資訊專題研討課程心得報告

講題:VIPAS: Virtual Link Powered Authority Search in the Web

演講人: 陳銘憲 教授 台灣大學電機工程系

報告撰寫人: 林經緯 政治大學 資科所 一年級 g9232@cs.nccu.edu.tw

2003/11/30

1 [演講主題與摘要]

在龐大的網際網路上能即時地找到符合自己想要的資訊 , 一直是每個網路使用者的夢想 。目前的搜尋技術在搜尋速度與資料量上已有不錯的水準 , 但是在資料搜尋的品質上卻仍有很大的發展空間。

本論文的研究動機在於走出傳統只考慮設計者觀點的搜尋方法 , 加入了"使用者觀點"的因子 , 藉此讓搜尋器能夠找到更貼近使用者需求的資訊。

2 演講要點整理

關鍵字搜尋是目前最常採用的搜尋方式 , 優點是程式寫起來簡單且較直接 , 但是卻有可能會遺漏掉部分不含關鍵字的網頁 , 因爲就語意上而言 , 有些網頁內容是使用者所想要的。

不同於以往分析關鍵字的方法 , google 分析網頁中的連結來研判其 "需求價值" , "需求價值"愈高代表能滿足使用者的期待。如此一來比較不會遺漏掉部分不含關鍵字的網頁 , 而且還可以根據此價值需求來做網頁呈現的排名。

google採用的是 HITS 的分析方式 , 也就是計算出 Hits Score 。一張網頁的 Hits Score = Authority Score + Hub Score 。舉例來說 , 網頁 k 的 Authority Score 指的是將所有連到 k 網頁的其它網頁 Hub 分數全部加總起來 ; Hub Score是指網頁 k 所連出的網頁 Authority Score 的加總 。Hits Score分數愈高者代表越有價值

本文作者認爲 Hits 的計分方式存在著兩個問題 : 第一 , 它只能反應出網頁設計者對其他網頁的"價值感" , 並不能充分代表衆多使用者心中眞正的期待 ; 第二 , 很可能因爲某些因素 , 眞正有"價

值”的網頁不見得能跟其它網站產生互動關係，也就無法在此計分方式下被彰顯出來，例如同行競爭的網站通常不會互設連結。

爲了解決這兩個問題，作者希望在計算網頁的價值時，能夠加入一般使用者的行爲因子，藉由大量統計一般使用者的行爲數據，找到更能貼近使用者所想取得的資料。

在作法上，作者提出了所謂 ”Hot Set” 的概念。在一段期間內被存取量高的網頁連結，會被暫存在稱爲”Hot Set”的資料結構中，而這些連接就是所謂的 ”Virtual link” ，這些 ”Virtual link” 代表的是使用者常點選的連結，而暫存在 Hot Set 的 virtual links 會不停地被更動，每隔一段時間被 skiped 頻率超過 a 的 link 將會被置換出去 (swap out)，或者被點閱的頻率超過 b 的 link 將會被置換進來 (swap in)，參數 a,b 的值是可以根據狀況作 tuning。這樣做是希望搜尋的結果可以更能反應出潮流。

爲了驗證 ”hot set” 的方法是否較 google Hits 的方式來的好，作者利用 Discrepancy coefficient 與 Grouping coefficient，分別計算”hot set” 與 ”google Hits” 的網頁排名，如果計算出來的結果愈小，代表愈能貼近使用者的預期，最後結果證實 ”hot set” 確實較目前現行的”google Hits” 優良。

3 評論與心得

驗證過程中所採用的”使用者預期”的假設標準會不會過於主觀，是我比較懷疑的地方，這部分或許透過大量行爲數據的蒐集，比較可以訂定出一個客觀的標準。

這篇論文雖然將使用者的行爲因子計算進去，但並未能防範 ”大量點選” 軟體的干擾，能否透過一些機制或是其他的計分方式來防範這種行爲，這個部分我想是未來可以再補強的地方。

目前網路上的資料儲存格式五花八門，想要在這些資料堆中精確地搜尋出使用者想要的資料，必須付出相當程度的代價。因此如何統一網路搜尋用”結構化”的資料儲存標準 (如 XML)，我想也是未來在追求搜尋技術時應該要注意到的地方。

另外，目前以關鍵字爲主的搜尋方式，對一般人來說顯得並不夠自然，如何讓搜尋介面更顯得智慧，人性化，個人化 (如語音，視覺等多媒體智慧搜尋)，我想這是未來的發展趨勢.

相關參考資料

[1] 微軟研發XML 搜尋技術 http://www.taiwan.cnet.com/news/ec/0,2000062974,20062001,00.htm

[2] 報導各家搜尋引擎的網站 http://www.searchenginewatch.com/

[3] 報導各家搜尋引擎的網站 http://www.searchengineshowdown.com

[4] 報導各家搜尋引擎的網站 http://websearch.about.com

[5] 各家搜尋引擎比較表 http://www.lis.ntu.edu.tw/ pnhsieh/articles/SE.htm