Google 搜尋引擎的網頁排名因素及其權重的推估

陸承志
元智大學資訊管理研究所
imcjluh@saturn.yzu.edu.tw

廖良珩
元智大學資訊管理研究所
S986217@mail.yzu.edu.tw

摘要

本研究旨在以機器學習方法來找出逼近 Google 搜尋引擎排名的可操作性排序因素以及其權重。所謂可操作性，指的是網站擁有者或者網路行銷業者可以據以來做搜尋引擎最佳化 (Search Engine Optimization, SEO)，亦即適度調整網頁的內或外部品質， 以便在特定關鍵字的搜尋結果中獲得排名的提昇。我們關心的是那些可以從搜尋引擎提供的管理者工具或者客觀的第三方取得公開數據的排序因素，而非所有可能的排序因素。本研究以四類工業產品的關鍵字，蒐集 Google 搜尋結果前 20 筆網頁的實驗結果顯示，在 Keyword in Title， Page Rank，加上 Linkdomain 的這組因素組合，經過基因演算法所求出的權重值加權後，可以得到在所有因素組合中最好的排名 Precision 約 70%；同時在不同關鍵字與多種因素組合下計算出的權重值，一致地呈現 Page Rank 的權重值遠比其他因素來得高。

關鍵字：網頁搜尋、搜尋引擎排序、搜尋引擎優化、排序因素

1. 研究動機與目的

隨著網際網路的普及，當使用者需要找尋資料時，最常使用的工具為搜尋引擎，其中關鍵字 (Keyword-based) 搜尋引擎，例如 Google 和 Yahoo，是查詢資料時最常用的工具 (Internet World Stats, 2008)。許多的研究指出，大部分的使用者通常只會看搜尋結果的第一頁，而且會看超過三頁的比例越來越低 (Spink et al. 2001; Jansen & Spink, 2006; Lorigo et al., 2006; iProspect 2006, 2008)，所以在搜尋結果排名比較前面的網頁往往有較高的流量。對許許多多網站擁有者來說，如何在搜尋結果爭取較好的排名就成為 search engine marketing (SEM) 的主要目的。一般而言，SEM 的操作標的有兩種： 自然搜尋排名 (Natural or Organic Search Ranking) 和付費廣告排名 (Paid Search or Sponsored Links Ranking)。iProspect (2010) 研究指出，結合付費的關鍵字廣告和自然搜尋排名對品牌的提昇最有助益。又根據搜尋引擎業者提供的信息以及一項付費廣告排名的專利文獻 (Davis et al. 2008) 顯示，關鍵字廣告的排序只跟關鍵字的相關性和出價 (bid)， 即每次使用者點選廣告時，廣告主願意付的費用有關。所以只要預

算充足,關鍵字廣告排名就不是問題。所以關鍵字廣告比較適合有目標客群，而且需要短期效果的行銷活動。

相對的，提昇自然搜尋結果排名就比較適合來經營長期效益，但它的困難在於搜尋引擎業者並未公布排序演算法和排序因素，學術界對 Google 的排序演算法的數學基礎已有所討論 (Langville & Meyer 2006), 但這些理論說明仍然不具可操作性。所謂可操作性，指的是網站擁有者或者網路行銷業者可以據以來做搜尋引擎最佳化 (Search Engine Optimization, SEO) , 亦即適度調整網頁的內或外部品質， 以便在特定關鍵字的搜尋結果中獲得排名的提昇。

本研究的目的主要在探討搜尋引擎排序的因素以及其權重。我們關心的是那些可以從搜尋引擎提供的管理者工具或者客觀的第三方取得數據的排序因素， 而非所有可能的排序因素。本研究對 Google 的搜尋結果以機器學習的方法來取的可操作的排序的因素以及其權重；接著， 我們將評估依這些機器學習的結果計算出來的預測排名和搜尋引擎原始排名的接近程度，以驗證這些排序的因素及其權重的可用性。

2. 文獻探討

本節探討與本研究相關的文獻，包含搜尋引擎之使用者行為、Search Engine Optimization (SEO) 、搜尋引擎排名的因素、計算權重值等相關議題。

2.1. 搜尋引擎之使用者行為

Internet World Stats (2008) 指出，當使用者需要某些資訊時，搜尋引擎是他們最常用的工具。Nielsen (2009) 的研究結果也顯示，使用者利用搜尋引擎查詢資料的使用量越來越高，其中以 Google Search 來說，在 2009 年 5-7 月，其每月搜尋量約增加 5%。

在使用者搜尋行為方面， 學術文獻的數據顯示，2001 年只有 29%的使用者只看搜尋結果的第一頁(Spink et al. 2001)，2002 年升高到 73% (Jansen & Spink, 2006) ，2005 年時已來到 96% (Lorigo et al., 2006)。 產業界的調查數據亦顯示相同的趨勢，例如 iProspect Study (2006, 2008) 的調查結果指出，只看第一頁的使用者從 2002 年的 48%，2004 年的 60%，2006 年的 62%，一路提升到 2008 年的 68%；在 2008 年的 68%使用者中，有 27%的使用者只看少數的搜尋結果，41%只看第一頁的搜尋結果。而使用者會看超過三頁以上搜尋結果的比例，則從 2002 年的 19%，2004 年的 13%，2006 年的 10%，一路降為 2009 年的 8%。由此可見，使用者瀏覽的搜尋結果頁面逐年遞減。

再者，幾項針對使用者眼球追蹤的研究也顯示，使用者幾乎是相信搜尋引擎的排序結果，因此會依照 Google 建議的順序來瀏覽網頁，縱使網頁的敘述看起來不怎麼相關 (Joachims et al. 2005, Pan et al. 2007, Guan & Cutrell 2007)。

