新資訊時代的啟發性資訊服務

Informative Services for New Information Era

陳光華

Kuang-hua Chen

台灣大學圖書館學系

Department of Library Science

National Taiwan University

E-mail: khchen@ccms.ntu.edu.tw

【摘要 Abstract】

現今網際網路上的搜尋引擎通常給予使用者排序後的文件表列，部份的搜尋引擎則加上簡短的文件描述，期望使用者藉此判斷文件的相關性。然而，這些描述性文字或是過於簡短，或是未經過有效的文本模型處理，通常無法提供足夠的訊息，甚至有誤導使用者的情形發生。在邁向新資訊時代的關鍵時刻，必須提供使用者更好的資訊服務，才能由浩瀚的電子文件中取得適切的資訊。本文將探討提供使用者適切資訊的服務策略，並且將重點放在文件的自動摘要。文件摘要作為原始文件的代表，一直是圖書館學領域重要的研究課題，於資訊檢索應用加上文件摘要的服務，讓使用者輕易地判斷文件的相關性，將是提高檢索效率的重要方法。

The services provided by search engines in Internet is a list of ranked documents nowadays. Some search engines attach a short description for each retrieved document. However, on the one hand, these descriptions are too short to be informative. On the other hand, these descriptions are not considerate representations for original documents. As a sequence, it is not easy for users to judge the relevance. In addition, these descriptions maybe mislead users. In the meantime towards the new information era, much better services should be provided for users to easily choose what they want from a great amount of documents. This paper discusses some possible approaches that provided informative services and then focuses on the automatic summarization. Providing summarization services for applications of information retrieval will help users judge the relevance of a document in an effective way. No doubt, document summarization is an important strategy to promote the performance of information retrieval.

一、緒論

對於生活於 20 世紀的人類而言，世界的變化是急遽又富挑戰性的。各種訊息傳遞技術的發明，使得資訊的傳遞速率越來越快，如今通訊與電腦技術的結合，網際網路的全球連線，更將人類的生活方式帶進前所未有的境地。例如，吾人可以透過網路連結各大圖書館，待在家裡就能夠查閱圖書館的館藏，一旦確認有需要的圖書，亦可線上預約；若有電子版本，則更可以線上覽讀，將電子圖書館視為家中書房的延伸，足不出戶就能飽讀群書。

網際網路可說是世紀末超級的新興媒體，原本僅是學術上用於溝通訊息的通訊管道，經由 WWW 迷人的瀏覽介面，以新的面貌讓網路使用者更方便地使用網際網路。這種新型態的資訊載體展現超乎想像的功能，也讓普羅大眾體會了看不到、摸不到卻又無所不在的真實感受。網際網路發展到這樣的階段，事實上已經將媒體的掌握權解放了，任何人都可以將個人的意見、出版品放到網路上，不再受制於出版商。而吾人實際使用網際網路幾年後，更能夠掌握它的特性，各種型態的資訊陸續出現，各種格式的資料推陳出新，網際網路呈現百家齊鳴的現象。

網際網路仍然持續地蓬勃發展，透過網路傳輸的資訊量越來越大，使用者以往苦惱於無法取得資訊，今日卻面臨了資訊爆炸的問題，獲得有用資訊的代價越來越高。如何協助讀者或是尋求資訊的人們取得有用的資訊，成為圖書館學與資訊科學研究領域中非常重要的課題。而資訊檢索的相關研究已經進行很長的一段時間，早期在獨立而封閉的環境運作，今日則處於開放的環境，然而無論面對的環境如何改變，其協助讀者或使用者取得有用、適用資訊的目標卻無二至。

本文討論新資訊時代應有的資訊服務，第二節說明目前網際網路提供的服務，及其不足之處。第三節則討論一種更具啟發性的資訊服務 -- 自動摘要，並描述美國 TIPSTER 文件計畫相關的學術活動。第四節提出筆者使用的自動摘要模型，其中重要的考量因素為詞彙的關聯、句子的位置、以及線索詞彙的使用。第五節是簡短的結語。

二、網際網路服務

目前網際網路的服務大致可分為遠端登錄（Telnet）、電子郵件（E-mail）、新聞討論群（Usenet）、全球資訊網（World Wide Web，簡稱 WWW）等等，吾人可將上述服

