黃仁鵬、張貞瑩（2014），『運用詞彙權重技術於自動文件摘要之研究』，
中華民國資訊管理學報，第二十一卷，第四期，頁 391-416。

運用詞彙權重技術於自動文件摘要之研究

黃仁鵬*
南臺科技大學資訊管理系

張貞瑩
南臺科技大學資訊管理系

摘要

目前各個搜尋引擎所產生的網頁摘要，大多無法提供使用者充足的摘要內容判斷資訊，更可能造成使用者的誤導。本研究希望搜尋引擎將查詢結果回傳給使用者時，不只是給予一些片斷不全的訊息，取而代之的是一個比較有幫助的摘要，使用者可以藉由此自動摘要，了解全文的概要，然後決定是否需要讀取網頁之全文。本研究運用權重技術針對網頁的內容進行文字探勘，藉由中研院所開發的中文斷詞系統（CKIP）進行斷詞，利用 TF-ISF 與相似度權重技術分別進行摘要實作，並透過其聯集與交集分別產生「概略摘要」與「精準摘要」，藉以提升自動摘要的品質。由實驗結果可證實本研究所提出之系統方法可以有效的提升文件自動摘要的正確性。

關鍵詞：自動文件摘要、文字探勘、網際網路探勘、資訊檢索、TF-IDF 演算法

* 本文通訊作者。電子郵件信箱：jehuang@mail.stust.edu.tw
2013/07/19 投稿；2014/06/14 修訂；2014/06/27 接受

Huang, J.P. and Chang, C.Y. (2014), 'Automatic Text Summarization based on Weights of Words', *Journal of Information Management*, Vol. 21, No. 4, pp. 391-416

Automatic Text Summarization based on Weights of Words

Jen-Peng Huang*
Department of Information Management, Southern Taiwan University of Science and Technology

Chen-Ying Chang
Department of Information Management, Southern Taiwan University of Science and Technology

Abstract

Purpose－The objective of text document summarization is to extract essential sentences that cover most of the concepts of a document so that users are able to comprehend the ideas of the documents which try to address by simply reading through the corresponding summary. This study aims to develop an automatic text summarization technique to product the summary of the web pages by extracting the sentences which cover most of the concepts of the web pages.

Design/methodology/approach－The research framework was developed from CKIP (Chinese Knowledge Information Processing) system and automatic text summarization techniques. Two studies were designed to elicit and evaluate the accuracy and applicability of the five automatic text summarization techniques with 10 samples from 184 web articles.

Findings－Our results show that TF-ISF (Term Frequency-Inverse Sentence Frequency) is better than the others in the evaluation of "F-measure". Further, "Rough Summary" and "Accurate Summary" respectively is the best performance in the evaluation of "RECALL" and "PRECISION".

* Corresponding author. Email: jehuang@mail.stust.edu.tw
2013/07/19 received; 2014/06/14 revised; 2014/06/27 accepted

Research limitations/implications－This paper focuses on Chinese web articles. Hence, future research is recommended to develop an automatic text summarization system based on Ontology-based architecture.

Practical implications － This paper provides several automatic text summarization techniques to product the summary of the web pages by extracting the sentences which cover most of the concepts of the web pages. The experimental results indicate that the proposed approach outperform a significant improvement on the accuracy of automatic text summarization.

Originality/value－This paper is the first that applies the union and intersection of "Rough Summary" and "Accurate Summary" to improve the quality of automatic text summarization.

Keywords: automatic text summarization, text mining, Web mining, TF-IDF

壹、緒論

隨著網際網路的崛起、資訊科技的日新月異，網路上充滿著大量的資訊。面對如此龐大的資料量，如何從大量資訊中過濾、篩選出符合使用者所需資訊才是使用者關心的焦點。為了從網際網路大量的資訊量中，準確的獲取使用者所需資訊，文件自動摘要處理變得越來越重要。

目前搜尋引擎所搜尋出來的摘要結果，會依所輸入的搜尋文字（關鍵字）而有不同的差異，各個搜尋引擎所產生的網頁摘要的方法都是雷同的。各個搜尋引擎所產生的網頁摘要，大多無法提供使用者充足的摘要內容判斷資訊，更可能造成使用者的誤導。有鑒於此，本研究希望搜尋引擎將查詢結果回傳給使用者時，不只是給予一些片斷不全的訊息，取而代之的是一個比較有幫助的摘要，使用者藉由此自動摘要，判讀是否需要讀取網頁之全文。本研究運用權重技術針對網頁的內容進行文字探勘，希望藉由中研院所開發的中文斷詞系統（Chinese Knowledge Information Processing, CKIP）進行斷詞並透過 TF-ISF 與相似度權重技術分別進行摘要實作，並使用其聯集與交集分別產生「概略摘要」與「精準摘要」，藉以提升搜尋的品質。本研究以中文網頁為設計對象及實驗資料皆針對純文字型態的文件，故多媒體文件及網路文件等資料將不適用於本研究，若其網頁內容中夾雜英文詞句，將其標記為外來語而不處理。

本文共分為六章，第一章為緒論，第二章為文獻探討，第三章為研究方法，第四章為系統開發與實作，第五章為實驗評估，第六章為結論。

貳、文獻探討

一、文字探勘

文字探勘（Text Mining）一詞最早始於早年利用大量人力在文字資料上萃取出資訊的行為。文字探勘是資料探勘的延伸應用，其涵蓋資訊檢索萃取、計算語言學、自然語言處理、資料探勘、機器學習等跨領域知識，特別強調從非結構（un-structured）或是半結構（semi-structured）的文字中發掘出未知、隱含且有用的資訊。Sullivan 學者定義文字探勘為「一種編輯、組織及分析大量文件的過程，主要提供分析人員或決策者等特定使用者對特定資訊（如摘要、關鍵字），發現資訊特徵及其間的關聯性」（Sullivan 2001）。近年許多學者也應用文字探勘技術至各種研究（Wei 等 2013; 李俊宏與張興亞等 2007），以持續演進文字探勘的發展與應用，可見其發展與應用的趨勢。Losiewicz 學者提出一個文字探勘的架構（Losiewicz 等 2000），此架構包含了資料集（Data Collection）、資料倉儲（Data Warehousing）

與資料探索（Data Exploitation）三個功能，每一個中皆有兩種子功能。資料集主要涵蓋資料來源的選擇與文件選擇；資料倉儲主要涵蓋資料轉換與資料儲存；資料探索主要涵蓋資料探勘與探勘結果呈現。

