以口碑為基礎之餐廳推薦機制

洪智力　　林政輝
中原大學 副教授　　中原大學 研究生
chihli@cycu.edu.tw　Benny634@gmail.com

摘要

在資訊快速傳播的時代，人們習慣將生活中所發生的事情撰分享於網路中，當分享的內容敘述到一個產品或者餐廳的評論時，即為電子口碑。而人們找尋餐廳美食時，常常透過網路得到他人的電子口碑，但因為網路上的資訊眾多，必須花許多時間搜尋與過濾口碑資訊，並整理出對於這個餐廳的評價，才開始選擇餐廳。本研究希望提供一個餐廳搜尋機制，讓使用者省去瀏覽網路資訊分析口碑的時間，且較容易得到自己所需求的口碑及餐廳評價。本研究提出較正確取得部落格本文下載方法，利用本文標籤萃取部落格中的本文，並建立餐廳面向的控制字彙，分析文章中對於餐廳面向的評價程度，提供較客觀的評家，將所計算的餐廳權重值做為一個餐廳推薦系統的依據。

關鍵詞：口碑、推薦機制、文字探勘

Abstract

In rapid dissemination of information era, people accustomed to share all of the things that happen in our life in network when our share about the comments on products or restaurants on the internet, then the contents is electron word of mouth(e-WOM). but network information is very messy, people must spend more time to search e-WOM, and filter unnecessary information.

This paper use information extraction of text mining, collection blog word of mouth article from internet, raised by this paper web-tag-text-reative-proportions to determine blog's text tag, then use text tag to extract text of blog, and then collection restaurants properties's word from extracting text. Using text mining, information extraction and information fusion to compare article for controlled vocabulary, further, we get restaurants property's weight from different article, and use the weight to build restaurant's recommend system

Keywords: Word of Mouth, Recommend System, text mining

1. 前言

隨著時代的進步，現代人的生活水準提高，人們漸漸重視休閒、旅遊，使得旅遊美食的需求比起往年增加了許多。當消費者從未前往某餐廳用餐消費時，會憑藉著過去的印象來決定，而印象往往由廣告或者其他朋友曾消費的經驗所組成，其中廣告是餐廳所設計宣傳的，內容會突顯餐廳優點並省略缺點，但消費者希望能了解真正的餐廳資訊，因此習慣詢問親朋好友對於店家的評價，而人與人之間透過口語來溝通或者討論，即為口碑(Arndt, 1967)。

經過研究顯示，口碑長期以來一直是影響消費者購買行為的一個重要因素(Whyte et al., 1954; Godes and Mayzlin, 2004)，當消費者在選擇餐廳用餐時，會有價格高於期望、餐廳服務態不不佳等用餐上的風險，因此消費者在決策的過程中，為了控制消費上的風險，會從許多管道蒐集即將購買產品的口碑資訊(楊緒永,2009)，由此可知口碑資訊對於消費者是相當重要的。

資訊科技快速的進步，更加強網際網路的蓬勃發展，使得口碑傳遞從現今的面對面溝通方式，改變成透過網路當媒介傳送的電子化口碑，電子化口碑突破地理與人脈網絡的限制，將口碑的影響力從親朋好友中推廣到不認識的網友，其影響力更是不可同日而語(Goldsmith, 2006)，如何有效的收集口碑，以減少消費者的瀏覽成本，協助消費者做出快速且正確的判斷以成為一項值得研究的議題(Henning-Thurau et al., 2004; Hung, 2008)。

網路發達，使得網路上充滿了無數的口碑資訊，相對的也產生了許多無用的訊息，當人們在網路中尋找口碑資訊時，會因過濾眾多的資訊而增加判斷過程。在這個以知識為訴求的資訊時代，人們獲取資訊的過程，總是期望在最短的時間內有效的取得所需的資訊(吳志

宏,2004)。近年來，部落格的盛行，人們喜歡將自己的生活經驗紀錄在部落格中，無論是飲食經驗、旅遊經歷及細節都在部落格中分享，因此部落格中所記錄的往往都是較真實的一面，使得這些文章成為人們在尋找評價或口碑時的參考依據。因此人們通常會經由網路中不同的部落格(如：Xuite、無名小站、痞客邦…等)得到附近餐廳的相關訊息，藉由不同部落格提供的文章，經過人們判斷後產生對於餐廳的自我評價。藉此萃取正確的部落格文章減少其他非格主所撰寫的資訊，是增進分析的重要因素，過去的研究中在萃取網頁資訊時，所使用的機制較無法套用於不同部落格，若有機制能自動搜集不同部落格文章，相信能增進部落格文章分析的正確率。

由於部落格及網路上其他口碑的數量相當可觀，使得找尋口碑的方向眾多，造成尋找口碑資訊的人們搜尋時間增加。而目前國內的美食評價網頁(如：愛評網、奇摩生活+)均採用單一的網站提供網友直接對相關產品給予等級數字的評價，此種方式可能容易陷入過於主觀以及廣告式的假口碑操弄，使用者需要過濾不必要的訊息以及錯誤的資訊，才得以瞭解使用者所在意的餐廳因素，另一方面，目前大部分的文字口碑探勘模型，忽略了口碑來源與形式的多樣性，只針對單一訊息來源，擷取相關的口碑資訊，如單一網站的口碑探勘(e.g. Hu et al., 2007; Abbasi et al., 2008)，並未考慮不同網站的不同口碑呈現型式，若有一個關於餐廳的推薦機制，能夠蒐集網路上眾多消費者的口碑經驗，並分析口碑得到一個關於餐廳的客觀評價，提供消費者經過過濾的口碑資訊，讓消費者能較容易了解所分析的類別中餐廳之排序，相信會對消費者在餐廳的選擇上有所幫助。

綜合上述，若能達到下列的需求，在口碑推薦上能給予消費者更方便的機制：

(1)正確且較完整的抓取不同部落格的本文資訊。

(2)從多個部落格的口碑資訊中，整理出餐廳不同面向的好壞程度

因此，本研究動機是提供一個自動化的機制，將網路上蒐集到的口碑資料利用文字探勘、資訊檢索的技術，尋這些資料與餐廳的關連，當使用者搜尋某個感興趣的需求時(如：價錢低、氣氛…等)，提供一個推薦機制，將使用者有興趣的面向由強到弱給予排序。讓使用者可以減少搜尋的時間，直接進入選擇的階段。當使用過一段時間後。

2. 文獻探討

本章節針對餐廳推薦機制，以及口碑與文字探勘領域進行部份的回顧與討論。

2.1 口碑影響力

口碑是建立在使用者經驗的交換，此種訊息交換的目的不在營業推廣，而在於使用者主觀經驗的表達與分享 (Arndt, 1967)，且口碑是一種非正式的資訊傳送、交流，且任何一方皆不為了銷售目的而傳達的行為，和廣告相比對於消費者購買、偏好轉換與推薦行為具有重要的影響(Blackwell et al., 2001)。