2.2. Search Engine Optimization

Search Engine Optimization (SEO) 的目的是為了能提升網頁在特定關鍵字搜尋結果的排名，然而每個搜尋引擎對採用的排序因素和演算法皆不同，以 Google 為例，市場上傳說 Google 考慮的因素超過 200 個，其中有些是眾所皆知的，例如 PageRank; 但有些就不清楚。至於這 200 多個因素的權重以及演算法就被 Google 視為最高機密。以下我們就搜尋引擎抓取網頁到計算排序分數的過程以及學術界和網路行銷業界所重視的 SEO 要素做探討。

2.2.1. 搜尋引擎的五項過程

搜尋引擎運作的流程大致分為網頁資料抓取 (Crawling)、建立索引 (Indexing) 、評分 (Scoring) 、使用者搜尋與點選統計 (User Search & Click-through Statistics) 以及 使用者瀏覽統計 (User Browsing Statistics) 等五個步驟。以下我們簡略說明每個步驟，並介紹搜尋引擎業者所提供的相關工具。

- Crawling: 搜尋引擎透過網路蜘蛛 (spider) 或爬蟲 (Crawler) 程式在網路上抓取各網站的資料。通常搜尋引擎會從一個或多個起點開始，由這些起點的連結到眾多網站去抓取資料。目前搜尋引擎都會提供網站站長管理工具，例如 Google Webmaster Tools, Bing Webmaster Center 等，讓網站管理者來提交網站的 URL，以及該站網頁的結構，更新時間，頻率和抓取的權重的 Sitemaps。
- Indexing: 這個步驟將抓取回來的資料進行前處理，計算詞出現的總次數 (Term frequency, TF)，詞出現在不同文件的總次數 (Document Frequency, DF), 挑選特徵詞與建立索引。不同的搜尋引擎有不同的資料處理與索引建立方式，並且會根據統計資料來決定網頁的索引量。
- Scoring: 通常搜尋引擎對網頁的評分包含兩部分：

  Net-Score = Static Page Quality Score + Relevance Score

  其中 static page quality store 指的是搜尋引擎對網頁的整體評價，例如 Google 的 PageRank、Microsoft's Bing 的 Page Score、Alexa 的 AlexaRank 等，都是跟 query 無關；第二部分 relevance score 則是跟 query 相關，會依據網頁是否包含 query 的關鍵字，以及關鍵字出現的次數、位置等眾多因素來計算。
- User Search & Click-through Statistics: 各搜尋引擎的網頁都藏有各類技術來收集與統計關鍵字以及點選資料，這些資料也會形成評比的指標例如，點選率低的高排名網頁其排名會逐漸滑落；同時，高點選率的網頁會逐步提昇排名。
- User Browsing Statistics: 各搜尋引擎透過瀏覽軟體 （例如 Google Chrome）或 Toolbar 軟體來搜集使用者瀏覽網頁的資料，可以得知那些網頁是人氣網站，

或者與特定主題相關的延伸連結有些，因此最後統計結果亦可當成評比的參考指標。

2.2.2. 搜尋引擎排名的因素

搜尋引擎業者對其排序因素、權重與演算法均視為最高商業機密，很少對外透露。以 Google (2008) 正式對外發表的 SEO Starter Guide 來看，其中提及的 Title, Meta Description, Anchor text, Image Alt Structure…等這些因素通常稱為 On-page factors 或者 content-based features。對於 Off-page factors 或者 Query-independent features，例如 PageRank, External links 等，則完全未提及。

在學術研究方面，Zhang & Dimitroff (2005a) 的研究指出，在網頁的 title 和內文同時增加關鍵字出現的次數對提升排名有幫助，而且比單獨在 title 或者內文增加次數的效果更好，但在 title 部分關鍵字重複次數不可超過 4 次。同時，Zhang & Dimitroff (2005b) 的另一項研究也指出，有 metadata 元素的網頁排名會比沒有 metadata 元素的網頁來得好，而在 Metadata 用的 keywords 最好來自網頁的 title 和內文。該研究也認為關鍵字同時出現在 Metadata title, description 和 keyword 三個欄位的效果最好。

Fortunato et al. (2006) 的實驗認為我們可以透過某個網頁的 in-links（即指向該網頁的外部 URLs） 的數量來逼近 Google 的 PageRank；同時 Bifet et al. (2005) 早就提出我們查得到的 PageRank 似乎跟 Google 實際用的不一樣， 而且從 Google 查到的外部連結數偏低 (接近於下限) 等觀點。

微軟公司的研究人員則是在 Google 的 PageRank 之外，再加上 Page 層次的特徵（例如常用關鍵字，總字數，URL），以及網頁的 Popularity (一段時間內的造訪次數)，利用一個以 Neural net 為基礎的 RankNet 演算法，對從 Microsoft 搜尋引擎取得資料做排序，發現這樣的結果更適合使用者的偏好(Burges et al. 2005; Richardson et al. 2006; Agichtein et al. 2006)。

Evans (2007) 從 Google 的搜尋引擎可能用到的因素中選出七個最有影響力的因素， 然後對 50 個經過 SEO 專家特別優化過的網頁做分析，以便得知那些因素是 SEO 業者常用而且有些的技巧，其結果如下：

- Number of pages indexed: 被索引的網頁數量是一個被許多 SEO 業者使用的因素，但成果有限。網頁的品質似乎比數網頁數量來的重要。
- PageRank of a web site: 儘管 PageRank 有明顯的重要性，但不能保證具有特定 PageRank 的網站一定比其他較低 PageRank 的網站排名來得高；只能說有較高 PageRank 的網頁比較可能會排在低 PageRank 網頁的前面。
- Number of in-links: 這是指連結到一個特定網頁的外部連結數目，而不是整個網站。當排名下降時，一個網頁的外部連結數目一定也是下降。