務視為原生性服務（Primitive Services），提供吾人取得資訊的基本功能。其中 WWW 在 1993 年 Mosaic 瀏覽器推出後，迅速吸引學術界的眼光；再經由 Netscape 公司與 Microsoft 公司在瀏覽器市場的爭霸戰，以及商業體系的推波助瀾，WWW 的使用者迅速攀升，若參考相關的統計值，更可以獲得佐證。（註 1）

衡諸 WWW 的龐大使用群，各種建構於 WWW 的加值型服務應運而生，其中最受注目的是搜尋引擎（Search Engine）以及主題指引（Subject Directory）。透過搜尋引擎，使用者可以取得特定事件的相關資訊；透過主題指引，使用者可以取得相關主題的資訊。由於該類服務背後的商業利益相當龐大，新的搜尋引擎以及主題指引不斷推出，目前知名的業者如表一所示（註 2）。此外，特殊的搜尋引擎服務也陸續嶄露頭角，如 BigFoot、Four11 等找人的服務，或是搜尋 e-mail 的服務。綜而言之，上述的加值型服務對於處於世紀末資訊浪潮的人們有相當大的貢獻，已經可以初步地過濾資訊，減少吾人的資訊負載。

表一、WWW 之加值型服務

	搜尋引擎	主題指引
國外	AltaVista (http://altavista.digital.com/)	Yahoo (http://www.yahoo.com/)
	Lycos (http://www.lycos.com/)	Galaxy (http://www.galaxy.com/)
	OpenText (http://www.opentext.com/)	PlanetSearch (http://www.planetsearch.com/)
	Northern Light (http://www.northernlight.com)	StartPoint (http://www.stpt.com/)
	InfoSeek (http://info.infoseek.com/)	The WWW Virtual Library (http://vlib.stanford.edu/Overview.html)
	Excite (http://www.excite.com/)	Magellan (http://www.mckinley.com/)
	HotBot (http://www.hotbot.com/)	Deja News (http://www.dejanews.com/)
	SavvySearch (http://guaraldi.cs.colostate.edu:2000/)	
	WebCrawler (http://webcrawler.com/)	
國內	GAIS（蓋世） (http://gais.cs.ccu.edu.tw/)	Yam（蕃薯藤） (http://taiwan.ntu.edu.tw/)
	WhatSite（哇塞） (http://www.whatsite.com/)	
	聚寶盆 (http://spring.nii.nchc.gov.tw/Search/)	

雖然眾多搜尋引擎與主題指引已經提供吾人相當大的幫助，並且提供簡短的文件描述，然而檢索所得的文件仍然相當的多，而且這些描述通常無法判斷該文件是否為相關的文件，使用者必須連結檢索所得的文件，真正閱讀之後才能夠知道文件是否適用。這個情形造成的影響可以從兩個角度觀察：第一是吾人並沒有真正享受上述服務帶來的好處，很可能隨著文件的來來往往，卻沒有需要的文件，心情越來越沮喪；第二是文件的來來往往，使得網路流量大增，卻沒有達到實際的效用，造成網路不必要的負擔。

在進入 21 世紀的關鍵時期，在 NII、GII、NGI 等口號震天價響的新資訊時代（註 3），享用更好的資訊服務，並非是過分的要求。是否有啟發性的資訊服務讓吾人更有效地取得所需的資訊？筆者認為有兩個重要的研究目標：第一是資訊擷取（Information Extraction）；另一為自動摘要（Automatic Summarization）。筆者在臺灣大學圖書館學刊第十二期已經發表有關資訊擷取的文章（註 4），在此不予贅述；本文則著重於自動摘要的相關研究。

三、文件摘要

輔仁大學蘇遠教授於中國圖書館學會會報第 56 期發表的文章「自動摘要法」指出，摘要具有以下的功能：（註 5）

- 宣示功能（Announcement）：宣示原始文件的存在性
- 篩檢功能（Screening）：判定原始文件的相關性
- 取代功能（Substitution）：取代原始文件
- 回溯功能（Retrospective）：查詢原始文件