口碑為何會如此快速的傳遞，其中最主要的原因是口碑具影響力，口碑具影響力主要有四個因素：可信度高、可雙向溝通、能降低風險、具相關性與完整性；第一個因素，可性度高主要是因為傳統的口碑大多來自消費者所熟識的人，因此消費者會將對傳遞口碑者的信賴投射到口碑上(Wilkie, 1990)，而第二個因素，可雙向溝通意旨：口碑不像電視廣告一樣的單向傳遞，讓消費者能有雙向溝通的效果，第三個因素，口碑能降低風險，消費者能從口碑資訊中蒐集許多曾使用過產品的意見，這些意見能準確而正確的了解這個產品的好壞，因此能降低消費者的購買風險，最後具相關性與完整性，口碑中所傳遞的資訊，與消費者欲購買的產品有較高的相關性，且傳遞的資訊會因為即時對話而相對完整(Silverman, 1997)。

口碑可分為正面口碑與負面口碑，正面的口碑通常附有許多消費者推薦的意願並且帶有說服其他消費者的效果，而負面口碑所包含的資訊往往帶有消費者抱怨與情緒的抒發(楊緒永, 2009)。消費者使用過某項商品或感受過某種服務後，對這樣的產品感到滿意產生好感，經由好東西與好朋友分享的想法，將使用過的經驗與感受傳遞給他人；負面口碑則是消費者使用過某項商品或服務，感到相當不滿意，欲尋找發洩的管道而將情緒轉化為口語，傳達到他人耳中，並勸說他人不要購買此商品或服務。

本研究希望能綜合正負面的口碑，使得消費者得到正確的影響力，讓使用者增加一個較客觀的口碑評價。

2.2 餐廳推薦要素

英文的餐廳(Restaurant)是由法國所傳遞

而來，是由法語的 restaurer 延伸而來，而依照法語的意思來說，餐廳是一個提供營養、恢復精神、休息、飲食的場所(林香君、高儀文, 1999)。

在繁華的現代餐廳種類繁多，但也能依飲食內容分為中式餐廳、西式餐廳、速食餐廳、咖啡簡餐等種類。在不同餐廳的選擇上，消費者往往會前往自己所偏好的餐廳，使得消費者能夠在用餐時達到自己所想像的滿足，因此餐廳推薦要素必須符合使用者需求。麥當勞在全球連鎖餐廳業中是一個相當成功的企業，此公司要求旗下餐廳必須做到 Q(品質)、S(服務)、C(清潔)、V(價值感)四個要求，認為一個餐廳的品質與服務是相當重要的。在過去的研究中，Pettijohn et al.(1997)提出品質、清潔和價值為三個餐廳較重要的屬性，而品質包括餐廳的服務品質與產品品質，清潔指的是餐廳環境的清潔與整潔，消費者對餐廳價值的相關因素，除了價格外，服務、環境、食物味道都是相關因素(Park, 2004;Wall and Berry, 2007)。

透過統計分析的方式也能發現顧客知覺價值是由產品品質、服務品質、價格、顧客觀感等因素構成(林桂田, 2008)。曹瓊文(2000)在研究中提出，服務的提供者想要傳遞的感受除了提供服務的本身外，還希望能給服務的對象產生信任感，並且因為這樣的信任感而產生向他人推薦的行為。在環境的要求上，學者 Baker(1987)根據環境心理的分析，將服務分為三大構面，第一為周圍因素，指的是會影響潛意識中對於被背景環境的認知，第二為設計因素，指在視覺刺激上較明顯影響顧客的設計，因此若有較特別的內部設計可讓顧客留下較正面的知覺，第三為社會因素，係指在餐廳服務的人員包括外貌、行為等。

在連鎖店年鑑(1999)中關於連鎖店顧客來店的因素調查裡，發現商品品質管理、店員服務態度與店內、外環境是消費者最重視的。因此本研究將消費者所關心的餐廳因素統整之後，依據 Park(2004)提出消費者在意的餐廳因素，將美食口碑分為「美味」、「服務」、「價格」、「環境」，四個面向。美味主要分析餐廳在食物味道、食物品質的滿意度，服務主要分析餐廳提供的客戶服務是否合宜，價格主要是分析消費者對於這家餐廳所提供的產品或服務所要求的價格消費者是否能夠接受，環境則是分析餐廳的用餐環境給使用者產生的感受是否合宜。

2.3 推薦機制

推薦機制是一個過濾資訊增加搜尋效率的機制之一，可以分析使用者的瀏覽紀錄建立使用者的喜好，根據使用者的喜好與所關心的興趣，來達到資訊過濾的效果(Resnick and Varian, 1997)。過去的搜尋機制，僅提供使用者相關的資料，但因為資料過於龐大，造成使用者搜尋上的困難以及時間的浪費，因此推薦機制就是解決這樣的問題的方法之一。除此之外，如果網路業者將推薦機制結合到網頁中，記錄使用者瀏覽行為、過去購買的商品來當作推薦預測的參考，也能帶來相當大的商機(吳志宏, 2004)，如:Amazon.com、CDNow.com 等。

Schafer et al.(1999)認為推薦系統除了能夠解決資訊超載的問題外，還能為電子商務資加銷售量，原因有以下三點：

(1) 從瀏覽者變成消費者：很多的消費者往往只是在網路上搜尋想要購買的商品，但最後並沒有購買任何產品，主要的原因是商品太多，無法搜尋到真正想要的商品，而推薦機制可以幫助這些消費者找到他們想要的商品，因此可以提升電子商務的銷售量。

(2) 提高交叉銷售的機會：當消費者購買完商品後，推薦機制可以推薦消費者額外的產品來銷售量，若推薦機制推薦的商品剛好符合消費者的喜好，則也就提高了交叉銷售的機會，例如：當消費者在瀏覽某樣感興趣的商品時，推薦機制可以根據消費者所瀏覽的商品推薦相關的產品給消費者。

(3) 增加消費者的忠誠度：消費者的忠誠度是相當重要的，當消費者瀏覽電子商務網站一段時間後，推薦機制能夠建立消費者的喜好資料庫，透過這樣的資料庫可以給予消費者個人化的推薦系統，當消費者重複瀏覽的時間越長，推薦機制的推薦效果就越好，因此也能提消費者的忠誠度。