二、中央研究院中文斷詞系統

基於建構中文自然語言處理的資源與研究環境，本研究使用中央研究院跨所成立的詞庫小組（Chinese Knowledge Information Processing Group, CKIPG）所開發的中文斷詞系統為工具，該系統為結合詞庫斷詞法及 N-Gram 選詞法之優點的混合斷詞法，將使用者所輸入之文章或文句自動斷詞後，再標示出每個詞的詞類標記，且該系統具有新詞辨識能力與附加詞類標記的功能。根據實驗數據顯示，其系統斷詞的精確度為百分之九十以上。此一系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料（鄒明城等 2010）。

三、文件摘要種類與產生方式

文件摘要大致上可分為「單文件摘要」（Singular Document Summarization）與「多文件摘要」（Multiple Documents Summarization）。「單文件摘要」是將單篇的文件內容作精簡化與重點化，注重的是能否有效的刪除不必要的資訊，留下真正的文件內涵資料，以達到摘要之精簡化與重點化；而「多文件摘要」則是將多篇探討類似主題或事件的文件整合在一起，除了刪除不必要的資訊外，尚需有效率地過濾重複在多篇文件中所出現的資訊。本研究的摘要則是以單文件摘要為主。摘要產生方式可以分為：「以抽象為基礎的（abstract-based）摘要」方式與「以抽取為基礎的（extract-based）摘要」方式。（Dalal & Zaveri 2011; Das & Martins 2007; Gupta & Lehal 2010; Mani & Maybury 1999）。由於「以抽象為基礎的摘要」作法較為困難，目前研究大多「以抽取為基礎的摘要」為主。本研究亦採用「以抽取為基礎的摘要」的作法。

四、文件自動摘要相關研究

文件自動摘要方法是利用文字探勘技術，由電腦自動地從原始資料中精練出最重要資訊且不失原意的過程。其可以提供使用者資訊擷取後簡潔的文件內容，以減少使用者閱讀原始文件的時間與精力，得到所想要的資訊，幫助使用者快速的過濾不需要的資訊（黃純敏與吳郁瑩 1999；黃純敏等 2011；魏玲玉與曾守正 2006）。文件自動摘要相關研究大致可以分為 4 大類（Dalal & Zaveri 2011）：(1)探索式技術（Heuristic Techniques）、(2)語義基礎技術（Semantics-based Techniques）、

(3)查詢導向技術（Query-oriented Techniques）、(4)分群導向技術（Cluster-based Techniques）。

本研究主要是屬於探索式技術。萃取過程通常會選取適當的特徵表達句子，此特徵為評斷句子重要程度的依據，以便篩選出文中的重要句子。較常考量的特徵可包含以下幾種：(1)以統計方法計算詞彙權重（黃純敏等 2002），如 TFIDF 法（黃純敏等 2011；黃純敏等 2002）；(2)考量句子於文中的位置；(3)計算句子之間的相似度（陳姿妤等 2007）；(4)利用 intra-document linkage pattern 找出文章特徵結構並萃取出文章摘要（Salton, Singhal, Mitra, & Buckley 1997）。

文件自動摘要在作業前，不需要事先取得特定領域的語法架構，因此可廣泛的應用於任何領域。其中 A. Harris and M. Oussalah 等學者的研究(Harris & Oussalah 2008)，在自動摘要作業的初期字詞處理階段，將關聯性高的句子，視為相同的句子計算權重，並且依照句子的位置做加權的處理，所以摘要的可讀性與流暢性較差，為此有學者提出選取文章句子中資訊涵蓋量較多的句子做為自動摘要的依據（Abdel Fattah & Ren 2008; Ji 2008; Ren, Li, & Kita 2001）。

而單文件摘要研究的文件也大多以技術文件，所以研究方法大多是以探索式技術為主。例如特殊字詞（例如 Keywords, Title-words）出現的頻率或顯著因素的高低做為句子的選擇準則（Das & Martins 2007; Luhn 1958）。也有學者（Baxendale 1958）以句子的位置去找出文件中顯著的部份。這個方法被廣泛用於很多複雜的機器學習系統中。在中文單文件自動摘要研究方面，陳姿妤等人利用 TFIDF 做為基本詞彙權重計算的依據（陳姿妤&魏世杰 2007）。

探索式技術用於一般評估文章的重要性，可以考慮詞彙在文章中所出現的頻率以及其關鍵詞彙所在位置，以推估其文章的重要性，因此多採用權重計算方式。上面所描述的方法皆各有其優缺點，而他們所做的研究概念皆是來自於 TFIDF、相似度（Salton 1989）和 GBP 幾個自動摘要的架構，下面我們將針對這幾種摘要的作業方式分別做詳盡的探討。

（一）TF-IDF

TF-IDF（Term Frequency-Inverse document Frequency）是一種利用字詞的出現頻率，來加以衡量字詞權重的一種方式，常被應用於 Data Mining 與 Text Mining 相關研究領域上，用來計算某個字詞在文章中的相對重要程度。由於每篇文章中的詞，占整篇文件的重要性是不相同的。依 TF-IDF 定義，一篇文章中，任一字詞的重要性與該字詞在文章中的出現頻率成正比，但與該字詞在文件集中出現的頻率成反比（Salton & McGill 1983）。利用 TF-IDF，可以藉此找出某一特定網頁內高出現頻率的詞彙，但在整個網頁內容中具有低出現頻率的詞彙，此詞彙可以代表作為該網頁的關鍵詞代表，同時該詞彙可以產生出高權重的 TF-IDF 值。

（二）相似度計算

相似度計算運用非常廣泛，不論是文件的群聚、分類、檢索或者是自動摘要等等，均需要透過相似度計算進行處理（Salton 1989）。一般要瞭解句子之間的關聯程度，均透過關鍵字來求得彼此的相似度，其作法會設定一個門檻值作為基準，以判斷是否已達到此基準值。通常衡量相似度均採用向量內積公式，但因其所產生的值較離散且不平均，於是多位學者提出正規化的改正。Salton 學者於 1989 年指出常用的相似度計算公式（Salton 1989）如向量內積、Dice 係數、Cosine 係數、Jaccard 係數，其共同特點是分子均以兩個資料交集，相當於計算詞彙共同出現在兩篇文章的次數，其中 Cosine 係數與 Jaccard 係數，是適用於任兩個句子相似度的計算，而 Dice 係數則用來計算兩個句子中詞彙之間的相似度（柯淑津 2003），通常相似度值越高代表兩句子意思越相近。