至於摘要的類型則分為指示性摘要、資料性摘要、評論性摘要、以及摘錄。（註 6）指示型摘要通常具有宣示功能與篩檢功能；資料性摘要則主要是具有取代功能；評論性摘要則比較特殊，這類型摘要的自動化處理非常困難；摘錄則是直接抽取文件的句子，其功能則視情形而定，很可能具有宣示、篩檢、以及取代的功能。至於回溯功能則是四種類型摘要皆具有的功能。

顧名思義，摘要是文件的精緻版（Finer Version），亦即以較少的文字表述原始文件所欲傳達的訊息。所謂的較少文字，圖書館學與資訊科學大辭典對此做的解釋為：「研究報告及專論，摘要宜少於 250 字，附錄及簡訊性質之資料，以 100 字為佳，至於社論

或讀者來函只需要一個句子即可，長篇論著，如：技術報告、學位論文，其摘要以一頁以內，且以 500 字為限」。（註 7）因此，如何在有限的文字內表達原始文件的微言大義，便是從事摘要研究的學者專家必須面對的重要課題。

至於「自動摘要」則是以自動化的程序製作原始文件的精緻版。若從自動摘要模型的角度檢視所謂的自動摘要，可以分為兩種作法：第一種可由文件中挑選適當的段落或句子構成摘要，亦即製作所謂的「摘錄」；第二種則可由分析原始文件的角度出發，抽取文件的「概念表意」（Conceptual Representation），再進行「摘要的產生」（Summary Generation）。這兩種作法各有其優缺點，基本上，第二種作法牽涉所謂「文件理解」的過程，若能夠真正達成所謂的「理解」，應該可以製作品質較高的摘要。然而對於網際網路的應用，筆者認為時間是一個非常重要的限制條件，而理解通常必須花費相當長的時間，因此前述第二種作法比較不可行。但是若採用離線處理（Off-line Processing）的方式，仍然是值得期待的作法，而且可以將自動摘要的流程模組化（Modularized），釐清造成摘要良莠的茫點。

在網際網路急遽發展的情況之下，文件的自動摘要逐漸受到學者專家的注意。在美國國防高等研究計畫機構（DARPA）TIPSTER 文件計畫的支持之下，將於今年首次舉辦自動文件摘要學術會議（Automatic Text Summarization Conference，簡稱 SUMMAC），廣邀世界各地相關的研究人員參與競賽。今年共有 21 個不同的研究團隊參加，美國以外地區的參賽團隊僅有英國、日本、以及臺灣，而臺灣地區僅有臺灣大學資訊工程學系陳信希教授與筆者組隊參賽。該項會議將評比三種不同用途的文件摘要，第一種稱為 Categorization Task；第二為 Adhoc Task；第三為 Q&A Task。

Categorization Task 的目標是評估自動摘要系統對於文件關鍵概念的掌握能力。參與競賽的團隊會取得大會準備的 500 篇文件，其中每 100 篇各與某一 Topic 相關，總共有五個不同的 Topic。必須稍加說明的是，這裡所稱的 Topic 並不是一般圖書資訊界認知的主題，而是對於資訊需求的描述，圖一是 Topic 的例子，而圖二則是 SUMMAC 要求參賽者製作摘要的原始文件。參賽的系統將製作完成的摘要送回大會，大會則聘請為數甚多的評估人員閱讀摘要，並要求他們據以設定該摘要的屬於哪一個 Topic，如果無法決定則設定為第六個 Topic。（註 8）大會接著依據評估人員的評估結果，計算參賽系統的績效。

<top>
<head> Tipster Topic Description
<num> Number: 001
<dom> Domain: International Economics
<title> Topic: Antitrust Cases Pending
<desc> Description:
Document discusses a pending antitrust case.
<narr> Narrative:
To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as well as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.
<con> Concept(s):
1. antitrust suit, antitrust objections, antitrust investigation, antitrust dispute
2. monopoly, bid-rigging, illegal restraint of trade, insider trading, price-fixing
3. acquisition, merger, takeover, buyout
4. Federal Trade Commission (FTC), Interstate Commerce Commission (ICC), Justice Department, U.S. Securities and Exchange Commission (SEC), Japanese Fair Trade Commission
5. NOT antitrust immunity
<fac> Factor(s):
<def> Definition(s):
Antitrust - Laws to protect trade and commerce from unlawful restraints and monopolies or unfair business practices.
Acquisition - The taking over by one company of a controlling interest in another, also called a takeover. The action may be friendly or unfriendly.
Merger - The acquisition by one corporation of the stock of another. The acquiring corporation then retires the other's stock and dissolves that corporation. Therefore, only one corporation retains its identity in a merger.
</top>