- Domain age: 這是根據 domain 註冊日期來計算，據說 Google 認為較年長的 domain 名稱比較可靠，因此應該要比年輕的 domain 排名更高。但實驗結果顯示，Domain age 只有部分效果。
- DMOZ directory submissions: 被列入 DMOZ 是成功的 SEO 業者慣用的技巧之一，但它的效果一直被質疑。
- Yahoo directory submissions: 由於登錄 Yahoo directory 每年需付 300 元美金的費用，這樣的機制會造成只有少數的 SEOs 才會使用此做為他們的因素。實驗結果也無法證實這個因素的重要性。
- Del.icio.us bookmarks: 許多高排名網站通常會比低排名網站有更多的 del.icio.us 書籤，但這並不意味著在 del.icio.us 的書籤數量必定會影響排名。

在業界方面，根據 SEOmoz[1] 發布的 2009 年資料顯示，各搜尋引擎排名因素的重要性已逐漸變化。2002 年最重要的因素 Raw PageRank/Link juice，在 2009 年變成較不重要的因素；相對的，2002 年最不重要的因素 Trust/Authority，反而在 2009 年變成最重要的因素。

SEOmoz 針對各搜尋引擎使用的數個排名因素做調查，調查結果[2] 顯示前五名最常使用的排名因素如下：

- Anchor text from external links\
- External Link Popularity (quantity/quality of external links)
- Diversity of Link Sources (links from many unique root domains)
- Keyword use anywhere in title
- Trustworthiness of the Domain Based on Link Distance from Trusted Domains (例如：TrustRank, Domain mozTrust... 等)

2.3. Ranking Factors 的資料來源

為了能夠進行實驗，以判定各項排序因素的權重，我們需要取得上述各個排序因素的資料。取得的方式可以利用瀏覽器擴充工具， 例如 SeoQuake 會將搜尋結果中每個因素的數據呈現在 Google Chrome 頁面上，並以橫條 bar 的方式呈現，例如，PR（Google PageRank）、I（Google Index 數量）…等。 另外，我們也可以點擊單一網頁，以清楚地看到更詳細的各筆因子數據。其中比較常用的指標說明如下：

- PR: 當前頁面的 Google PageRank。
- Google I: Google Index，被 Google 收錄的頁面數量。
- LD: Yahoo Linkdomain，指向當前網站的外部連結數量。
- LD2: Yahoo Linkdomain2, 指向當前網頁的頂級域名的外部連結數量。
- a Rank:：Alexa rank。

其他亦有 Link: 指向內外部連結的數量，以及 Density 關鍵字密度等。

[1] SEOmoz, http://www.seomoz.org
[2] http://www.seomoz.org/article/search-ranking-factors

2.4. 權重值的計算

許多的研究指出，搜尋引擎的排序因素必須同時考量 static page quality (例如 PageRank) 和 Content based features (例如關鍵字密度，頻率等)，有的還會加上使用者點選的記錄等，但整體而言，到底搜尋引擎使用那些排序因素，以及這些排序因素的權重可以利用機器學習的方法來逼近。

Burges et al. (2005) 提出一個以 Gradient descent 方法來學習以機率成本為基礎的排序函數，並以 Neural net 建置的 RankNet 系統來實作。實驗結果顯示，以兩層 Neural net 的效果較佳，但作者承認實驗訓練過程太耗時，而且步驟也可以再簡化。Richardson et al. (2006) 提出一個以 RankNet 為基礎的 fRank, 這個系統在 Google 的 PageRank 之外，再加上網頁的 Popularity (一段時間內的造訪次數)，Page 層次的特癥 （例如常用關鍵字，總字數，URL），Anchor text 總數，和網頁整個 domain 平均對外連接數來計算網頁的 static ranking score。實驗結果顯示，在 fRank 在 pairwise accuracy 為 67.3%, 比單用 PageRank 的 56.7% 好很多。所謂的 pairwise accuracy 指的是排序演算法和專家對同一組網頁的排序是一致的百分比。Agichtein et al. (2006) 也是在 RankNet 基礎上，加入使用者的行為資訊來調整排序的分數。他們考量的使用者回饋資訊包括使用者對特定 URL 的點選次數、在特定排序位置上預期點選次數與實際點選次數的差異、以及停留在網頁時間等使用者行為的特徵值。在超過 3000 個詢問的大規模實驗結果顯示，結合使用者行為特徵可提升約 31%的準確度 (accuracy)。

Haveliwala (2003) 提出一個 Topic Sensitive 的 PageRank 演算法，這個演算法會根據使用者下的查詢 keywords，來計算這些 keywords 和每個特定 Topic 的關聯度分數，然後再以線性方式將這些權重分數組合成一個情境相關總分，以做為排序的基礎。實驗結果顯示，使用多個維度的情境總分比使用單一 PageRank 可以得到更精確的排序結果。

Bifet et al. (2005) 對 Google 的排序因素做分析，並試圖推導出一個逼近 Google 評分函數的估計函數，但結論是: 這是件很困難的工作。該研究把評估函數的問題重新歸納為一個有線性決策邊界的二元分類問題，亦即把搜尋結果中的任兩筆網頁當成一組，把排名在前的網頁當成”+”，排名較後的網頁當成”-”，然後訓練 SVM 分類器或 Logistic Regression Model 來分辨每組網頁排名。實驗結果顯示，最佳 Precision 大約為 65%。這個研究採用的特徵有 context, formatting, link, 和 metadata 四大類，使用的特徵數有 22 個，但其中有些非常細微，例如文章的位元數、詞的位元長度、關鍵字是否大寫等。