（三）Global Bushy Path

Salton 等人所提出的本文關聯地圖（text relationship map）（Salton 等 1997），是將相似度較高的段落相互鏈結（Link）。在地圖中，每一個節點代表一個段落。若兩段落具較高的關聯性，則產生鏈結，它是以 Jaccard 係數做為計算段落間相似度的方法，因此相似度值就會介於 0 與 1 之間，全篇文章的分段落權值，則為該段落與其他段落的相似度總和，亦即連結點越多的段落權值越高，也代表該段落越重要。Salton 所提出的內文關聯法除了 GBP 外，還有提出 Depth First Path（DFP）與 Segmented Bushy Path（SBP）等方法，因 DFP 所產生的摘要語意雖較為連貫，但所包涵的主題則不如 GBP 廣泛。而 SBP 主要是探討相鄰近的幾個段落之間的相似性，卻容易忽略各個段落與所有句子的關係，因此本研究採用 GBP 做為其中一種實作方法。

五、自動摘要評估

近年來，文件自動摘要的方法被廣泛探討與研究（黃純敏等 1999），許多研究提出不同的摘要方法，來提昇摘要對於表達文件內容的正確性及代表性。目前被大家所廣泛使用的摘要評估方法，大致可分為主觀評估與客觀評估二種。主觀評估是指以人工的主觀判斷與評估作為自動摘要評估結果。客觀評估則是以計算摘要的正確率作為評估結果，其方法是比對自動摘要與參考摘要之間的相關程度，作為摘要的正確率，常見的評估方法如召回率（Recall）、精確率（Precision）與 F-measure（李麗華等 2009；陳姿妤&魏世杰 2007）。使用探索式技術的文件自動摘要的相關研究大多會結合多種不同的探索式技術，然後再使用不同的評估方法去評估自動摘要的品質，本研究也是使用多種探索式技術，因此也是採用此方式去評估自動摘要的品質。

參、研究方法

本研究之系統主要包含「網頁搜尋與擷取模組」、「網頁前置處理模組」、「自動摘要模組」與「摘要呈現模組」四個模組，其系統流程如圖 1 所示。

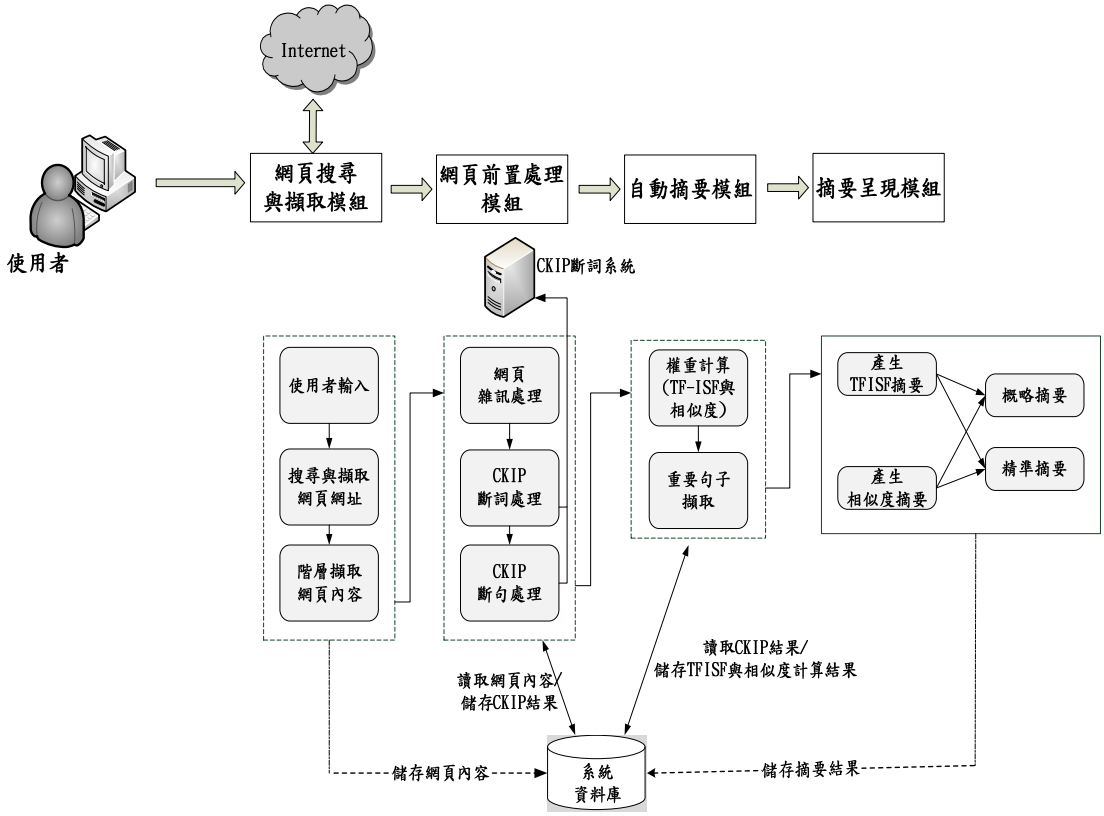

圖 1：系統流程圖

一、網頁搜尋與擷取模組

網頁搜尋與擷取主要是在取得與關鍵搜尋條件相合符的網頁，並存入到資料庫中。網頁搜尋與擷取模組主要分為以下兩個部分：

1. 使用者輸入：在使用者輸入部份，主要是讓使用者可以依需求，來輸入搜尋字串。
2. 搜尋與擷取網頁位址：將搜尋字串輸入 Yahoo 等搜尋引擎，並將所搜尋結果網頁的 URLs、網頁名稱以及網頁摘要儲存至資料庫中，以提供其他模組使用。