圖一、SUMMAC 使用的 Topic

Adhoc Task 的目標則是評估自動摘要系統是否能夠提供使用者找尋的資訊，是一種使用者導向（User-directed）的文件摘要。大會將提供參賽團隊 20 個 Topic，每一個 Topic 有 50 篇文章，共計 1000 篇文章，參賽的系統必須視 Topic 為使用者資訊需求的描述，依據 Topic 建構每一文件的摘要。當大會接獲參賽者製作完成的文件摘要，評估人員必須閱讀每一篇摘要，並且判定摘要是否與 Topic 相關。（註 9）

<DOC>
<DOCNO>WSJ911028-0008</DOCNO>
<DOCID>911028-0008.</DOCID>
<TEXT>
DALLAS -- Texas Utilities Co. reported a $765.7 million loss for the third quarter, reflecting a $1 billion nonrecurring charge taken because regulators won't let it recoup certain costs associated with its Comanche Peak nuclear power plant.

The utility blamed the quarterly loss almost entirely on the $1.01 billion after-tax charge, which it in August announced that it would take at the close of the quarter. In addition, the company also said it was recording a nonrecurring, after-tax charge of $37 million for fuel costs disallowed by the commission order.

On a per-share basis, the utility's loss was $3.66. Revenue in the quarter was $1.45 billion. In the year-ago quarter, Texas Utilities reported net income of $344.7 million, or $1.77 a share, on revenue of $1.41 billion. Excluding the effect of the disallowances, the company said earnings for the third quarter would have been $1.35 a share.

The utility said the charge results from a disallowance in the rate order issued by the Public Utility Commission of Texas in August for the company's principal subsidiary, Texas Utilities Electric Co. The commission ruled that $472 million of the expenditures reviewed in the construction of the Comanche Peak nuclear plant were imprudently incurred in 87.8% of the plant. The commission ordered an additional disallowance of $909 million of the expenditures related to the repurchase of a 12.2% interest in the plant from former co-owners.

Texas Utilities Electric plans to appeal the commission order to state district court, the utility said.

In New York Stock Exchange composite trading Friday, Texas Utilities rose 37.5 cents to $38.25 a share.
</TEXT>
</DOC>

圖二、典型的 SUMMAC 文件

參與前述兩類 Task 競賽的團隊，可以選擇製作定長摘要（Fixed-length Summary）或最佳摘要（Best Summary），或是兩者皆予以製作。所謂定長摘要，其長度不可超過原文的 10%；最佳摘要則無限制。然而文件摘要的長度為評比的項目，評估人員閱讀摘要的時間也是評比的項目，因此，過長的摘要是參賽團隊必須極力避免的。

Q&A Task 難度相對較高，SUMMAC 將 Q&A Task 產生的摘要假想為撰寫報告過程中所需的資訊，亦即為了撰寫有價值的報告，撰寫人員必須具有某些特定問題的相關資

訊，因此若有一文件自動摘要系統能夠針對特定主題摘錄所有相關文件中的相關資訊，將於莫大的助益。顯然 SUMMAC 也知道這個競賽項目並不容易，因此聲明這個競賽項目仍處於初期設計的階段。

四、自動摘要模型

有關自動摘要法的文獻探討，蘇誠教授發表的「自動摘要法」一文已有詳盡的討論（註 10），有興趣的讀者可以參閱該論文，本文不再贅述。本節將著重於筆者自己提出的自動摘要模型，筆者採取的作法是製作「摘錄」型摘要，亦即直接由文件擷取重要的句子，自動製作原始文件的摘要。