2.4 本文資訊萃取

文字探勘(Text Mining)也稱為文字知識發覺(Knowledge Discovery in Text, KDT)，主要是將非結構或半結構的文字資料中，進一步的發現尚未得知的資訊，或者解讀文字中隱含的訊息(Fayyad and Uthurusamy, 1996)，使用的技術有資訊檢索(Information Retrieval, IR)、資訊萃取(Information Extraction, IE)、自然語言處理(Natural Language Processing, NLP)等技術。

過去在網頁本文的資訊萃取中，會遇到一些困難，主要原因是網頁中往往記載著許多不

必要的雜訊例如:很多商業網站會提供他們的服務管道，以及版權和隱私公告，又或者在每個部落格網頁中都有出自於商業目的的廣告(Lin et al., 2002)。因此，對於自動化擷取有用的網頁資訊是一個困擾的挑戰。

在過去的研究中，Hammer et al. (1997) 使用重要的文字特徵，自動化擷取有用的氣象數據，但此研究所運用的方法僅能使用於特殊的氣象網站，利用特定的網頁標籤擷取特定的值，若套用到其他的網頁本文萃取中，則可能產生本文資料萃取不到的困境；在特定網頁中，判斷特定標籤取得所需資料的方法，過去有學者利用 Kylin's Information Extraction 來獲得更多的維基百科(Wikipedia)資訊(Wu et al., 2008)，但這樣的方法就如同抓取氣象網站一樣，只能專門抓取某個網站，而不能套用到每個不同的網站中。也有學者提出使用網頁所呈現出來的區塊空間大小來判別本文所在的位置(Lin and Ho, 2002)，概念上認為本文應佔據較大的版面空間，舉個例子，在圖 2.1 中 C1 為網頁的主題，C2 為文章本文位置，C3 為作者簡介，可以輕易的發現 C2 所包含的區塊大小較其他兩個區塊大，因此判定 C2 的內容為部落格本文。

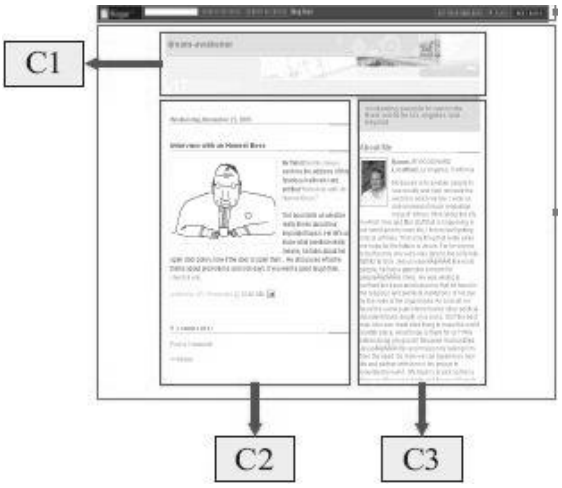

圖 2.1 網頁區塊示意圖
資料來源： Cao et al., 2008

但本文所出現的位置不一定都是佔用網頁最多的區塊，另有學者使用嵌入本體(Ontology)的方式擷取網頁，並利用這個方法收集網頁資訊，也考慮到網頁在本體概念中的強度與本文位置(蔡明原, 2004)，這樣的方式雖然能夠在內容萃取上較為正確，萃取出研究人員所關心的內容，但因為概念強度強的位置不完全為本文位置，因此抓取了不必要的雜訊。

Geng et al.(2007)提出利用 XML(eXtensible Markup Language)中的 DOM 技術，DOM 是一個能夠將 XML 的原始碼掃描過一次後，將 XML 文件儲存成樹狀結構的樣式(http://www.w3.org/TR/2000/WD-DOM-Level-1-20000929)，完成 HTML 的樹狀結構，利用這樣的結構內容去搜尋將要建立成 RSS 所要抓取的值。這樣的方式不但能夠快速的萃取到所需的資訊，並且減少本文位置誤判而萃取到雜訊的機會。本研究將利用 DOM(Geng et al., 2007)的概念，分析每個標籤內所包覆的內容大小，建立網頁標籤文字相對比例法完成部落格本文萃取。

2.5 控制字彙

文字探勘領域中，在分析一篇文章的內容時，會利用關鍵字比對的方式，分析文章中重要且相關的字眼，進而了解文章內容的意涵，而關鍵字建立的方法有：

(1) 關鍵字詞適度字詞頻論(resolving power of significant words)(Luhn,1958)此學者利用字彙出現的頻率經過統計後，將字彙出現頻率較高以及出現頻率較低者去除，因為出現太多次的字彙有可能是連結詞或者語助詞，較無太大意義(如：「的」)，最後留下出現頻率剛好的的字彙，這些字彙多半是較有意義的，將過濾後的字彙建立成關鍵字。字詞的頻率計算公式如下：

$$TF_{ij} = \mathbf{n_j}$$

$\mathbf{n_j}$ ：表示字詞 j 出現的次數

$\mathbf{TF_{ij}}$ ：表示字詞 j 在文章 i 中出現的次數

(2) 利用 TFIDF 也就是 TF 乘上反向文件頻率(inverse document frequency, IDF)得到每個字詞在這些文件中的權重值，取權重值較高的字詞建立關鍵字，利用這樣的權重值來加以運算比對文件(Salton and Gill, 1983)。

IDF 公式如下：

$$IDF_j = \log \frac{\mathbf{N}}{\mathbf{df_{all}}} + \mathbf{1}$$

N ：代表所有文章數。

$\mathbf{df_{all}}$：代表字詞有出現在整個文章集內的文章數目。

$\mathbf{IDF_j}$ ：代表字詞 j 有出現過的文章總數。

(3) 除了自動化的建立外，可加入人為判斷使用控制字彙提高準確度，當自動化建立的關鍵字很多時，可能會蒐集到許多與主題無關的字彙，因此關鍵字並不一定是重要的字彙，利用人為判斷建立控制字彙讓文章分析準確度提

高(Wattenbarger et al., 1977)。控制字彙使用主要是預分析的文件是屬於較科學性的文章(如：醫學類)，或者當文章中的字彙是比較多變的(如：當下流行術語)時，透過 TF 或 TFIDF 可能無法有效正確的抓取有意義的字彙，因此利用控制字彙建立有意義的關鍵字與預分析的文章關聯程度較大的字詞，來解決這樣的問題。Gray et al.(2009)提到使用控字字彙的好處有下列幾點：

A. 能可靠的轉換成一個查詢的字彙概念。

B. 允許使用者使用他們熟悉的字彙和概念。

由於網路上的用詞常常不是辭典上的標準字彙，對於非規律行的詞組，則需要依賴事先建立的詞庫字典或控制字彙(Gray et al., 2009; Binali et al., 2009)，因此本研究採用控制字彙的方式建構關鍵字。