二、網頁前置處理模組

網頁前置處理模組主要是在對網頁內容做雜訊處理、中文斷詞以及萃取有用的關鍵字或特徵詞。網頁前置處理模組主要分為以下三個部份：

1. 網頁雜訊處理：由於抓取回來的網頁除了我們看到的文字內容之外，還包含了很多的 Html 語法設定。故將取回回傳的網頁原始碼後，針對網頁的純文字原始檔進行雜訊之去除，以整理出乾淨的網頁內容，最後僅網頁內文的部分，以增加文字探勘的確準率。
2. 網頁 CKIP 斷詞處理：進行網頁內容探勘之前，需先找出文件中的詞彙，特別是重要的關鍵詞彙。為了提高擷取詞彙的正確率，本研究在詞彙擷取部分，以中研院資訊科學研究所詞庫小組（CKIP）所提供之中文斷詞系統來進行中文文件之斷詞與詞性標注服務。因為在一個句子裡，動詞與名詞通常是句子的核心，在自動摘要的研究中，有學者採用動詞與名詞當作重要詞彙的依據（黃純敏、吳郁瑩）。本研究也是只擷取動詞與名詞，將這兩詞性視為與內文最具有相關性的重要詞彙。
3. 網頁 CKIP 斷句處理：通常摘要的形成是以句子為基礎，因此首要的工作是將網頁內文進行斷句。

三、自動摘要模組

自動摘要模組使用文字探勘技術與權重技術產生網頁摘要。自動摘要模組主要分為權重計算（權重計算又細分 TF-ISF 與相似度的計算方式）與重要句子擷取兩個部份：

（一）權重計算

雖然透過中文斷詞系統判斷出所有網頁內容中的詞彙，但是所斷出來的詞彙仍然無法代表是網頁的關鍵詞彙。因此，若要選出具有代表性的關鍵詞，就得計算各個詞彙在網頁內容中的權重值。本研究透過陳姿妤與魏世杰學者所提出的 TF-ISF（陳姿妤&魏世杰 2007）與 Salton 學者（Salton 1989）所提出的 Jaccard 與 Cosine 相似度計算方法計算出詞彙的權重值。

1. TF-ISF

   根據 TF-IDF 的計算方法，其詞彙的權重與詞頻成正比，而其出現的文章篇數成反比，依此方法計算出詞彙的權重值。但因為本研究為單篇網頁摘要，無須考慮詞彙的跨文件詞頻（IDF）。因此，本研究以詞彙在單篇網頁內文中的出現句子數倒數，即 inverse sentence frequency 取代 IDF，計算每篇詞彙之權重。其 TF-ISF 計算方式如公式 1 所示：

$$W_{ij} = \frac{n_j}{n_{all}} * \log_2 \frac{N}{df_j} \quad (1)$$

n_j：詞彙 j 在此網頁內文中出現的次數
n_{all}：網頁 i 中所出現的詞彙總數
N：網頁內文中所有句子的總數
df_j：網頁內文中包含詞彙 j 的句子數

本研究利用 TF 演算法，藉此找出某一特定網頁內文出現頻率的詞彙，其計算方式如公式(2)所示：

$$tf_{ij} = \frac{n_j}{n_{all}} \quad (2)$$

但在整個網頁內容中具有低出現頻率的詞彙，此詞彙可以代表作為該網頁的關鍵詞彙代表，其計算方式如公式(3)所示：

$$isf_j = \log_2 \frac{N}{df_j} \quad (3)$$

2. 相似度計算

句子的相似度計算，則是利用句子間關鍵詞重疊多寡來求句子之間的相似度大小，一般會設定一個門檻值作為判斷句子之間是否達到高相似度值，而在文章中與較多句子具有高相似度時，相對地代表此句子在文章中愈重要。其目的是將兩個句子之間的關聯性及相似度量化，使我們可以很快的藉由數字了解兩者之間的依存性。若是兩句子皆存在數個重要詞彙，就將這些重要詞彙的相似度權重值相加後，即表示此句子的權重值。本研究將以 Jaccard 係數與 Cosine 係數兩個公式（Salton 1989）下去做驗證，並擷取前 30%句子做為後續處理。

（二）重要句子擷取

網頁句子的權重皆以 TF-ISF 與相似度計算出後，即可以句子為主，進行重要句子的擷取。在此句子的重要性與排列的優先順序也就成為摘要成敗的重要關鍵。而在不失原本文義的原則下，我們選擇出擁有較高權重值的句子，並按照網頁內容原本句子的順序組合，以產生該篇網頁之 TF-ISF 與「相似度摘要」。

四、摘要呈現模組

摘要呈現模組則是將所產生的摘要呈現給使用者。產生 TF-ISF 與「相似度摘要」之後，我們利用 TF-ISF 與相似度產生的摘要以聯集與交集的方式，分別又產生出「概略摘要」與「精準摘要」,「概略摘要」可以以互補的方式，將 TF-ISF 與「相似度摘要」其中一方沒有擷取到重要句子加以彙整；而「精準摘要」可以將 TF-ISF 與「相似度摘要」,均擷取到的句子呈現給使用者，以加強擷取之命中率。故本研究將產生四種摘要：「TF-ISF 摘要」、「相似度摘要」、「概略摘要」、「精準摘要」,提供給使用者做為參考。為了與之前的研究做比較，在實驗中加上「GBP 摘要」(Salton 等 1997）做為比之參考依據。

五、自動摘要評估

由於自動摘要取決於使用者之主觀認定，有關自動摘要的評估，大多將系統產出的摘要與人工摘錄進行比較，自動摘要與人工摘錄重疊愈高，即表示該自動摘要成效越佳（黃純敏等 2011)。依據本研究所建置之系統，並蒐集網路上的網頁來驗證此系統的效益，藉由使用者對摘要的評選，以探討系統摘要結果是否能夠真正的代表該篇網頁摘要。評估之目的主要是針對本研究建置完成之摘要系統進行效益評估。

本研究將透過主觀評估與客觀評估二種評估方式進行評估。在客觀評估方面，實際讓使用者進行文章的評選，蒐集使用者對文章所挑選出的句子與系統摘要所得到的結果，以分析人工摘要與系統摘要之間有無差異，藉此驗證本研究所提之摘要系統的可行性，並透過召回率與精確率來進行網頁摘要內容重要性評估。在主觀評估方面，透過五等地評分法，依序為非常同意、同意、普通、不同意、非常不同意，進行評估系統摘要內容是否能夠有效表達原文大意以及語句順暢部分，讓評估人員依照個人的想法決定摘要結果所屬的等級，以反映本研究是否能達到所預期的功效。

在實驗過程中，每篇網頁將由 3 位評估人員進行評選，為避免單一人的評估過於主觀而造成誤差，故取 2 位以上共同認定的網頁 為評估準則，使評估能得到較客觀的結果。

六、實驗設計

依據本研究所提出的自動摘要法，建置文字探勘技術與權重技術自動摘要系統，並收集網路上之純文字型態的網頁來驗證此自動摘要系統的效益評估。藉由使用者的評選，以探討本研究所提出之自動摘要法的成效與適用性。