一般而言，組織完善、意念完整的文件，其名詞與名詞以及名詞與動詞的關係相當密切，模型的建構是基於下列的假設：

名詞與動詞共存於述語參數結構；而名詞間的關係是建構於言談層次。

欲自動建構文件摘要必須瞭解構成書面語的要素，也就是一般人撰寫文章的過程。文件是有生命的文字組合，並非是任意文字的交替出現，若能夠探究文字之間的關係，計算出哪些文字是文件的核心，如此可以大略知道作者的意念。因為意念的表達是以詞為單位，應該以詞彙的層次而非字與字之間的關係作為建構文件模型的基礎。（註 11）筆者使用四種詞彙的統計值，如下所示：

- 詞彙的重要性
- 詞彙的重複性
- 詞彙的共現性
- 詞彙的距離

詞彙的重要性代表的是，當它出現於文件時，做為作者意念核心的機會，也就是當讀者重建作者創作時的心智活動，由文件挑選適當詞彙做為文件主題的機會。並不是所有的詞彙都一樣重要。例如，若是將文件中的冠詞、副詞、以及介系詞等詞彙刪除，仍然能夠知道這份文件的梗概，這說明了上述的詞彙並不十分重要。反觀之，名詞與動詞就十分重要了。詞彙的頻率常常可以代表某種程度的重要性，這種情形，尤以一般的資訊檢索系統為最。然而，詞彙的重要性無法由 TF 完全顯示，因為所謂的重要性是針對

文件而言，並非詞彙本身重要與否。因此 IDF 才能代表詞彙對文件的重要程度。（註 12）當訓練語料的數量夠大時，IDF 值具有相當高的穩定性，可據以計算詞彙的重要性，IDF 可以使用下列的數學式計算求得。

$$IDF(w) = \log((P\text{-}O(w))/O(w)) \quad (1)$$

P 是某一文件集合的文件總數，$O(w)$是包含詞彙 w 的文件總數。當詞彙 w 出現於一半以上的文件，則其 IDF 小於等於 0，吾人可以認為這個詞彙一點都不重要，對文件集合中的文件不具有鑑別性。

意念一致的文件資料，作者使用的詞彙必然趨向某一個語意範疇。從統計的觀點，這表示該語意範疇的詞彙一起出現的機率比較大。判斷那些詞彙屬於同樣的語意範疇是相當困難的工作，但是由大規模的語料庫計算詞彙的共現程度就很簡單。可以使用共容訊息（Mutual Information，簡稱 MI）計算詞彙的共現，其數學式分別如下所示：（註 13）

$$MI(t_i, t_j) = \log\frac{P(t_j \mid t_i)}{P(t_j)} = \log\frac{P(t_i, t_j)}{P(t_i)P(t_j)} \quad (2)$$

共容資訊的意義是，當詞彙 t_i 與詞彙 t_j 經常一起在語料庫出現，聯合機率 $P(t_i,t_j)$會甚大於 $P(t_i)\times P(t_j)$，因此 $MI(t_i,t_j)$會甚大於 0；當 t_i 與 t_j 出現的方式是背道而馳時，$MI(t_i,t_j)$會甚小於 0；當彼此沒有什麼關係時（以機率論的術語而言，也就是互相獨立），因此 $P(t_i,t_j)\cong P(t_i)\times P(t_j)$，所以 $MI(t_i,t_j)$接近於 0。

詞彙的位置也很重要。基於文件是有生命的文字組合的觀點，相關的詞彙其出現的距離必定不會太長。因為，一旦相隔太遠，彼此之間的相乘效果就大打折扣，這不會是一般作者的用意。引入距離的因素，比較能夠忠實反應寫作的行為。距離的計算可採用如下的方式，首先為每一個名詞與動詞設定一個編號，以下面這一段文字為例：

蘇聯$_1$ 許多 製造$_2$ 民生$_3$ 日用品$_4$ 的 工業$_5$ 得到$_6$ 政策性$_7$ 的 補貼$_8$，其 目的$_9$ 是 保持$_{10}$ 物價$_{11}$ 的 平穩$_{12}$。但 補貼$_{13}$ 勢 難 普及$_{14}$ 於 各行各業$_{15}$，因此 又 造成$_{16}$ 某 些 日用品$_{17}$ 不足$_{18}$ 或 完全 缺乏$_{19}$ 的 後遺症$_{20}$ 。 現在 既然 要 引進$_{21}$ 市場$_{22}$ 經濟$_{23}$，補貼$_{24}$ 政策$_{25}$ 又 勢 難 繼續$_{26}$，一旦，放棄$_{27}$，許多 民生$_{28}$ 物資$_{29}$ 的 價格$_{30}$ 必然 上漲$_{31}$，於是 又 引出$_{32}$ 民間$_{33}$ 屯積$_{34}$ 物資$_{35}$ 與 通貨膨脹$_{36}$ 的 壓力$_{37}$。