3. 研究方法

本研究採用資訊擷取與控制字彙的技術結合內容式推薦，建構一個個人化餐廳推薦機制。本章將研究方法分為四節，第一節說明本研究的研究架構；第二節敘述如何達到口碑文章蒐集；第三節說明口碑文章分析；第四節解說如何進行評估。

3.1 研究架構

研究架構圖如圖 3.1，主要分為四個部份：

1. 口碑文章蒐集

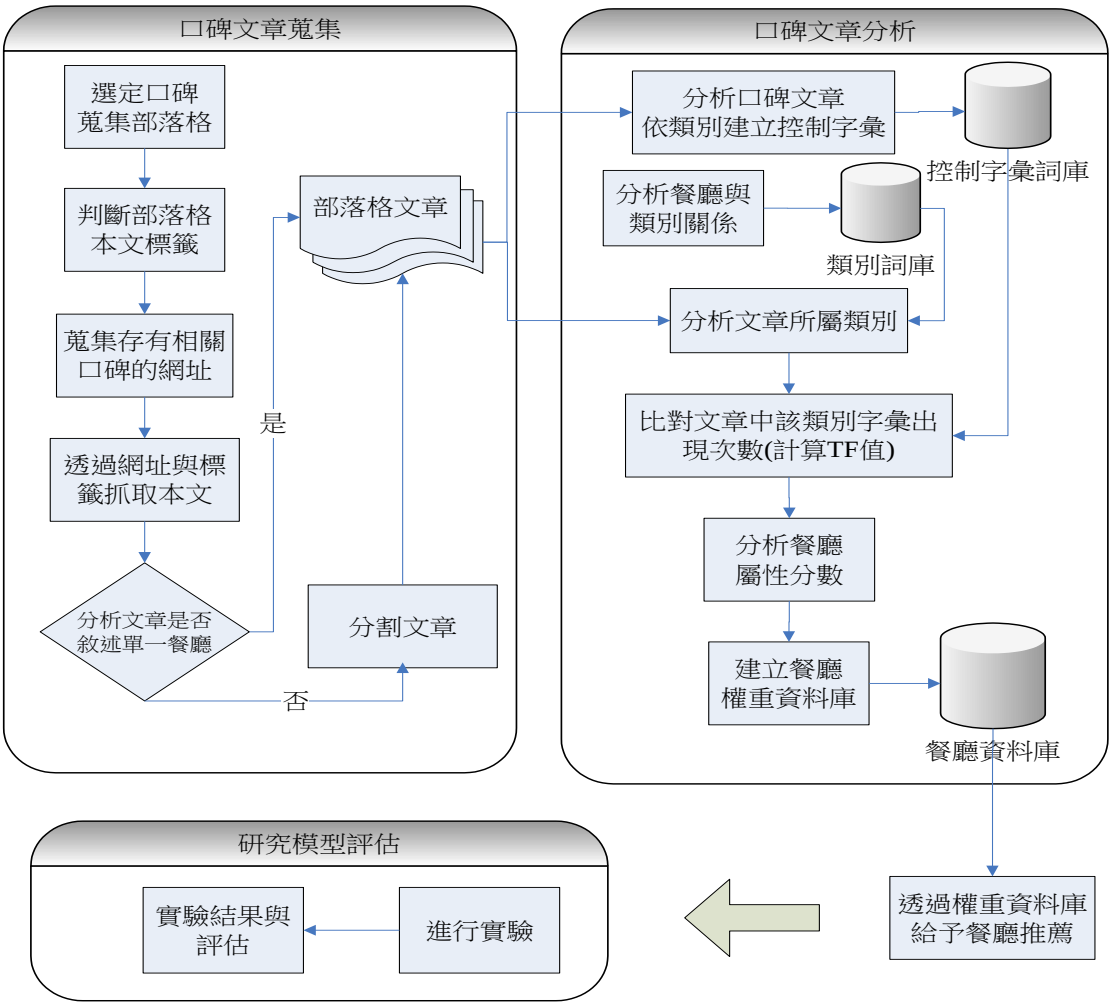

圖 3.1 研究架構

利用 DOM(Document Object Model)的概念提出網頁標籤文字相對比例法，從網路上萃取部落格中所記載之資料。

2. 口碑文章分析

本研究針對 Park(2004)學者提出使用者在意的餐廳因素(價格、服務、環境、美味)，透過這四個因素分析已下載的口碑文章，來建立控制字彙，並分析控制字彙建立餐廳權重。

3. 模型評估

將本研究所提出的機制實做後，發放問卷，給使用過的使用者填寫，藉此評估本研究架構。

3.2 口碑文章蒐集

本節主要將口碑從不同的部落格中萃取，由於部落格能套用許多格式，為了讓格式順利套用於部落格中，部落格提供者利用本文標籤將本文的位置標示出來，因此利用本研究提出的網頁標籤文字相對比例法，判斷部落格記載本文的標籤，透過標籤萃取出部落格本文，完成口碑文章蒐集。

本研究將口碑文章蒐集分為四個步驟：

(1)選定部落格

選定主要資料蒐集的部落格資訊，而創市際市場研究顧問公司調查發現台灣前五大部落格為：無名小站、雅虎奇摩部落格、Hinet Xuite、Yam 天空部落格、痞客邦，因此本研究將以這些部落格為資料蒐集對象。

(2)判斷部落格本文標籤

本研究提出網頁標籤文字相對比例法，將包含字數最多的標籤視為網頁的本文標籤，共分為五的步驟，分別是 a.蒐集部落格原始碼，b.分析 html 標籤，萃取出各個標籤，c.去除每個標籤內多餘的指令(如:Java Script、CSS)，d.