Craswell et al. (2005) 提出一個簡單的密度分析方法來建置可將 query independent 特徵 (例如內容，Links, 使用狀況) 轉換成相關度權重 (relevance weight) 的模型。該研究對 BM25 資料集的實驗結果顯示，在單一特徵的貢獻度上，PageRank > In-degree > URL length > Click

Distance；但在組合的貢獻度，在 PageRank 之外，只有加上 URL 長度會有些許改善，其他的 in-degree 數量和 Click Distance 幾乎對相關度權重毫無貢獻。(ClickDistance 指的是從一個 authoritative site 例如，大學網站連到目標網站所需經過的 link 個數)。

Bao et al. (2007) 探索如何使用 Social bookmarking 來改善搜尋結果。該研究提出 (1) SocialSimRank 來計算書籤和查詢關鍵字之間的相似度，(2) SocialPageRank 計算網頁的人氣。該研究對從 del.icio.us 取得的資料進行實驗，試驗結果認為這兩個新的資訊對搜尋結果的 MAP (Mean Average Precision) 和 NDCG (Normalized Discounted Cumulative Gain)都有明顯改善。

Beel et al. (2010) 根據三個 Google Scholar 搜尋結果的個案研究，指出在學術搜尋引擎 (Academic Search Engine) 做最佳化的重要因素包含：Citation count，Relevance 和 Author and Publication Name 三個因素。

3. 研究方法

此章節說明本研究系統設計之構想、流程、排序因素的選擇、權重值的計算， 以及系統效能的評估。

3.1 系統架構

本研究的系統架構如圖 1，2 所示，系統流程共分為兩階段：第一階段取得搜尋結果的前 30 筆網頁的 on-page 和 off-page 資料；第二階段則將這些資料匯入基因演算法軟體， 推估出可用排序因素的權重，以及每筆網頁的預測排名。 最後再將預測排名與實際排名做比較， 以求出這個方法的精準度。

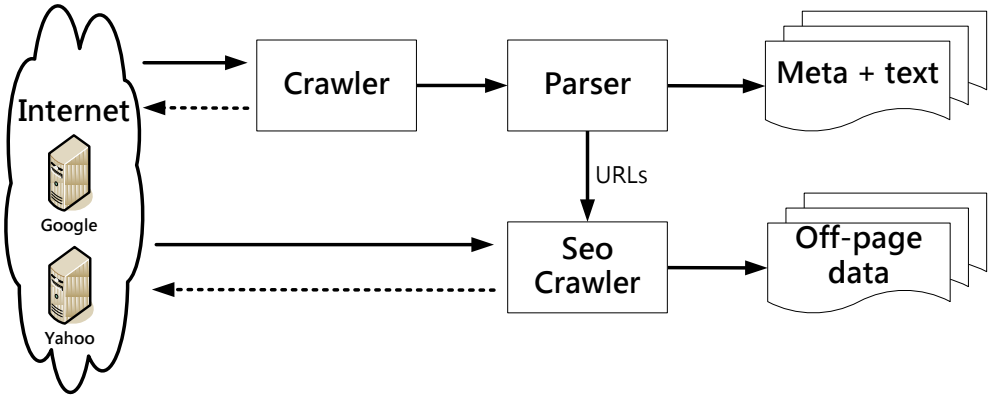

圖 1. Stage1：Automatic Crawling and Parsing of On-page and Off-page data

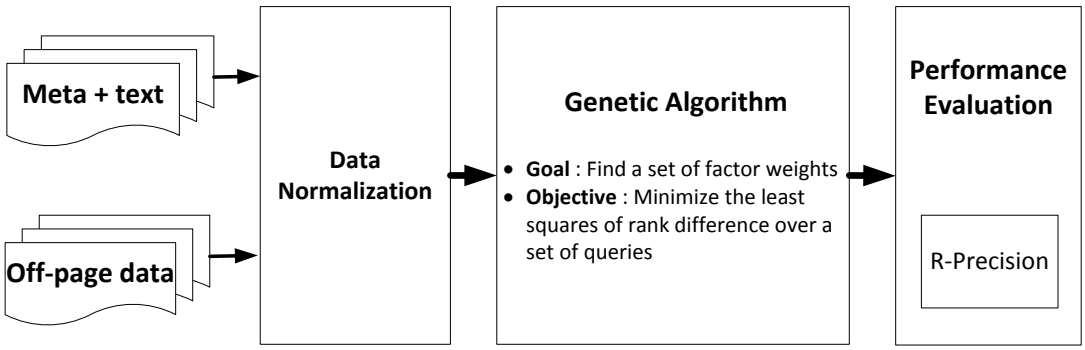

圖 2. Stage2： Factors' Weights Solving and Performance Evaluation

3.2 網頁擷取與網頁剖析

第一階段 Pre-processing 包括搜尋結果網頁擷取、網頁剖析和 SEO 擷取資料三個步驟。本研究以 Google 搜尋引擎為研究對象，輸入數組相關的關鍵字以取得實驗資料。

(1) 網頁擷取： 這個步驟針對搜尋結果中的前 20 筆網頁，**擷取**每筆網頁的 html 原始檔。我們會先檢查搜尋引擎是否提供網頁的暫存檔(cache)，若有就抓取暫存內容，若無才去直接去抓取網頁內容。

(2) 網頁剖析：這一步驟可再細分為兩個小步驟：首先，我們針對搜尋結果頁面中每筆網頁的標題 (Title)、網頁片段 (Snippet)、超連結(URL)進行剖析。接者，再進入各網頁的原始檔，抓取網頁標籤(Meta Tag) 中的關鍵字(Keywords)、描述 (Description)及網頁內文 (Page content)。