詞彙 X 與 Y 的距離 $D(X,Y)$可以用以下的方式計算：

$$D(X,Y) = \mathrm{ABS}(C(X)-C(Y)) \tag{3}$$

ABS 為絕對值函數，$C(X)$代表詞彙 X 的編號，如 C(政策性) = 7，而 C(目的) = 9，所以 D(政策性,目的) = 2。

綜合以上因素，計算名詞重要性的模型為：

$$CS(n) = pn \times SNN(n) + pv \times SNV(n) \tag{4}$$

$CS(n)$為名詞 n 的聯結強度（Connective Strength）；$SNV(n)$為名詞 n 與其他動詞的強度；SNN 為名詞 n 與其他名詞的強度；pn 與 pv 分別為 SNN 與 SNV 的權重，可藉由消去內插法（Deleted Interpolation）計算。（註 14）$SNN(n)$與 $SNV(n)$的計算方式如下：

$$SNN(n_i) = \sum_j \frac{IDF(n_i) \times IDF(n_j) \times f(n_i, n_j)}{f(n_i) \times f(n_j) \times D(n_i, n_j)} \tag{5}$$

$$SNV(n_i) = \sum_j \frac{IDF(n_i) \times IDF(v_j) \times f(n_i, v_j)}{f(v_i) \times f(v_j) \times D(n_i, v_j)} \tag{6}$$

$f(w)$為詞彙 w 的頻率，$f(w_i,w_j)$為詞彙 w_i 與 w_j 共同出現的頻率；$D(w_i,w_j)$為 w_i 與 w_j 之間的距離。可以看出整合了前述的四項考量因素，事實上，$f(w_i,w_j)/(f(w_i)\times f(w_j))$即為計算詞彙共現的程度，與 MI 具有相同的型式，或許可稱之為共容頻率（Mutual Frequency，簡稱 MF）。

一旦求得每一個名詞的聯結強度，便能夠進而得到每一個句子的重要性。假設某一個句子 s_i 有 m 個不同的名詞，該句子被摘錄的可能性度量，若以摘錄強度（Extraction Strength，簡稱 ES）稱之，可以用下列數學式度量：

$$ES(s_i) = \sum_{j=1}^{m} CS(n_{ij}) / m \tag{7}$$

文件的句子經由前述的方式可以排成有序集合（Ordered Set），文件的摘要就可以由該有序集合擷取數量適當的句子組成。若要製作 SUMMAC 所稱的定長摘要與最佳摘要，則可以設定一個門檻值（Threshold），刪除有序集合中摘錄強度小於門檻值的句子即構成最佳摘要，從最佳摘要再刪除部份的句子，使得有序集合中句子數小於原文的 10% 即構成定長摘要。

純粹藉由文字間相互關係建構自動摘要的模型，到此可說是已經完成。然而，無論是從文獻的討論或是個人閱讀的經驗，吾人可以發現句子的位置事實上扮演重要的角

色，而某些線索詞彙（Cue Word）也具有舉足輕重的份量。因而若考慮這兩個因素可進一步修正數學式(7)為：

$$ES(s_i) = w_1 \times \sum_{j=1}^{m} CS(n_{ij}) / m + w_2 \times POS(s_i) + w_3 \times CW(s_i) \qquad (8)$$

(8)式中的 *POS* 為句子 s_i 位置的度量；*CW* 為線索詞彙的度量；w_1，w_2，w_3 則為相對的權重。依據經驗法則（Heuristic Rule），文件的第一個段落與最後一個段落通常傳遞文件的重要訊息，而第一段與最後一段的第一個句子又特別重要。所謂的線索詞彙又分為增益詞（Bonus Word）與損益詞（Stigma Word）（註 15），例如重要、顯著等等為增益詞；不可能、幾乎不等等為損益詞。（註 16）