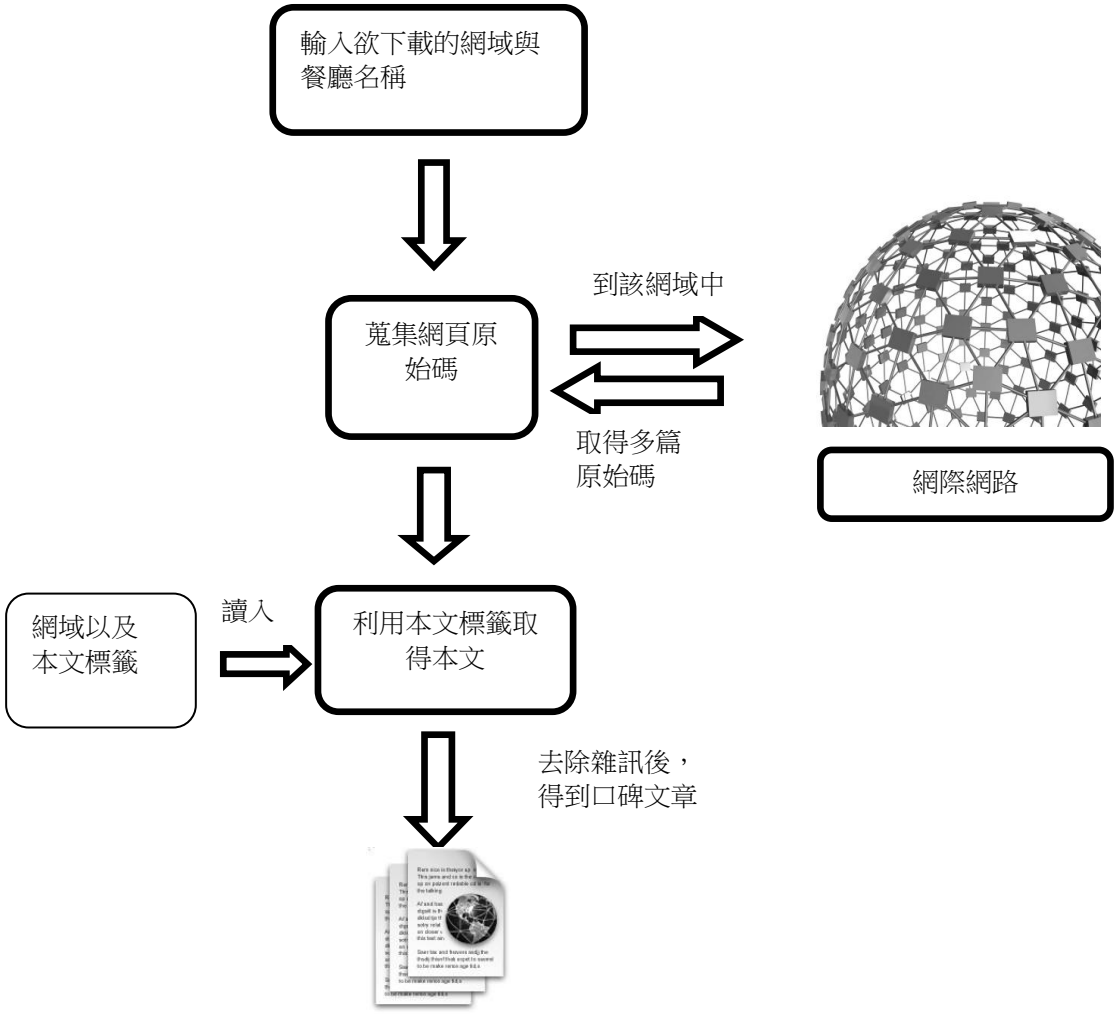

圖 3.2 部落格本文下載過程

分析每個標籤內所包含的字數，e.選定此部落格的本文標籤。

(3)蒐集與餐廳相關的部落格網址：

將欲搜尋的餐廳名稱(如：貴族世家)以及部落格網域(如：www.wreth.cc/blog)提供給GOOGLE 搜尋引擎，透過搜尋引擎取得相關的網址(如圖 3.2)。

(4) 透過網址萃取部落格口碑文章，運作方式如圖 3.2：

當進行文章下載時，先進行網頁原始碼蒐集，接著判斷原始碼中的標籤，並讀取步驟(2)分析部落格本文標籤的判定結果，透過本文標籤將其標籤內容萃取出來後，去除其餘的HTML 標籤，留下沒有 HTML 標籤的本文並儲存。如此一來就能讓部落格中的口碑文章減少雜訊，且能蒐集不同部落格的口碑資料。

(5) 分析文章是否敘述單一餐廳：

若口碑文章中所敘述的為單一餐廳，則能直接進入口碑文章分析階段，但若口碑文章中不僅敘述單一餐廳，必須將文章依餐廳劃分成兩篇文章。本研究透過先前所建立好的餐廳名稱與文章進行比對，當比對到某一餐廳名稱(如：貴族世家)時，會將接下來的敘述認為是同一餐廳名稱的內容，直到下一個餐廳名稱出現為止，並將此篇口碑文章分成兩個餐廳口碑儲存，若口碑文章中只有敘述一個餐廳，則本研究將直接進行下一個階段的分析。

其中步驟(2)判斷部落格本文標籤方式如下，首先將欲搜尋的部落格網域傳送給 Google 搜尋器，Google 搜尋器會依本研究所給予的部落格網域傳回多個部落格網址，為了提高判斷的正確性，因此必須蒐集多個同網域內的網頁原始碼。

完成上述步驟後，分析 HTML 標籤利用 DOM 的概念分別將網頁原始碼依各標籤分割，萃取出每個網頁標籤。接著去除多餘的網頁標籤(如:Java Script、CSS 等)以及多於網頁標籤內的內容(如:CSS 程式碼)，去除多餘內容最主要的目的在於降低判斷過程中的雜訊影響，當網頁標籤內的雜訊未去除時，會將程式碼等不相關的字數判斷成網頁標籤包含的字數，因此使用網頁標籤文字相對比例法時，若有其它非本文文字的的出現時，會增加其標籤所判別出包含字數的多寡，進而影響本文標籤的判斷正確性。

開始判斷本文標籤，首先將每個標籤中所包含的文字計算出來，接著比較每個標籤所包含的字數，將包含字數最多的標籤視為此網頁的本文標籤，公式如下：

$$T = \max_{x \to n}(t_x) \quad \cdots(1)$$

T：部落格本文標籤

n：總標籤數

t_x：第 x 個標籤的字數

例如，原始碼的記載方式如圖 3.3，網頁標籤分別獨立存在，因此能夠容易的判斷出<Tag A>所記載的字數最多，而將此網頁的本文標籤設定為<Tag A>。但在尋找網頁本文標籤中，常常會出現不易判斷的問題，如果網頁的原始碼撰寫方式不是每個標籤單獨存在，而是標籤中又包含了其他標籤(如圖 3.4)，則在判斷過程中會出現標籤內容判斷錯誤的問題，在圖 3.4 的判斷問題一裡雜訊標籤<Tag A>中包含了本文標籤<Tag B>，如此一來<Tag A>的長度將會遠遠大於<Tag B>，因此造成了標籤判斷錯誤，誤將<Tag A>辨別為記載本文的標籤，另一種判斷錯誤的狀況(如圖 3.5)，本文標籤<Tag A>中包含了雜訊標籤<Tag B>，雖然得到的結果會是正確的本文標籤，但在部落格本文抓取時會將<Tag B>內的值一併萃取出來，導致萃取後的文章出現雜訊問題。