(3) **SEO 資料擷取：**這個步驟根據上述步驟剖析出來的網頁 URLs，透過 SEO Crawler 到 Google 和 Yahoo! 提供的管理者工具箱以及其他第三方網站取得相關的數據，有關參數的設定則如同 2.3 節所示。

3.3 排序因素的選擇

本研究主要依據 Evans (2007) 及 SEOmoz 調查結果 (2009) 提到的數個因素，以 On/Off-page 的方式做分類(表 1)。 其中我們把 On-page 和 Off-page 因素再細分為 primary 和 second 兩類，例如 Keyword use in Title 是 primary on-page 因素，它可在細分為 Keyword Density、 Keyword in Domain 等細項。

表 1 排序因素的分類

Factors	Primary	Second
On-Page	Keyword use in Title	Keyword Density Keyword in Domain
Off-Page	PageRank of a web site Number of External links Number of link domains Anchor text in external link	Trustworthiness of the Domain Domain age Del.icio.us bookmarks

在 On-Page factors 部分，有些是根據我們觀察得來的，例如 Keyword in Domain 似乎在 Google 的排序也很重要，例如當我們查 "Hand Trolley" 時，這兩個網站：www.handtrolley.org 和 www.hand-trolley.co.uk 都排在第一頁，但其實他們的 PageRank 都是 0。

在 Off-Page factors 部分，許多的研究都認為 PageRank 是最重要的因素 (Burges et al. ,2005; Bifet et al. ,2005; Agichtein et al.,2006; Richardson et al. ,2006, Evans, 2007)，所以我們將之列為 Primary Off-page 因素。雖然 Craswell et al. (2005) 指出，在 PageRank 之外，加上 in-degree（即外部連結）數量對相關度權重幾乎毫無變化。但 Bifet et al. (2005) 指出，利用 Google toolbar 查得到的 PageRank 似乎跟 Google 實際用的不一樣；Richardson et al. (2006) 認為 PageRank 重視的外部連結數量與品質通常需要長時間來建置，所以我們需要其他資訊來呈現比較即時的外部連結數量。上表中的 Number of External links 和 Number of Link Sources 就是為了這個目的而加上的， 而且我們可從 Yahoo! 的網站管理者工具取得上述資料。

3.4 權重值與網頁分數的計算

本研究將採用的網頁排序評分公式如 Formula 1, 2 ,3 所示：

$$\text{Score}_j = \alpha \times \text{Weight}_{j,\text{OffPage}} + \beta \times \text{Weight}_{j,\text{OnPage}} \quad (1)$$

$$\text{Weight}_{j,\text{OffPage}} = W_{11} \times \text{PageRank}_j + W_{12} \times \text{NumberofLinkDomains}_j + \cdots \quad (2)$$

$$\text{Weight}_{j,\text{OnPage}} = W_{21} \times \text{KeywordInTitle}_j + W_{22} \times \text{KeywordDensity}_j + \cdots \quad (3)$$

亦即網頁的評分由 on-page 和 off-page 權重決定，又 off-page 權重主要由表 1 中的 primary off-page 因素決定；同樣的，on-page 權重主要由表 1 中的 primary on-page 因素決定，至於最終的 w_{ij} 則交由基因演算法來決定。由於不同因素的範圍值域不同，例如 PageRank 是介於 0-10 之間 ，NumberofLinkDomains 可能會高達數千或數萬，因此我們需要把各個因素正規化到 0-1 之間， 那麼最終跑出來的 w_{ij} 才能相互比較，以判斷出各個因素的相對重要性。

在基因演算過程，我們的目標是要找到一組 w_{ij}，使得 the least squares

function (Formula 4) 的數值最小化。

$$S = \sum_{k=1}^{m}\sum_{l=1}^{n}(Actual\_Rank_{kl} - Predicted\_Rank_{kl})^2 \quad (4)$$

其中 m 代表實驗的關鍵字總數；n 為每個關鍵字搜尋結果中的前 n 筆網頁；$Actual\_Rank_{kl}$ 代表在第 k 個關鍵字的搜尋結果中第 l 個網頁的真實排名；$Predicted\_Rank_{kl}$ 代表在第 k 個關鍵字的搜尋結果中第 l 個網頁的預測排名。這個預測排名是在每一個基因演算世代中，當我們計算出每個網頁的 Score 之後，按此 Score 將同一批搜尋結果中的 n 筆網頁做遞減排序，所得到新的網頁排名。當基因演算停止，使得 m 個關鍵字的新舊排名差異平方總和最小的那一組 w_{ij}，就是各個排序因素的權重值。

4. 實驗評估

4.1 實驗的排序因素

本研究從最常用的排序因素中挑出 Keyword in Title (簡稱 KIT) 和 Page Rank (簡稱 PR) 來做實驗用的基本考量因素。由於 Google 的 Page Rank 會經過很長一段時間才更新，因此我們另外考量更新較快的 Yahoo Links (簡稱 L)、 Yahoo Linkdomain (簡稱 LD) 、以及 Yahoo Linkdomain2 (簡稱 LD2)。我們進行實驗的排序因素組合共有下列四種 (1) KIT + PR；(2) KIT + PR + L; (3) KIT + PR + LD; (4) KIT + PR + LD2。其中 Keyword in Title 計算關鍵字在標題 (Title) 中佔的比重，它由三個元素：(1) Keyword presence 、(2) Keyword Proximity 、(3) Keyword prominence 組成，定義如下：

$$\begin{aligned}&\text{Keyword in Title}\\&= \text{Keyword Presence}\\&\times \text{Keyword Proximity}\\&\times \text{Keyword prominence}\end{aligned} \quad (5)$$

