊科學研究所, 2002。

[7] B. H. Hammo, H. Salem, and M. Evens, "A Hybrid Arabic Text Summarization Technique Based on Text Structure and Topic Identification", *International Journal of Computer Processing of Languages,* vol. 23, no. 1, pp. 39–65, 2011.

[8] J. Kupiec, J. Pedersen, and F. Chen, "A Trainable Document Summarizer", *Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 68-73, Seattle WA, USA, 1995.

[9] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization", *Information Processing & Management,* vol. 33, no. 2, pp. 193-207, 1997.

[10] S. Wang, W. Li, F. Wang, H. Deng, "A Survey on Automatic Summarization", *Proc. Information Technology and Applications,* pp. 193-196, 2010.

[11] J. Y. Yeh, H. R. Ke, W. P. Yang and I. H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis", *Information Processing & Management*, vol. 41, no. 1, pp. 75-95, 2005.

[12] X. Wan, and J. Zhang, "CTSUM: Extracting More Certain Summaries for News Articles", Proc. *37th International ACM SIGIR Conference on Research and Development In Information Retrieval,* pp.787-796, Gold Coast, QLD, Australia, 2014.

[13] E. D. Yirdaw, and D. Ejigu, "Topic-based Amharic Text Summarization with Probabilisic Latent Semantic Analysis", *Proc. International Conference on Management of Emergent Digital EcoSystems,* pp. 8-15, Addis Ababa, Ethiopia, 2012.

[14] L. Page, S. Brin, R. Motwani, T. Winograd, "*The PageRank Citation Ranking: Bringing Order to the Web*", Technical report, Stanford Digital Library Technologies Project, 1998.

[15] G. Erkan, and D. R. Radev, "LexRank: graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, vol. 22, pp. 457-479, 2004.

[16] O. Gross, A. Doucet, and H. Toivonen, "Document Summarization Based on Word Associations", *Proc. 37th International ACM SIGIR Conference on Research and Development In Information Retrieval,* pp. 1023-1026, Gold Coast, QLD, Australia, 2014.

[17] R. L. Cilibrasi, and P. M. B. Vita´ny, "The Google Similarity Distance", *IEEE Transactions on Knowledge and Data Engineering*, vol.19, no.3, 2007.