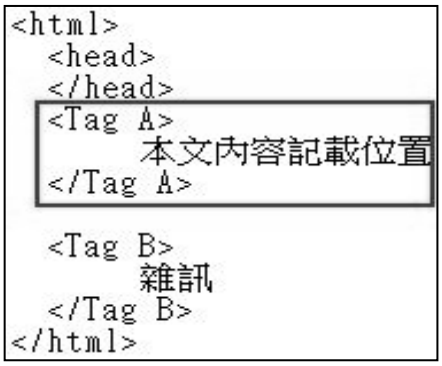

圖 3.3 原始碼撰寫方式

圖 3.4 判斷問題一

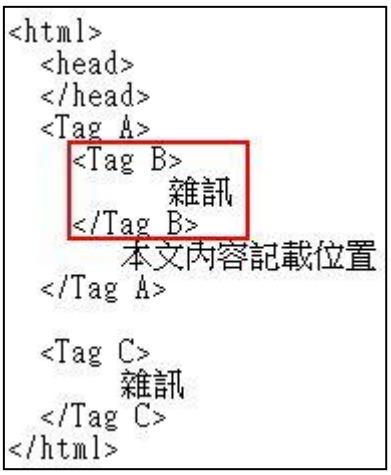

圖 3.5 判斷問題二

本研究將比對每個標籤的內容，當標籤的內容出現重複，則從涵蓋其他標籤的內容中刪除重覆的標籤所包含的內容，舉例來說，利用 DOM 的概念處理完圖 3.5 判斷問題二的原始碼後，其中某兩個節點會呈現如圖 3.6，此時會取出標籤一的內容，並去除包含標籤二的內容部分，再放回原來的標籤一中(如圖 3.7)。經過此步驟後，就能將每個標籤內所涵蓋的內容清楚劃分，除去標籤判斷的模糊問題，以提高判斷本文標籤的正確率。

標籤編號	標籤一	標籤二
標籤內容	<Tag A> <Tag B> 雜訊 </Tag B> 本文內容記載位置 </Tag A>	<Tag B> 雜訊 </Tag B>

圖 3.6 網頁部分標籤內容

標籤編號	標籤一	標籤二
標籤內容	<Tag A> 本文內容記載位置 </Tag A>	<Tag B> 雜訊 </Tag B>

圖 3.7 處理後的網頁標籤內容

當確定每個標籤所包含的字數時，便可比較單一網頁中，那個標籤內所含的字數最多，進而將包含字數最多的標籤定為此網頁的本文標籤。雖然知道了單一網頁的本文標籤為何，但所判斷出包含字數最多的標籤，可能是該網頁回覆字數比本文字數多，造成判斷結果並不一定是正確的本文標籤，因此本研究在判斷單一部落格標籤時，會抓取多篇部落格原始碼，每篇原始碼都透過上述的方式判斷出該篇的本文標籤，接著判斷萃取出來的標籤選出代表該網域的本文標籤，利用標籤出現的頻率計算，頻率計算的公式如下：

$$\text{標籤出現頻率} = \frac{\text{標籤萃取次數}}{\text{總萃取次數}} \times 100\% \quad ...(2)$$

將每個萃取出來的標籤出現頻率計算後，比較每個標籤的出現頻率，接著把頻率最高的標籤判定為本文標籤。

3.3 口碑文章分析

在口碑文章分析部分，本研究分為兩步驟分別為：(1)建立控制字彙與餐廳類別，(2)文章處理。步驟一，必須建立控制字彙，本研究針對 Park(2004)學者提出使用者在意的餐廳因素，將餐廳美食口碑，分為四個面向討論分別為「美味」、「服務」、「價格」、「環境」。由於網路上的用詞往往不是辭典上的標準詞彙，對於非典型的詞組，須依賴事先建立的詞庫字典或控制字彙(Gray et al., 2009; Binali et al., 2009)，因此本研究採用半自動化的方式建立控制字彙，使用 CKIP 中文斷詞(http://ckipsvr.iis.sinica.edu.tw/)，去除無用詞(stop words)後，累計出現頻率較高的詞彙，以協助蒐集描述這四個面向詞彙，並加上愛評網中所分析的餐廳類別(如：鍋類、燒烤類等)，分別蒐集類別中描述這四個面向的詞彙，再結合其他美食雜誌常用來描述食物或餐廳的詞彙，將這些詞彙建立成每個面向的控制字彙，最後再將這些詞彙依正向與負向語意分為八種美食描述詞彙，為每個餐廳類別做出八個詞彙分類(如圖 3.8)。

美味正面	美味負面	服務正面	服務負面	價格正面	價格負面	環境正面	環境負面
好吃	冷掉	全心全意	冷漠	便宜	貴	超讚	舊舊暗暗
特別好吃	單薄	大方	機車	低價位	小貴	裝潢過	路邊
很香	油膩	超讚	白木	免費	爆貴	乾淨	攤子
獨特	粉粉的	態度	白目	無限食用	搶錢	有冷氣	小攤位
回味無窮	水水的	開朗	速度慢	合理	冤大頭	有氣氛	騎樓底下
獨門	寡淡	健談	惡劣	很俗	黑店	裝飾好看	無冷氣
獨門配方	無味	和善	沒禮貌	公道	坑錢	有風格	髒亂
唇齒留香	死鹹	會改進	臭臉	超便宜	很搵	偏遠	吵雜
好好吃	太少	親切	囂張	優惠	搵	明亮	有小強
很好吃	有異物	貼心	兇	實在	漲價	布置	蟑螂
料多	難吃	窩心	火大	特價	價格高	情境式餐廳	黯淡
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

圖 3.8 每個面向的正負語意分類

完成上述詞彙蒐集後，本研究會考慮語意詞彙的強烈程度，透過問卷分析強烈程度，預計受測者為曾使用網路搜尋餐廳口碑者，將採用網路問卷進行資料收集，於部落格與台大批踢踢實業坊(telnet://ptt.cc)之美食版發放網路問卷，實施兩個禮拜的問卷調查，透過反向題來測試是否為有效問卷，預計回收 30 份問卷，將每個控制字彙由不同的受測者填入每個詞彙的強烈程度，接著分析問卷的結果，將每個詞彙的語意強烈程度分成 3 個等級(1 為強烈程度正常，3 為語意最強烈，0 則表示該美食詞彙與該面向無關，如圖 3.9)。

建立餐廳類別，由於本研究將以類別推薦餐廳，因此必須先將餐廳分別分類到不同的類別中，首先利用評價網站(如愛評網，http://www.ipeen.com.tw/rank/)中對於美食選擇的共享標籤，如「鍋類」、「吃到飽」、「燒烤」、「中式料理」、「異國料理」、「小吃」、「速食」、「飲品」、「冰品」等，此共享標籤表是網友搜尋美食的常用字彙。餐廳與類別是多對多的關係，意旨一家餐廳會販售多種類別，一種類別會被多家餐廳販售，本研究將使用類別名稱對應餐廳的方式，自動產生餐廳和類別的關係。