- Keyword presence (%) = (Number of words in the title tag that match with each of the words in the search query) / (Total number of words in the search query).
- Keyword Proximity (%) = (Maximum number of words in the title tag that exactly match with part of the search query) / (Total number of words in the search query)
- Keyword prominence (%): = Average of the targeted search terms' prominence in the title tag

$$\text{Term Prominence} = \left(\$\text{TotalWords} - \left(\frac{\$\text{PositionSum}-1}{\$\text{PositionsNum}}\right)\right) \times \left(\frac{100}{\$\text{TotalWords}}\right) \quad (6)$$

4.2 實驗資料集

我們從臺灣出口的產品中挑選 EE (電機) 、 Metal（金屬）、 LED 及 Machine （機械） 四類中，各類都挑選三個關鍵字，然後從 Google 的搜尋結果各抓取 20 筆網頁[3]，並且用人工與程式計算各個數值之後，再進行實驗。實驗用的關鍵字如下表：

[3] Google 搜尋結果抓取日期為 2011/02/26.

表 2 實驗用關鍵字

EE	dc power supply, switching power supply, frame power supply
Metal	CNC lathe machining, die casting molds, stainless steel tube
LED	led down light, led driver circuit, power led driver
Machine	blow molding machine, injection molding machine, plastic injection machine

此外，我們把抓取到的 Page Rank, L, LD，LD2 等數值都做正規化處理。Page Rank 原本的範圍在 0 -10 之間，因此我們只簡單地將 Page Rank 除以 10，將其值正規化到 0-1 之間。其他 L, LD，LD2 三者，我們都是先取 log，再用 min-max normalization 將其範圍值對應到 0 – 1 之間。

4.3 實驗設定

本研究使用基因演算法套裝軟體 Evolver 來求解排序的權重。由於 Evolver 可和 Excel 搭配，因此我們除了把取得的 KIT, PR，L,（或 LD，LD2）等數值輸入至 excel 中，本研究還加入利用求解權重值所計算出來的分數 (Score)、網頁原始排名 (Original Rank)、根據 Score 降冪排序得到的新排名 (New Rank) 以及原始排名和新排名的排名差距 (Rank Difference)。

(1) Evolver 設定範圍: 我們採用 Evolver 提供的兩種解法：Recipe 和 Budget，其設定分別如下：

a. Recipe: KIT，PR，L， LD，LD2 設定範圍：0~1。

b. Budget: KIT，PR，L， LD，LD2 設定範圍：0~1； 而且所有因素權重總和等於 1。

(2) Evolver 求解目標：我們設定 Min Sum (Di) = Rank Difference 平方和，最終目標是找尋 Min Sum (Di) 的最小值。

4.4 效能評估

本研究採用 R-Precision（Manning et al.,2009）來衡量系統在求解之後，重新計算排名後的效能。R-Precision 表示在檢索出第 R 篇文件時的 Precision，本研究設定 R=10，用來衡量在關鍵字 k 的查詢結果中，原來排在前 10 名網頁，在系統求解後仍然排在前 10 名的比例，計算方式如公式 (7)，其中 New Rank 為系統求解後的新排名。