（一）實驗目的與方法

實驗目的主要是針對本研究所建置完成之自動摘要系統進行評估。本實驗設計了兩個實驗，透過主觀評估與客觀評估二種評估方式進行評估，它們主要目的與作法，如表 1 所示。根據文獻指出，自動摘要與人工摘要重疊愈高，即表示該自動摘要成效越佳（黃純敏等 2011）。在實驗一部份，即是利用客觀評估的方式，藉由使用者對文章進行評選，以分析人工摘要與系統之間有無差異，以驗證自動摘要方法的成效。且每篇抽樣由 3 位評估人員進行評選，並選取有 2 位評估人員均認同之樣本做為評估的準則，使評估能得到較客觀的結果；在實驗二部份，即是利用主觀評估的方式，將實際地讓使用者針對本研究所產生的五種自動摘要進行評選，以驗證五種自動摘要法的適用性。

表 1：實驗目的與方法

	目的	方法
實驗一	驗證所提出的自動摘要法之成效。	客觀評估：本實驗將系統所產出的摘要與人工摘要進行比較。
實驗二	驗證所提出的自動摘要法之適用性。	主觀評估：請評估人員檢視系統所產生的摘要且填答問卷。

（二）實驗對象

本實驗評估人員的數量上，會因為實驗的目的與方法不同而有所差異。在實驗一部份，將邀請 10 位評估人員進行自動摘要的成效評估實驗，實驗對象為南台科技大學資訊管理系之研究學生，每位評估人員針對三個不同的樣本進行摘要之評選；在實驗二部份，將邀請 150 位評估人員進行自動摘要之適用性評估實驗，實驗對象為經常使用搜尋引擎做搜尋的使用者，每位評估人人將針對一份問卷進行評選，最後蒐集所有資料，以進行分析探討。

（三）實驗工具

本實驗工具，會因為實驗的目的與方法不同而有所差異。實驗一的部份，將利用紙本的方式，讓評估人員進行摘要之評選，並利用系統計算出召回率與精確率；在實驗二部份，將利用 mySurvey（http://www.mysurvey.tw）進行線上問卷調查，蒐集使用者對於自動摘要系統之評估資料。

（四）問卷信度分析與效度分析

為了確保問卷具有較高的可靠性與有效性，在形成正式問卷之前，本研究透過 SPSS 統計分析軟體對問卷測試結果進行信度分析（Reliability）和效度分析（Validity）。信度分析是指測量結果是否具有一致性（Consistency）或穩定性

（Stability）的程度，本研究的各個因素構面之信度 Cronbach's α 係數皆大於 0.7，代表本問卷具有良好的信度，結果如表 2 所示。

表 2：信度分析與效度分析

因素構面	因素命名	特徵值	解釋變異量%	累積解釋變異量%	α 係數
一	「精準摘要」效益評估	0.898 0.829	23.253	23.253	0.862
二	「概略摘要」效益評估	0.880 0.869	21.268	44.521	0.792
三	「相似度摘要」效益評估	0.868 0.787	20.008	64.529	0.803
四	「TF-ISF 摘要」效益評估	0.891 0.728	19.584	84.113	0.738
五	「GBP 摘要」效益評估	0.852 0.796	22.588	43.557	0.783

肆、系統開發與實作

本系統功能之建置，主要依上述各種功能模組進行整合，為了提供便利的實驗管理介面供評估使用，本研究以 C#的程式開發了 Windows Form 程式，此程式不僅擁有各種模組功能，還能依據系統流程展現系統運作情況，以實證系統的完整性與可用性，其實驗結果如圖 2 所示。

伍、實驗評估

一、實驗資料來源

研究中用來產生自動摘要之樣本，由「網頁搜尋與擷取模組」取得，由人工篩選純文字型態的網頁文章，隨便收錄了核四公投（22 篇）、蔡經濠（6 篇）、H7N9 禽流感（38 篇）、人貓橋（19 篇）、Google 眼鏡（40 篇）、板橋愛心待用麵（8 篇）、媽媽嘴咖啡店（23 篇）、南北韓分裂（17 篇）與原萃綠茶（11 篇）等 9 個議題之純文字型態網頁文章，共計 184 篇，利用簡單隨機抽樣從中挑選 10 篇樣本進行評估。

圖 2：整體功能模組展示圖

二、實驗結果分析

本實驗資料分析方式，會因為實驗的目的與方法不同而有所差異，在實驗中有採用 TFISF 方法（陳姿妤&魏世杰 2007）做為與之前單文件摘要的比較。在實驗一部分，系透過召回率、精確率及 F-measure 來進行網頁摘要內容重要性評估；在實驗二部分，則透過李克特量表，依序為非常同意、同意、普通、不同意、非常不同意，進行評估系統摘要內容是否能夠有效表達原文大意以及語句順暢。

（一）實驗一：驗證所提出的自動摘要法之成效

實驗一中利用系統所產生之「TF-ISF 摘要」、「相似度摘要」、「概略摘要」、「精準摘要」與「GBP 摘要」，透過召回率、精確率以及 F-measure 來進行網頁摘要內容重要性評估。實驗中將請評估人員以逗號、句號與問號作為斷句基礎，藉由評估人員勾選所認定的句子與系統所產生的五種自動摘要比對。並假設自動摘要與人工摘要重疊愈高，即表示該自動摘要成效越佳，其實驗結果如表 3 至表 12 所示。

表 3：樣本一實驗結果

樣本一					
	樣本總句數：50				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	19	19	19	19	19
系統摘要句數	13	13	15	11	13
正確的句子數	10	10	11	9	10
召回率	52.63%	52.63%	**57.89%**	47.37%	52.63%
精確率	76.92%	76.92%	73.33%	**81.82%**	76.92%
F-measure	62.50%	62.50%	**64.71%**	60.00%	62.50%

根據表 3 樣本一的實驗結果顯示，在召回率的表現得知「概略摘要」擁有最佳的句子擷取能力。若以精確率的角度來看，「精準摘要」因為是取「TF-ISF 摘要」與「相似度摘要」的交集，因此擷取的句子最少，因而表現最佳。透過 F-measure 來評估系統，可看出「概略摘要」的表現最佳。

表 4：樣本二實驗結果

樣本二					
	樣本總句數：45				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	15	15	15	15	15
系統摘要句數	16	13	17	12	14
正確的句子數	10	7	10	7	8
召回率	**66.67%**	46.67%	**66.67%**	46.67%	53.33%
精確率	**62.50%**	53.85%	58.82%	58.33%	57.14%
F-measure	**64.52%**	50.00%	62.50%	51.85%	55.17%