步驟二文章處理，判斷口碑文章中所敘述的內容是評價哪個類別，本研究利用類別詞庫比對分析，一篇口碑文章中，可能敘述著不同類別的評價，例如：同樣在敘述「鍋大爺」的餐廳口碑文章，其中提到火鍋很好吃，但燒烤很難吃。為了處理這樣的問題，本研究先將文章斷句，以句子為單位比對類別詞庫的詞彙，將每個句子依不同的類別儲存成一個單一類別的口碑文章，當句子未比對到詞庫中任何一個詞彙時，本研究便將此句子分配到所有類別中，以完整的呈現每個口碑文章所要表達的評價。完成口碑文章類別分割後，本研究將蒐集到的文章透過不同類別的控制字彙比對分析，在同一面向中，若比對到兩個以上的特徵描述詞彙，則使用長詞優先原則，如口碑文章中出現「特別好吃」，則忽略「好吃」的描述。接著整理出每篇文章在每個面向的控制字彙中出現的次數(如圖 3.10)，也就是詞彙的 TF 值，將每個比對的結果記錄下來。但每個詞彙所代表的語意強烈程度不同，應經過語意加權

	美味正面				美味負面				
	好吃	特別好吃	很香	獨特	冷掉	超難吃	死鹹	油膩	……
美味正面	2	3	1	2	0	0	0	0	……
美味負面	0	0	0	0	-3	-2	-1	-2	……
服務正面	0	0	0	0	0	0	0	0	……
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

圖 3.9 美食語意強烈程度圖

	美味正面				美味負面				
文章	好吃	特別好吃	很香	獨特	冷掉	超難吃	死鹹	油膩	……
文章1	3	4	0	1	0	0	0	0	……
文章2	0	0	0	0	1	2	0	3	……
文章3	0	2	4	2	0	0	0	1	……
文章4	2	4	3	1	0	0	0	0	……
文章5	0	0	0	1	3	4	2	1	……
文章6	4	2	1	1	0	0	0	0	……
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

圖 3.10 文章字詞出現頻率圖

運算才能較為準確的分析每個面向個權重，因此本研究將進行加權運算，將所選擇的美食語意詞彙之強烈程度與文章出現的詞彙作加權計算。

加權運算主要是計算文章所包含的字詞和面向關係字詞權重的方法，將每個文章所提及的詞彙次數乘上該詞彙的權重後，可得到此文章所描述的餐廳產品類別在每個面向的權重。由於文章的長短將影響詞頻值，本研究將以文章出現詞彙的個數平減，即能得到此文章在該面向的權重值，計算出的權重便是本研究作為餐廳推薦的依據，公式如下：

$$X = \frac{\sum TF \cdot T_w}{T_f} \quad \cdots(3)$$

X:此文章在某一面向的權重。

T_f:文章在此面向出現不同詞彙的次數。

TF:文章中 T 詞彙出現次數。

T_w:T 詞彙的權重。

例如現在文章 1 跟美味正面有關的詞彙有四個詞彙為：好吃、特別好吃、獨特，因為很香在文章中沒有出現，所以出現次數為零，而其他詞彙的出現次數分別為：3、4、1， (如表 3.1)，而美味正面的四個詞彙權重分別為：3、1、2、1(如表 3.2)

表 3.1 文章字詞出現頻率

	特別好吃	好吃	很香	獨特
文章 1	3	4	0	1

表 3.2 面向字詞權重

	特別好吃	好吃	很香	獨特
美味正面	3	1	2	1

計算方法如下：

先將每個正向詞彙的權重 3、1、2、1 乘以文章面向 3、4、0、1 後，得出

$$\sum TF \cdot T_w = 3 \times 3 + 4 \times 1 + 0 \times 2 + 1 \times 1 = 14$$

因此可求出，文章 1 在美味正面的權重為 。

得到每個文章在每個面向的正負面權重後，將每個文章與美食語意詞彙運算結果記錄起來。接著開始計算該餐廳在各類別中每一面向的權重值，首先將關於同一餐廳相同類別中面向的正面權重與負面權重相加，此目的為綜合不同消費者對於相同類別中同一面向的評價，相加後得到該餐廳在某一類別中相同面向的權重值，為了避免所蒐集的口碑文章無提及分析中的面向，導致單一面向的權重值相對其他面向低，進而影響分析的結果，因此本研究將加總起來的值標準化，標準化方法是將計算出來的值乘上與該面向相關的文章數再除上總文章數，計算出的權重便是本研究作為餐廳推薦的依據，公式如下：

$$R_{aw} = \frac{\sum ra}{\sum r} \cdot \sum Xa \quad \cdots(4)$$

R_{aw}：餐廳某面向權重。

$\sum ra$：與餐廳和面向相關的文章數。

$\sum r$：餐廳相關文章數。

$\sum Xa$：每個文章在此面向的權重加總。

以圖 3.11 為例，假設文章 1 到文章 6 皆敘述同一餐廳，因此此餐廳在美味這個面向的權重為 0.29664-0.35253+03962=0.34031，因此與此餐廳相關的文章有六篇，而這六篇中有三篇文章與美味有關，因此求出權重：

$$\frac{3}{6} \cdot 0.34031 = 0.17016$$

	美味正面	美味負面	服務正面	服務負面	環境正面	環境負面	價格正面	價格負面
文章1	0.29664	0	0	0	0	0	0	0
文章2	0	-0.35253	0	-0.235	0	0	0	0
文章3	0	0	0.43254	0	0	0	0	0
文章4	0	0	0	0	0.7683	0	0.5242	0
文章5	0.3962	0	0.4792	0	0.2637	0	0	0
文章6	0	0	0	0	0	-0.6237	0	-0.4598
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

圖 3.10 文章權重圖

透過上述方式完成每個餐廳在每個屬性裡面所產生權重值後，即能得到每個餐廳的四個面向權重值，接著就能利用這些權重值當成給予使用者推薦的餐廳依據，權重值高表示這個餐廳在這個屬性內是較占優勢的，反之權重值低表示此餐廳在此屬性較為弱勢。

3.4 評估

評估本研究所提出的方法對於使用者在餐廳的屬性分析上是否有正面幫助，以及評估對於使用者在餐廳搜尋上是否有正面幫助。當使用者使用本機制搜尋後，利用問卷的方式，給予使用者填選是否提升搜尋效率，是否為使用者所想要的結果，以完成問卷評估。