$$\text{R - Precision} = \frac{\{\text{New Rank} \le 10\}}{10} \quad (7)$$

實驗結果如圖 3， 4 所示，其中 Average R-Precision 代表的是每個實驗類別中三個關鍵字的平均 R-Precision。

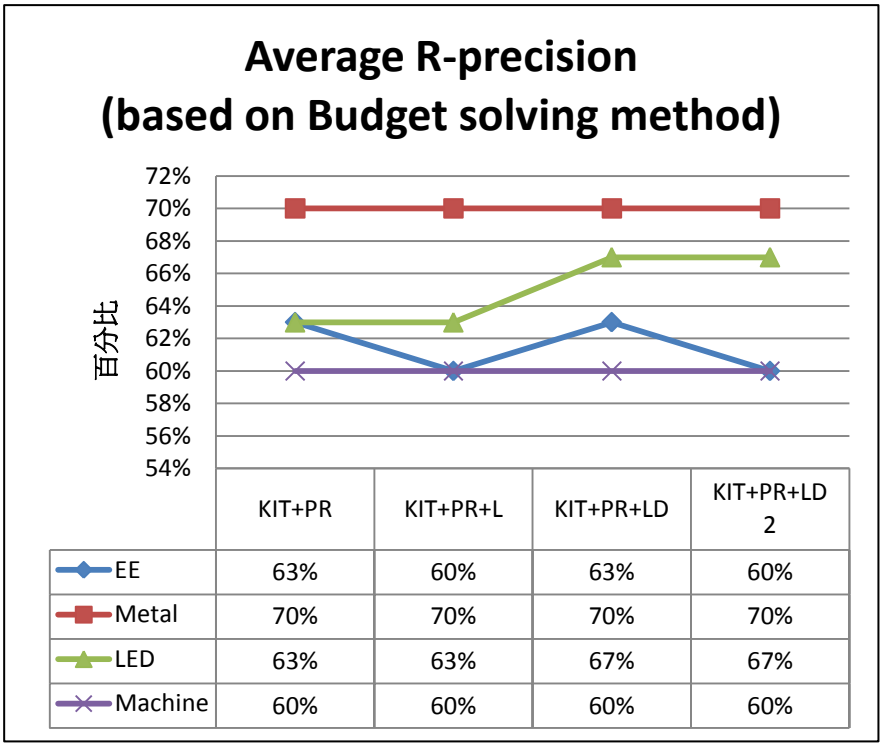

圖 3 不同因素組合的 Average R-precision （Budget solving method）

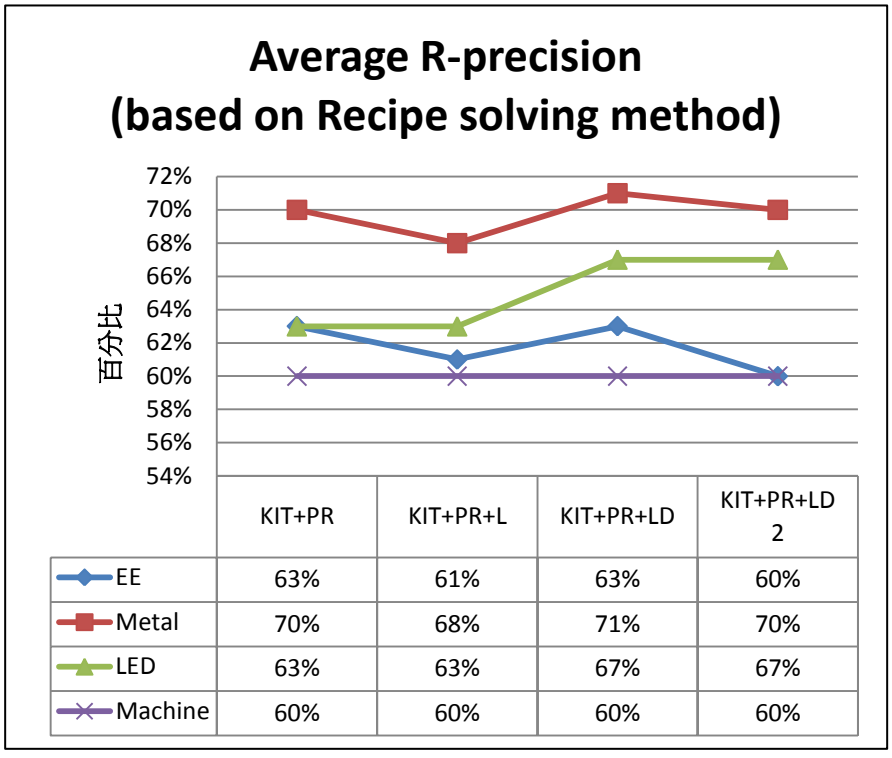

圖 4 不同因素組合的 Average R-precision （Recipe solving method）

從實驗結果，我們發現所有的關鍵字類別，無論是用 Budget 或者 Recipe 解法，Average R-precision 都是在 KIT+PR+LD 這個因素組合下達到最高。不過，比起其他組合的實際差異並不大。此外， 我們也發現所有的實驗結果中 （範例如表 3），無論是何種實驗關鍵字類別以及因素組合，我們得到的因素權重值都是 PR 遠大於 KIT，而且 L, LD，或 LD2 的權重值都很小，亦即 $W_{PR} >> W_{KIT} >> W_{L}, W_{LD}, W_{LD2}$。

表 3. 部分實驗數據 (Metal 類)

Evolver Results									
Keyword in Title + PR					Keyword in Title + PR + L				
K in Title	PR	N/A	Min Sum(Di)	AVE (R-P)	K in Title	PR	L	Min Sum(Di)	AVE (R-P)
0.026	**0.142**	N/A	1201	70%	0.155	**0.831**	0.008	1195	67%
0.072	**0.4**	N/A	1201	70%	0.077	**0.425**	0	1192	70%
0.057	**0.315**	N/A	1201	70%	0.14	**0.776**	0.002	1186	70%
0.102	**0.568**	N/A	1201	70%	0.058	**0.324**	0	1186	70%
0.031	**0.175**	N/A	1201	70%	0.177	**0.966**	0.006	1195	67%
0.08	**0.44**	N/A	1201	70%	0.179	**0.983**	0.005	1195	67%
0.175	**0.975**	N/A	1201	70%	0.112	**0.6**	0.006	1195	67%
0.14	**0.771**	N/A	1201	70%	0.11	**0.596**	0.004	1195	67%
0.165	**0.919**	N/A	1201	70%	0.153	**0.834**	0.006	1195	67%
0.16	**0.896**	N/A	1201	70%	0.175	**0.973**	0.002	1186	70%
Keyword in Title + PR + LD					Keyword in Title + PR + LD2				
K in Title	PR	LD1	Min Sum(Di)	AVE (R-P)	K in Title	PR	LD2	Min Sum(Di)	AVE (R-P)
0.1	**0.757**	0.15	1216	73%	0.146	**0.867**	0.011	1187	70%
0.121	**0.877**	0.181	1216	73%	0.101	**0.563**	0.004	1191	70%
0.111	**0.918**	0.221	1210	73%	0.052	**0.28**	0.002	1191	70%
0.1	**0.749**	0.159	1216	73%	0.181	**0.969**	0.004	1191	70%
0.13	**0.745**	0.004	1195	70%	0.142	**0.794**	0.007	1191	70%
0.08	**0.671**	0.164	1210	73%	0.092	**0.513**	0.004	1191	70%
0.048	**0.445**	0.047	1216	67%	0.18	**0.985**	0.009	1191	70%
0.033	**0.464**	0.078	1209	70%	0.005	**0.026**	0	1191	70%
0.088	**0.96**	0.21	1216	70%	0.101	**0.585**	0.003	1191	70%
0.16	**0.929**	0.024	1202	70%	0.173	**0.951**	0.01	1191	70%

5. 結論

本研究以基因演算法來找出逼近 Google 搜尋引擎排名的可操作性排序因素及其權重。我們以四類工業產品的 12 個關鍵字，蒐集 Google 搜尋結果前 20 筆網頁的實驗結果顯示，在 Keyword in Title， Page Rank，加上 Linkdomain 的這組因素組合，經過基因演算法所求出的權重值加權後，四類關鍵字的 Average R-precision 都比其他因素組合來得好，但差距不大，其中最好的 R-Precision 約 70%；同時在不同關鍵字與多種因素組合下計算出的權重值，一致地呈現 Page Rank 的權重值遠比 Keyword in Title 和其他因素來得高。