根據表 4 樣本二的實驗結果，在召回率的表現得知「TF-ISF 摘要」與「概略摘要」擁有最佳的句子擷取能力，表現較佳。以精確率的角度來看「TF-ISF 摘要」的表現最佳。透過 F-measure 來評估系統，可看出「TF-ISF 摘要」的表現最佳。

表 5：樣本三實驗結果

樣本三					
	樣本總句數：57				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	13	13	13	13	13
系統摘要句數	20	15	24	11	14
正確的句子數	8	8	8	8	8
召回率	**61.54%**	46.67%	**61.54%**	**61.54%**	**61.54%**
精確率	40.00%	53.85%	33.33%	**72.73%**	57.14%
F-measure	48.48%	50.00%	43.24%	**66.67%**	59.26%

根據表 5 樣本三的實驗結果，以精確率的角度來看，「精準摘要」的表現最佳。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「精準摘要」的表現最佳。

表 6：樣本四實驗結果

樣本四					
	樣本總句數：31				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	13	13	13	13	13
系統摘要句數	7	10	10	7	10
正確的句子數	5	5	5	5	5
召回率	**38.46%**	**38.46%**	**38.46%**	**38.46%**	**38.46%**
精確率	**71.43%**	50.00%	50.00%	**71.43%**	50.00%
F-measure	**50.00%**	43.48%	43.48%	**50.00%**	43.48%

根據表 6 樣本四的實驗結果顯示，全部方法的召回率均為 38.46%。在精確率的表現得知「TF-ISF 摘要」與「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「TF-ISF 摘要」與「精準摘要」的表現最佳。

表 7：樣本五實驗結果

樣本五					
	樣本總句數：44				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	19	19	19	19	19
系統摘要句數	11	14	14	11	13
正確的句子數	11	11	11	11	11
召回率	**57.89%**	**57.89%**	**57.89%**	**57.89%**	**57.89%**
精確率	**100.00%**	78.57%	78.57%	**100.00%**	84.62%
F-measure	**73.33%**	66.66%	66.66%	**73.33%**	68.75%

根據表 7 樣本五的實驗結果，召回率均為 57.89%。在精確率的表現得知「TF-ISF 摘要」與「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統可看出「TF-ISF 摘要」與「精準摘要」的表現最佳。

表 8：樣本六實驗結果

樣本六					
	樣本總句數：30				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	9	9	9	9	9
系統摘要句數	7	9	11	5	8
正確的句子數	6	6	7	5	6
召回率	66.67%	66.37%	**77.78%**	55.56%	66.67%
精確率	85.71%	66.67%	63.64%	**100.00%**	75.00%
F-measure	**75.00%**	66.52%	70.00%	71.43%	70.59%

根據表 8 樣本六的實驗結果在召回率的表現得知「概略摘要」擁有最佳的句子擷取能力。若以精確率的角度來看，「精準摘要」的表現最佳。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「TF-ISF 摘要」的表現最佳。

表 9：樣本七實驗結果

樣本七					
	樣本總句數：22				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	10	10	10	10	10
系統摘要句數	8	6	9	5	7
正確的句子數	6	4	7	4	5
召回率	60.00%	40.00%	**70.00%**	40.00%	50.00%
精確率	75.00%	66.67%	77.78%	**80.00%**	71.43%
F-measure	66.67%	50.00%	**73.69%**	53.33%	58.82%

根據表 9 樣本七的實驗結果，在召回率的表現得知「概略摘要」擁有最佳的句子擷取能力。若以精確率的角度來看，「精準摘要」的表現最佳。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「概略摘要」的表現最佳。

表 10：樣本八實驗結果

樣本八					
	樣本總句數：34				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	11	11	11	11	11
系統摘要句數	10	9	11	8	10
正確的句子數	7	6	7	6	7
召回率	**63.64%**	54.55%	**63.64%**	54.55%	**63.64%**
精確率	70.00%	66.67%	63.64%	**75.00%**	70.00%
F-measure	**66.67%**	60.00%	63.64%	63.16%	**66.67%**

根據表 10 樣本八的實驗結果，在召回率的表現得知「TF-ISF 摘要」與概略摘擁有最佳的句子擷取能力。若以精確率的角度來看，「精準摘要」的表現最佳。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「TF-ISF 摘要」與「GBP 摘要」的表現最佳。

表 11：樣本九實驗結果

樣本九					
	樣本總句數：36				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	12	12	12	12	12
系統摘要句數	9	9	11	7	9
正確的句子數	7	7	8	6	7
召回率	58.33%	58.33%	**66.67%**	50.00%	58.33%
精確率	77.78%	77.78%	72.73%	**85.71%**	77.78%
F-measure	66.67%	66.67%	**69.57%**	63.16%	66.67%

根據表 11 樣本九的實驗結果，在召回率的表現得知「概略摘要」擁有最佳的句子擷取能力。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。透過 F-measure 來評估系統，可看出「概略摘要」的表現最佳。

表 12：樣本十實驗結果

樣本十					
	樣本總句數：27				
	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
人工摘要句數	14	14	14	14	14
系統摘要句數	8	9	11	6	8
正確的句子數	7	6	8	6	7
召回率	50.00%	42.86%	**57.14%**	42.86%	50.00%
精確率	87.50%	66.67%	72.73%	**100.00%**	87.50%
F-measure	63.64%	52.18%	**64.00%**	60.00%	63.64%

根據表 12 樣本十的實驗結果，在召回率的表現得知「概略摘要」擁有最佳的句子擷取能力。在精確率的表現得知「精準摘要」所擷取的句子精確程度最佳。最後透過 F-measure 來評估系統，可看出「概略摘要」的表現最佳。

就整體而言，五種摘要方法的召回率之平均數均低於精確率之平均數，則說明本研究之自動摘要機制所擷取的重要句子質重於量，其可有效擷取到每篇網頁之重要句子，所擷取的句子皆為正確且是符合使用者所期待的。透過 F-measure 可看出「TF-ISF 摘要」所擷取的句子皆為正確且是最符合使用者所期待。四個摘要

之實驗平均數如表 13 所示。

表 13：四個摘要之實驗平均數

平均數	TF-ISF 摘要	相似度摘要	概略摘要	精準摘要	GBP 摘要
召回率	57.58%	51.93%	**61.77%**	49.49%	55.25%
精確率	74.68%	65.71%	64.46%	**82.50%**	70.75%
F-measure	**63.75%**	57.52%	62.15%	61.29%	61.56%