4. 結論

本研究提出部落格本文自動萃取機制，將網路上的部落格口碑萃取出來，利用網路部落格口碑建立控制字彙，透過問卷了解使用者對於控制字彙的影響程度建立字彙的權重，再利用文獻分析消費者注重的餐廳重要面向，加以分析每個餐廳在重要面向的權重值，提出餐廳推薦機制，讓搜尋口碑更有效率。

參考文獻

[1] 林香君、高儀文著，***餐飲實務***，揚智文化事業股份有限公司，1999，(ISBN:957-818-002-0)

[2] 林桂田，顧客知覺價值、顧客滿意與顧客忠誠關係之實證研究-以連鎖餐廳為例，***大葉大學碩士論文***，2008。

[3] 吳志宏，以隱性回饋為基礎的自動化推薦機制，***朝陽科技大學碩士論文***，2004。

[4] 楊緒永，品牌形象、知覺價值、口碑、產品知識與購買意願之研究-以手機為例，***南華大學碩士論文***，2009。

[5] 曹瓊文，專業服務接觸、實體環境線索、性別刻板印象與顧客反應關係之研究-以牙科醫病互動性服務為例，***國立成功大學碩士論文***，2000。

[6] 蔡明原，基於語義概念與網頁特徵之相關網頁擷取，朝陽科技大學碩士論文，2004。

[7] Arndt, J., Role of Product-Related Conversations in the Diffusion a New Product, ***Journal of Marketing Research***, Vol. 41, No. 16, pp. 291-295, 1967.

[8] Abbasi, A., Chen, H., and Salem, A., Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ***ACM Transactions on Information System***, Vol. 26, No. 3, 2008.

[9] Baker, J., The role of environment in marketing services: The consumer perspectives., ***In The Services challenge: Integrating for competitive advantage. John A. Czepiel, Carole A. Congram and James shanahan, eds. Chicagr: American Marketing Association***, pp. 79-84, 1987.

[10] Binali, H., Potdar, V., and Wu, C., A state of the art opinion mining and its application domains, ***ICIT 2009 IEEE International Conference on Industrial Technology***, pp. 1-6, 2009.

[11] Blackwell, R. D., Paul W. M. and James F. E., ***Consumer Behavior, Ninth Edition, Publisher: Ohio, Mike Roche***, 2001.

[12] Cao, D., Liao, X., Xu, H., and Bai, S., Blog Post and Comment Extraction Using Information Quantity of Web Format, ***AIRS***, pp. 298-309, 2008.

[13] Fayyad, U., Uthurusamy, R., Data Mining and Knowledge Discovery in Databases, ***Communications of the ACM***, Vol. 39,No. 11, 24-26, 1996.

[14] Geng, H., Gao, Q., and Pan, J., Extracting Content for News Web Pages based on DOM, ***IJCSNS International Journal of Computer Science and Network Security***, Vol. 7, No. 2, pp. 124-129, 2007.

[15] Godes, D., Mayzlin, D., Using Online Conversations to Study Word-of-Mouth Communication, ***Marketing Science***, Vol. 23, No.4, pp.545-560, 2004.

[16] Goldsmith, R.E., Electronic word-of-mouth, In Mehdi, K.-P. (Ed.), Encyclopedia of E-Commerce, ***E-Government and Mobile Commerce*** , pp. 408-412, 2006.

[17] Gray, A.J.G, Gray, N., and Ounis, L., Searching and Exploring Controlled Vocabularies, ***ACM***, pp. 1-5, 2009.

[18] Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., and Crespo, A., Extracting Semistructured Information from the Web, ***In Proceedings of the Workshop on Management of Semistructured Data***, pp. 8-25, 1997.

[19] Hennig-Thurau, T., Gwinner, K.P., Walsh, G., and Gremler, D.D., Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? ***Journal of Interactive Marketing***, Vol. 18, No. 1, pp. 39-52, 2004.

[20] Hu, N., Liu, L., and Zhang, J., Analyst forecast revision and market sales discovery of online word of mouth. ***Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS' 07)***, 2007.

[21] Hung, C., A personalized word of mouth recommender model, ***Webology***, Vol. 5, No. 3, 2008. http://www.webology.ir/2008/v5n3/toc.html.

[22] Lin, S. H., Ho, J. M., Discovering Informative Content Blocks from Web Documents, ***Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining***, pp. 588-593, 2002.

[23] Luhn, H.P., The Automatic Creation of Literature Abstracts, ***IBM Journal of Research and Development***, Vol. 2, pp. 157-165, 1958.

[24] Park, C., Efficient or enjoyable? Consumer values of eating-out and fast food restaurant consumption in Korea, ***International Journal of Hospitality Management***, Vol. 23, pp. 87-94, 2004.

[25] Park, C., Efficient or enjoyable? Consumer values of eating-out and fast food restaurant consumption in Korea, ***International Journal of Hospitality Management***, Vol. 23, pp. 87-94, 2004.

[26] Pettijohn, L.S., Pettjohn, C.E., Luke, R.H. , An evaluation of fast food restaurant satisfaction: determinants, competitive comparisons and impact on future patronage, ***Journal of Restaurant and Foodservice Marketing***, Vol. 2, No.3, pp. 3-20, 1997.

[27] Resnick, P., Varian, H.R., Recommendation Systems, ***Communication of ACM***, Vol. 40, No.3, pp. 56-58, 1997.

[28] Salton, G. and Gill, M., Introduction to Modern Information Retrieval, ***McGraw-Hill Book Co., New York*** , 1983.

[29] Schafer, J.B., Konstan, J., and Riedl, J., Recommender Systems in E-Commerce, ***Proceedings of the ACM Conference on Electronic Commerce***, 1999.

[30] Silverman, G., How to harness the awesome power of word mouth, ***Directing Marketing***, Vol. 60, No.7, pp. 32-37, 1997.

[31] Wall, E.A., Berry, L.L., The Combined Effects of the Physical Environment and Employee Behavior on Customer Perception of Restaurant Service Quality, ***Cornell Hotel and Restaurant Administration Quarterly***, Vol. 48, No.1, pp. 59-70, 2007.

[32] Wattenbarger, D.W., Bailey, J.A., and Martinez, S.J., Interactive System For Controlled Vocabulary Maintenance, ***ACM***, pp. 79-85, 1977.

[33] Whyte, W.H., Jr., The web of word of mouth, ***Fortune***, Vol. 50, pp. 140-143, 1954.

[34] Wilkie, W.L., Consumer behavior, ***New York Wiley & sons***, 1990.

[35] Wu, F., Hoffman, R. and Weld, D.S., Information Extraction from Wikipedia: Moving Down the Long Tail, ***Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining***, pp. 731-739, 2008.