筆者使用前述的模型處理 SUMMAC 大會提供的文件，製作完成的摘要已於 1998 年 2 月 16 日送回 SUMMAC。SUMMAC 將於同年 5 月 4 日舉辦評比成果的發表會，屆時可得知系統的績效，以及與其他團隊提出的自動摘要系統彼此間的差異，因此，本文目前無法說明評比結果。

五、結語

「知識就是力量」這句話精確又殘酷地說明目前世界文明發展的態勢，網際網路更加速了資訊的流通，縮短了資訊形成知識所需要的時間。然而網際網路膨脹地過於快速，資訊累積太快造成雜訊過多，卻又干擾了知識的形成。網際網路的各類加值型服務業者遂為處於世紀末的人們提供搜尋引擎與主題指引以及其他特殊的服務，希望讓網際網路使用者能夠有效地檢索文件、取得資訊。在即將邁入下一個世紀的當口，應該提供怎樣的新服務，讓使用者更容易判斷文件的相關性，應是資訊檢索研究人員必須注意的課題。

本文認為文件摘要將是新資訊時代一種重要的啟發性資訊服務，對於網際網路使用者而言，可以透過文件摘要快速地判讀文件的相關性，而不必取得完整文件之後才發覺文件根本不符合需求；此外，文件摘要的服務也能夠有效降低網際網路的流量。對於文件的使用者與整體網際網路環境，文件摘要都是值得期待的服務。雖然，人工製作文件摘要具有高品質的特性，但是緩不濟急，審視網際網路文件數量極為龐大的事實，自動化的文件摘要是無法避免的作法。本文提出一個可能的作法，該模型建構於詞彙的重要性、詞彙的重複性、詞彙的共現性、詞彙的距離等四項文件文字之間的要素，並且考慮

文件中句子的位置以及線索詞彙。然而，為了因應網際網路的特性，該模型仍有待進一步的實驗與修正，以適應各種不同類型的文件，並且必須更加縮短所需要的計算時間。

附　註

註 1： 請參考「蕃薯藤第二次台灣網際網路使用調查。」（http://taiwan.yam.org.tw/survey/survey97/）。

註 2： 目前部份的搜尋引擎業者整合了主題指引的功能；而部份的主題指引業者也整合了搜尋引擎的功能。

註 3： NII 為 National Information Infrastructure 的縮寫，GII 為 Global Information Infrastructure，NGI 則為 Next Generation Internet。

註 4： 陳光華，「資訊的組織與擷取」，臺灣大學圖書館學刊第十二期（民國 86 年 12 月），頁 127-141。

註 5： 蘇誴，「自動摘要法」，中國圖書館學會會報第 56 期（民國 85 年 6 月），頁 41-47。

註 6： 同註 5。

註 7： 國立編譯館主編，圖書館學與資訊科學大辭典（台北市：漢美，民國 84 年），頁 2002。

註 8： 此時評估人員並不知道文件的 Topic 為何，他們必須就摘要本身判定文件究竟屬於哪一個 Topic。

註 9： 此時評估人員已經知道文件的 Topic，他們必須判定摘要與 Topic 的相關性。

註 10： 同註 5。

註 11： 筆者亦使用該模型部份子系統處理文件的主題辨識，細節請參閱下列文章。陳光華，「電子文獻主題之自動辨識」，中國圖書館學會會報第 59 期（民國 86 年 12 月），頁 43-58。

註 12： Sparck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." Journal of Documentation, 28.1 (1972): 11-21.

註 13： Church, K.W., and P. Hanks. "Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics,16.1 (1990): 22-29.

註 14： Jelinek, F. Markov. "Source Modeling of Text Generation." Ed. J.K. Skwirzynski. The Impact of Processing Techniques on Communication, Nijhoff, Dordrecht, The Netherlands, 1985.

註 15： Edmundson, H.P. "New Methods in Automatic Extracting." Journal of Association for Computing Machinery, 16.2 (1968): 264-285.

註 16： 依據中華民國分詞標準，「不可能」應該是「不」及「可能」兩個詞，然而在某些應用時，有時可能會將之合併，本文所指稱的增益詞或損益詞亦可能有這種情形。若要更精確地區別，可以說增益詞或損益詞也可能是複合詞。