若單純以 TF-ISF 方法與相似度方法兩者互相比較，「TF-ISF 摘要」在召回率與精確率的表現均優於「相似度摘要」，證實單文件自動摘要結果以 TF-ISF 方法為首選。若同時採用 TF-ISF 方法與相似度方法產生自動摘要可增設「概略摘要」與「精準摘要」其召回率與精確率都有不錯的表現，可藉此增進摘要內容的廣度與深度。而因「GBP 摘要」的計算公式與「相似度摘要」類似，但是「GBP 摘要」有採用本文關聯地圖，因而召回率與精確率均略優於「相似度摘要」。

由於自動摘要的結果會因本文的結構句子而有不同的召回率、精確率與 F-measure。實驗一的整體結果，因概略摘要所得的句子包含正確的句子最多，因而有較高召回率；因精確摘要所得的句子較少，因而有較高精確度；因「TF-ISF 摘要」的平均表現較佳，因而有較高的 F-measure。

（二）驗證所提出的自動摘要法之適用性

為了驗證本研究自動摘要的適用性，藉由線上問卷 mySurvey，進行問卷調查，由受測者主動填寫問卷。本實驗共抽取 30 個網頁以及其自動摘要結果並設計 30 份問卷作為評估。經過一個禮拜的開放時間，共計有 124 為受測者填寫問卷，其中共有 3 份無效問卷。在受測者資料方面，男性為 59 人占 47.58%，女性為 65 人占 52.42%；在學歷的部分高中（職）為 6 人占 4.84%，專科為 10 人占 8.06%，大學為 61 人占 49.19%，研究所以上為 47 人占 37.90%。

本實驗評估是在比較不同的自動摘要方法，將實際地讓使用者針對本研究所產生的五種自動摘要進行評估分為：以上四種摘要那些所表達的內容是否足以代表本文、是否能有效表達原文大意與所摘錄的語句是否順暢，以驗證五種自動摘要法的適用性，其評估結果如表 14 至表 16 所示。

每個問卷由 3~5 位受測者進行評選，並選取有 2 位以上受測者均認同之樣本做為評估的準則，使評估能得到較客觀的結果。根據表 14 評估結果顯示，本研究所產生之「TF-ISF 摘要」、「相似度摘要」、「概略摘要」、「精準摘要」與「GBP 摘要」足以代表本文且普遍符合使用者所期待的。

表 14：2 位以上受測者均認同之樣本數

摘要類型	2 位以上受測者均認同之樣本數（x）	百分比（x/30）
「TF-ISF 摘要」	20	63.33%
「相似度摘要」	17	56.67%
「概略摘要」	18	60.00%
「精準摘要」	18	60.00%
「GBP 摘要」	17	56.67%

表 15：評估四種摘要是否足以代表本文

摘要類型	四種摘要哪些所表達的內容是否足以代表本文
「TF-ISF 摘要」	57.84%
「相似度摘要」	46.80%
「概略摘要」	54.76%
「精準摘要」	45.21%
「GBP 摘要」	48.51%
以上皆不足以代表本文	2.10%

根據表 15 評估四種摘要是否足以代表本文結果顯示，本研究所產生之「TF-ISF 摘要」、「相似度摘要」、「概略摘要」、「精準摘要」與「GBP 摘要」足以代表本文且普遍符合使用者所期待的。

表 16：評估四種摘要是否能有效表達原文大意與所摘錄的語句是否順暢

摘要類型	適用性	有效表達原文大意	語句順暢
「TF-ISF 摘要」	非常同意與同意	79.84%	72.58%
	普通	16.94%	22.58%
	不同意與非常不同意	3.23%	4.84%
「相似度摘要」	非常同意與同意	74.19%	70.16%
	普通	21.77%	25.81%
	不同意與非常不同意	4.03%	4.03%
「概略摘要」	非常同意與同意	68.55%	66.13%
	普通	29.03%	29.03%
	不同意與非常不同意	2.42%	4.84%

摘要類型	適用性	有效表達原文大意	語句順暢
「精準摘要」	非常同意與同意	61.29%	63.71%
	普通	32.26%	35.48%
	不同意與非常不同意	6.45%	0.81%
「GBP 摘要」	非常同意與同意	76.61%	68.55%
	普通	20.16%	26.61%
	不同意與非常不同意	3.23%	4.84%

根據表 16 評估四種摘要是否能有效表達原文大意與所摘錄的語句是否順暢結果顯示。由表 16 評估可知「TF-ISF 摘要」其「摘要能夠有效表達原文大意？」與「摘要所摘錄的語句順暢？」的適用性最高。就整體而言，本研究所產生的「TF-ISF 摘要」、「相似度摘要」、「概略摘要」、「精準摘要」與「GBP 摘要」皆獲得 60%以上的認同，這樣的結果證實所提出自動萃取網頁摘要方法，適用於網頁自動摘要，可藉以挑選網頁中涵蓋較多資訊的句子並自動展現足以表達網頁內文的簡短摘要。避免使用者漫無目的的找尋網頁，節省使用者逐筆進入網頁瀏覽的時間花費。

陸、結論

由於各個搜尋引擎所產生的網頁摘要，大多無法提供使用者充足的摘要內容判斷資訊，更可能造成使用者的誤導。因此要在搜尋到的眾多資訊當中，有效快速地整理出所需資訊，便是一個重要的議題。因此本研究希望搜尋引擎將查詢結果回傳給使用者時，不只是給予一些片斷不全的訊息，取而代之的是一個比較有幫助的摘要，使用者可以藉由此自動摘要，決定是否需要讀取網頁之全文。

本研究運用權重技術針對網頁的內容進行文字探勘，並藉由中研院所開發的中文斷詞系統與文件自動摘要的技術，自動萃取出網頁摘要，藉以提升搜尋的品質。依據本研究所提出的 TF-ISF 與相似度自動摘要方法分別進行摘要實作，並透過 TF-ISF 與相似度自動摘要方法的聯集與交集分別產生「概略摘要」與「精準摘要」。本實驗部分收集網路上之純文字型態的網頁來驗證此自動摘要系統的效益評估，並藉由人工摘要語句做為實驗對照，以探討本研究所提出之自動摘要法的成效與適用性。