目前我們使用的都是可以從搜尋引擎提供的管理者工具或者客觀的第三方取得公開數據的排序因素，未來我們希望能夠加入更多的排序因素，例如 Anchor text in link，keyword density 等，並且進行更多樣化關鍵字的實驗，以便評估排序因素及其權重是否具有普及性或者特殊性。

6. References

[1] Agichtein, E., E. Brill, and S. Dumais (2006), Improving web search ranking by incorporating user behavior information, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 19-26

[2] Bao S., X. Wu, B. Fei, G. Xue, Z. Su, and Y. Yu (2007), Optimizing web search using social annotations, *Proceedings of the 16th international conference on World Wide Web*, pp. 501-510.

[3] Beel, J., B. Gipp, and Erik Wilde (2010), *Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.*, *Journal of Scholarly Publishing*, p. 176-190

[4] Bifet, A., C. Castillo, P. Chirita, and I. Weber (2005), An Analysis of Factors Used in Search Engine Ranking, In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005, pp. 1-10

[5] Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005), Learning to rank using gradient descent, *Proceedings of the International Conference of Machine Learning*, Bonn, Germany, 2005, pp. 89-96.

[6] Craswell, N., S. Robertson, H. Zaragoza, and M. Taylor (2005), Relevance Weighting for Query Independent Evidence, *Proceedings of the 28th annual international ACM SIGIR*, pp. 416-423

[7] Davis, J.D., M. Derer, J. Garcia, L. Greco, T. E. Kurt, J. C. Lee, K. L. Lee, P. Pfarner, and S. Skovran (2008), System and method for influencing a position on a search result list generated by a computer

network search engine, *United States Patent*, No.: 7,363,300 B2, April 22, 2008.

[8] Evans, M. P. (2007), Analysing Google rankings through search engine optimization data, *Internet Research*, Vol. 17 No. 1, p. 21-37

[9] Fortunato S., M. Boguñá, A. Flammini and F. Menczer, (2008), Approximating PageRank from In-Degree, *ALGORITHMS AND MODELS FOR THE WEB-GRAPH, Lecture Notes in Computer Science*, 2008, Volume 4936/2008, 59-71

[10] Google (2008), *Google Search Engine Optimization Starter Guide*, Version 1.1, published 13 November 2008, and available online at *http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf*

[11] Guan, Z. and E. Cutrell (2007), An eye tracking study of the effect of target rank on web search, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 417-420

[12] Haveliwala , T. H. (2003), Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Transactions on Knowledge and Data Engineering*, pp. 784-796

[13] Internet World Stats, (2008), Internet Usage Statistics—The Big Picture, *Available* at *http://www.internetworldstats.com/stats.htm*

[14] iProspect.com, Inc. (2008), *iProspect Blended Search Results Study*, available online at *http://www.iprospect.com/about/researchstudy_2008_blendedsearchresults.htm*

[15] iProspect.com, Inc. (2010), *Real Branding Implications of Digital Media - an SEM, SEO, & Online Display Advertising Study*, available online at *http://www.iprospect.com/about/researchstudy_2010_digitalmedia.htm*

[16] Jansen, B.J. and Spink, A. (2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *Information Processing and Management* (42:1), p. 248-263

[17] Joachims, T., L. Granka, B. Pan, and G. Gay (2005), Accurately interpreting clickthrough data as implicit feedback, *Proceedings of the 28th annual international ACM SIGIR*, August 15–19, 2005, Salvador, Brazil.

[18] Langville, L. N. & C.D. Meyer (2006), *Google's PageRank and Beyond: the Science of Search Engine Rankings*, Princeton University press, New Jersey and Oxford, 2006.

[19] Lawrence, S. (2000), Context inWeb Search, *Internet Computing*, IEEE, p. 25-32

[20] Lawrence, S. and C.L. Giles (1998), Context and page analysis for improved Web search. *Internet Computing*, IEEE, p. 38-46

[21] Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L. and Gay, G. (2006), The influence of task and gender on search and evaluation behavior using Google, *Information Processing and Management* (42:4), p. 1123-1131

[22] Manning, C., F. Raghavan and H. Schütze (2009), *An Introduction to Information Retrieval*, Cambridge Universty Press, Cambridge England, 2009.

[23] Pan, B., Hembrooke, H., and Joachims, T. (2007), In Google We Trust: Users' Decisions on Rank, Position, and Relevance, *Journal of Computer-Mediated Communication*, 12: 801–823. doi: 10.1111/j.1083-6101.2007.00351.x

[24] Nielsen Company (2009), *Nielsen MegaView Search—U.S. Search Share Rankings*, available online at *http://www.nielsen.com/us/en/insights/press-room/2009/*

[25] Spink, A., Jansen, B.J, Blakely, C. and Koshman, S. (2006), A study of results overlap and uniqueness among major Web search engines, *Information Processing and Management* (42:5), pp:1379-1391

[26] Richardson, M., A. Prakash, and E. Brill (2006), Beyond PageRank: machine learning for static ranking, in *Proceedings of the 15th international conference on World Wide Web*. p. 707-715

[27] Zhang, J. & A. Dimitroff (2005a), The impact of webpage content characteristics on webpage visibility in search engine results (Part I), *Information Processing and Management*, pp. 665-690

[28] Zhang, J. & A. Dimitroff (2005b), The impact of metadata implementation on webpage visibility in search engine results (Part II), *Information Processing and Management*, p. 691-715