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: 黃啓峰 政治大學 資科所 一年級 g9216@cs.nccu.edu.tw

2003/12/2

1 演講主題及摘要

這一次的演講主題是關於在網際網路上搜尋我們所 key in 資料時, 使用傳統的 Keyword-based search 方法的缺點及後來的 Link-based search 的介紹, 然後提出一種 algorithm 來決定 search 結果的排列優先權, 企圖把 user 最想得到的資訊擺在第一位, 增加 user 的查詢資料的效率。

2 要點整理

u傳統 keyword-based search 的缺點 以前在網際網路上的搜尋方法是大都是用傳統的 Keyword-based search,keyword-based search 是指在搜尋引擎上輸入我們所要找尋的關鍵字,engine 會根據所輸入的關鍵字去從資料庫中去尋找有相關資料的 page, 當把相符的結果傳回來時, 通常會有一大筆的搜尋結果, 有時甚至會有數千筆符合條件的網頁被找到, 而 user 必須從衆多的網頁當中找尋所想要的資料, 這相當的費心費力, 是可以用 and/or 增加 key word 來增加搜尋的正確性, 但這不是最好的方法, 所以如何把搜尋結果加以整理並排列, 而讓 user 在搜尋結果的前幾筆就能找到所想要得到的資料, 這是比較重要的地方。除了使用 keyword-based search 之外還會加入搜尋同義字詞的功能來找出那些沒有包含特定的關鍵字, 但是卻包含關鍵字同義字詞的網頁內容。例如陳教授有舉例:Netscape 在它的網頁中可能沒有 web browser 這些關鍵字, 如果在搜尋 web browser 的結果中沒有 Netscape, 但我們知道它的確是 browser 的一種, 這樣的搜尋結果似乎不太正確。所以就有了 HITS 演算法的出現。uHITS演算法的優缺點 uAuthority pages 照字義上來說就是較具有權威性的網頁, 權威性的意思表示這樣的網頁是比較容易被其他的網頁所指到, 指到的意思是在其

他的網頁上有其 Hyper link, 所以一個被其他網頁的超連結所指到的網頁, 稱爲 Authority Page。uHub pages 像是一個 Hub 一樣, 有許多 Hyper link 指到別人的網頁稱做 Hub Page。一個被很多 Hub Page 指到的 Authority Page 是一個等級比較高的 Authority Page; 而一個指向很多 Authority Page 的 Hub Page 則視爲是一個比較有價值的 Hub Page, 所以 HITS 的目標是要找出一組具有等級較高的 Authority Page。所以使用了 Iterative algorithm 來計算出等級較高的 Authority Page: (1) 用 keyword-based query 得到一個 web pages 的集合, 然後 expand 它形成一個 base set。(2) assign base set 中的每一個 page 一個初始值爲1的 authority 跟 hub value。(3) 根據它的 link,update 每一個 page 的 score。(4) 將所得到的 score 做正規化 (5) 重覆 (3)(4) 步驟直到 score 收斂爲止。使用這個演算法便能依據網頁之間的超連結來評斷出哪個網頁是比較好的網頁, 以網頁之間的連結比重解決了單單使用 keyword 的網頁排序問題, 讓搜索引擎找出來的結果可以將比較重要的網頁排在搜尋結果頁面的前面, 然而,HITS 演算法還是有需要改進的地方。(1) 若以網頁上的連結來判斷網頁的重要性所反映出來的是網頁設計師的想法並不一定是 user 的需求。(2) 在有些相關的網頁中, 它們並不會有指到對方的連結, 陳教授舉了一個麥當勞跟肯德基這個很好玩的例子, 在互相競爭的公司的網頁上就不會有競爭對手的網頁連結。uVirtual Links 因此若只使用連結來分析網頁的重要性, 會錯失掉某些重要相關網頁的資訊。因此我們提出了藉由加入使用者習慣的 Virtual Links 這個排序方法,Virtual Links 的觀念是認爲在一段時間當中, 有許多 user 瀏覽過的網頁應該是 user 比較感興趣的網頁, 因此這些較常被瀏覽的網頁應該要被加上比較多的分數。把這些網頁收集起來稱爲 Hot Set。在原有的 HITS 演算法當中無法改變 Hot Set 中網頁連結的數量和積分, 因此我們就針對在集合當中的網頁 create 虛擬的 Virtual Links 並使字 HITS 演算法來計算這些網頁的分數。所以我們就比須先紀錄使用者在網路上瀏覽的行爲, 再依 user 所點選過的網頁, 在這些網頁上 create Virtual Links,create Virtual Links 有兩種方式:

(1) 第一種是由搜尋引擎所產生的 Virtual Hub, 可以想像成是搜尋結果的那個網頁, 利用這個 Virtual Hub 來計算各個網頁積分。(2) 另一種是選出現有分數最高的前 n 個 Hub Pages, 在計算分數時在這些 Page 上加上 Virtual Links, 藉以幫這些 Hot Set Pages 加分。此外, 另外設計一個 data warehouse 來幫助確認這些權威性高的網頁。uAlgorithm VIPAS (Virtual Linked Powered Authority Search) uInitial Phase 在 Query 的階段, 先用正常的 HITS algorithm 來對 Query string 作分析, 計算出 base set 中 pages 的 Authority 和 Hub score 後, 收集起來存到資料庫去。uVirtual Link Collection Phase 監控 User 的行爲, 看看在 page 中是哪些 URL 被點選了, 在一段時間的觀察後, 收集那些常被點選的 pages, 把這些 URL 加到 Hot Set 裡面, 在 pages 和 Hot Set 之間 create Virtual Links。uRefinement Phase 對在 Hot Set 中的每一

個 page 重新計算 Authority 和 Hub Scores, 在 Base Set 中重複執行更新 page score 的動作。u2 flavors nVIPAS-VH(VIPAS with virtual links from a Virtual Hub)nVIPAS-TH(VIPAS with virtual links from Top Hubs) uResults 如此便可以利用 Virtual Link 來達到在網路搜尋上考慮 user behavior 排序的目的, 以提供更好的搜尋結果給 User。

3 評論與心得

在這次的演講中, 從教授的研究主題讓我了解到整個搜尋引擎的架構及其所加入了 user 使用行爲後, 可以大幅增加 user 找尋資料的便利性, 平常我們只是很單純的想找尋資料, 並沒有想到原來還有那麼多的因素會影響到搜尋的結果, 所以搜尋結果的排序就變得相當的重要, 在現在這個時間就是金錢的時代, 如何在最短的時間獲取最豐富的資源, 就可以決定一個人的能力如何, 若搜尋引擎能幫助 user 提高生產力, 那就是一個好的 search engine, 所以搜尋結果的排序是最重要的一個 point, 在教授的實驗中, 把搜尋結果的排序做得好需要考慮到 user 的使用行爲, 我覺得這眞的是一個很棒的 idea, 以往只是從 web programmer 的角度來看, 而忽略了 user 眞正的需求, 現在把這加進去後, 讓整個系統更人性化了, 也提昇了整個社會的經濟效益及生產力。教授有提到 google 之所以能在 search engine 中排行較前面的原因是因爲他們有考慮到人爲的因素, 有時一個關鍵字就有好幾千筆的搜尋結果被產生,google 把這些結果做較完善的排序, 讓 user 在前面的 page 中就可以得到想要的資料, 然而教授也提到要做到這樣, 必須要有大量的 pc 來計算如此龐大的資料量, 來提供 user 較好的服務品質, 但事實上, 我認爲我們不太可能以一套單一或是有效的方法來針對使用者的需求和習慣對搜尋的結果做最佳化, 原因在於使用者的數量太多, 而且使用者和使用者之間可能彼此的差異性很大, 要以一套有效的演算法來處理這些因素, 而要求所得出來的結果要能夠盡量的符合每一個使用者的需求, 我認爲是難以達成的, 很多時候在這方面所做的努力也許改善了其缺點, 但是在另一方面有可能會造成相對的負面影響, 所以我覺得能做到大多數 user 所期望的結果, 那我想這就應該算是最大的成果了。在這次的演講中最大的收獲是得到一些新的觀念, 包括搜尋的演算法, 怎樣去改善原來的機制, 還有學到從 user 的需求角度去衡量如何讓 user 使用起來更加方便, 這些在往後的設計專題或論文上相信會讓我受益良多。

4 相關資料

(1) Web reference Search Engine Tutorial http://webreference.com/content/search/how.html

Overview of how search engines work, with some tips for designers.

(2) Web Ranking Tools http://www.bruceclay.com/web$_r ank.htm$

Bruce Clay provides a page outlining his optimization strategies.