本研究針對所提出的自動摘要法，進行客觀效益評估與主觀效益評估二種評估方式之實驗。在主觀效益評估之實驗結果顯示，本研究之各種自動摘要的召回率之平均數均低於精確率之平均數，則說明本研究之自動摘要機制所擷取的重要句子質重於量，其可有效擷取到每篇網頁之重要句子，所擷取的句子皆為正確且是符合使用者所期待的。若單純以 TF-ISF 方法與相似度方法兩者互相比較，

「TF-ISF 摘要」在召回率與精確率的表現均優於「相似度摘要」，證實單文件自動摘要結果以 TF-ISF 方法為首選。若同時採用 TF-ISF 方法與相似度方法產生自動摘要可增設「概略摘要」與「精準摘要」其召回率與精確率都有不錯的表現，可藉此增進摘要內容的廣度與深度。

在主觀效益評估之實驗，在 124 位受測者及 30 個問卷中，僅有 2.10%的受測者認為以上五種摘要不足以代表本文，代表本研究所產生之各種自動摘要足以代表本文且普遍符合使用者所期待的。

在其「摘要能夠有效表達原文大意？」與「摘要所摘錄的語句順暢？」的適用性部份，本研究所產生之各種自動摘要皆獲得 60%以上的認同，這樣的結果證實所提出自動萃取網頁摘要方法，適用於網頁自動摘要，可藉以挑選網頁中涵蓋較多資訊的句子並自動展現足以表達網頁內文的簡短摘要。避免使用者漫無目的的找尋網頁，節省使用者逐筆進入網頁瀏覽的時間花費。由實驗結果可證實本研究所提出之系統方法可以提升搜尋結果的正確性。

誌謝

本研究作者感謝國科會專題研究計劃所提供之部份經費支援補助（NSC 100-2410-H-218-003）。

參考文獻

李俊宏、張興亞（2007），『一個以 Ontology 為基礎的 Web-Mining 技術應用於供應鏈競爭分析之研究』，*電子商務學報*，第九卷，第三期，頁 435-160。

李麗華、李富民、詹尚驥、周裕健（2009），『以學術部落格為主之個人化推薦系統』，*資訊科技國際期刊（IJAIT）*，第 3 卷，Vol. 3，頁 56-75.

柯淑津（2003），『從詞網出發的中文複合名詞的語意表達』，*中文計算語言學期刊*，第八卷，第 2 期，頁 93-107。

陳姿妤、魏世杰 （2007），『運用重複具排除技術於中文文件自動摘要之研究』，*第十八屆國際資訊管理學術研討會論文集（ICIM 2007）*，臺北，臺灣，5 月 26 日。

黃純敏、吳郁瑩（1999），『網路中文文件自動摘要』，*網際網路研討會（TANET99）論文集*，高雄，臺灣，10 月 22 日。

黃純敏、楊存一、邱立豐（2002），『英文網路文件自動摘要之研究』，*第十三屆國際資訊管理學術研討會論文集（ICIM 2002）*，台北，台灣，5 月 20-23 日。

黃純敏、黃世源、盧韋秀（2011），『自動摘要方法於新聞解讀之比較』，*2011 商管與資訊研討會論文集（TBI 2011）*，新北市三峽，臺灣，4 月 28-29 日。

鄒明城、韓慧林、邱景星（2010），『網頁地理資訊檢索與探勘—以民宿主題為例』，*中華民國資訊管理學報*，第十七卷，第三期，頁 19-44。

魏玲玉、曾守正（2006），『以文件倉儲概念實現動態群聚與多重文件摘要之研究－以中文電子新聞為例』，*中華民國資訊管理學報*，第十三卷，第三期，頁 153-173。

Abdel Fattah, M. and Ren F. (2008), 'Probabilistic neural network based text summarization', *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (IEEE 2008)*, Beijing, China, October 19-22, pp. 1-6.

Baxendale, P.B. (1958), 'Machine-made index for technical literature: an experiment', *IBM J. Res. Dev.*, Vol. 2, No. 4, pp. 354-361.

Dalal, M.K. and Zaveri M.A. (2011), 'Heuristics based automatic text summarization of unstructured text', *Proceedings of the International Conference & Workshop on Emerging Trends in Technology (ICWET 2011)*, Mumbai, India, February 25-26.

Das, D. and Martins A.F. (2007), 'A survey on automatic text summarization', *Literature Survey for the Language and Statistics II course at CMU*, Vol. 4, pp. 192-195.

Gupta, V. and Lehal G.S. (2010), 'A survey of text summarization extractive techniques', *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, pp. 258-268.

Harris, A. and Oussalah M. (2008), 'Automatic document summarizer', *Proceedings of the 7th IEEE International Conference on Cybernetic Intelligent Systems (CIS 2008)*, London, UK, September 9-10, pp. 1-6.

Ji, X. (2008), 'Research on the Automatic Summarization Model based on Genetic Algorithm and Mathematical Regression', *Proceedings of the International Symposium on Electronic Commerce and Security (ISECS 2008*, Guangzhou, China, August 3-5, pp. 488-491.

Losiewicz, P., Oard D.W. and Kostoff R.N. (2000), 'Textual data mining to support science and technology management', *Journal of Intelligent Information Systems*, Vol. 15, No. 2, pp. 99-119

Luhn, H.P. (1958), 'The automatic creation of literature abstracts', *IBM Journal of research and development*, Vol.2, No. 2, pp.159-165.

Mani, I. and Maybury M.T. (1999), '*Advances in Automatic Text Summarization*', *Vol. 293*, Cambridge: MIT press.

Ren, F., S. Li, and Kita K. (2001), 'Automatic abstracting important sentences of web articles', *IEEE International Conference on Systems*, *Man*, *and Cybernetics (IEEE*

SMC 2001), Tucson, Arizona, October 7-10, pp. 1705-1710.

Salton, G. and McGill M.J. (1983), '*Introduction to modern information retrieval*', McGraw-HIII Book company.

Salton, G. (1989), '*Automatic text processing*', Addison-Wesley Publishing Company

Salton, G., Singhal A., Mitra M. and Buckley C. (1997), 'Automatic Text Structuring and Summarization', *Information Processing &Management*, pp. 193-207.

Sullivan, D. (2001), 'Document Warehousing and Text Mining', Wiley.

Wei, C.P., Chen L.C., Chen H.Y. and Yang C.S. (2013), 'Mining Suppliers from Online News Documents', *Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2013)*, Jeju Island, Korea, June 18-22.