

教育部獎助國立清華大學選送學生赴國外頂尖大學（機構）研修學生

成果報告書

Query Log Query Engine

獲獎生姓名：楊惠婷

出國地點：United States

研修大學（機構）：University of Illinois at Urbana-Champaign

研修系所（單位）：Department of Computer Science

出國期間：2007/09/01 ~ 2008/02/22

報告日期：2008/03/26

摘要

在 2007 年 9 月我到美國伊利諾大學香檳校區（University of Illinois at Urbana-Champaign）進行了六個月的短期研究。在那裡我加入了張振川老師的研究團隊，專研於分析網路搜尋日誌的相關研究，並建造一個搜尋日誌的查詢系統，讓有興趣的人都可以查詢分析的結果。而這樣一個系統要支援的查詢理當相當複雜也相當多樣，因此我們決定要先從支援簡單的查詢－『找出一組物件的“共同概念（Common Aspects）”跟“相異概念（Differences）”』開始，也就是在資料量龐大的查詢日誌中找出使用者針對某一組物件進行搜尋時，有什麼概念是會共同使用的，而又有什麼概念是只專屬於某一特定物件的。這樣的查詢結果有許多應用範圍，如商業行銷，或讓搜尋引擎拿來作查詢推薦，還有更多可能的應用等著我們繼續去發掘。

除了專注於研究之外，我還抽空旁聽了一些課程，並在張老師的公司實作了一段時間。旁聽課程、實作作業讓我親自體會到美國的自由學風，對於以往自己的學習缺點做了省思；參與張老師的公司運作更是難能可貴的經驗，我累積了一些業界實作經驗，也讓我有許多國際合作的機會。

此報告書將簡短的描述我在美國的研究成果，還有我的研修心得與建議，以此記錄我過去六個月的所見所得。

目 次

摘要	i
研修目的	1
研修方法及過程	3
研修心得	5
建議	7

研修目的

隨著網際網路的普及與流行，網路世界的網頁以驚人的速度正快速地成長，曾有研究報告指出網頁數量正以每年 145%的比例成長，這種資訊爆炸的現況會讓使用者會越來越不容易找到特定的資料，因此搜尋引擎的重要度日益增加，根據研究估算每天大約有七百億筆的搜尋，而搜尋數量同樣驚人的以每年 131%的比例成長。如此驚人的成長速度造成傳統的搜尋方式越來越不足以滿足使用者，許多更新更先進的技術在陸續的開發，以提供更好的搜尋結果給使用者。

在眾多先進的技術中，分析並瞭解使用者在網路上的搜尋習慣，以此掌握使用者查詢的喜好及意向，是目前許多研究著重的方向。瞭解使用者習慣及意向可以促進許多先進搜尋技術的發展，如個人化的搜尋排序就是針對每個個別使用者搜尋習慣及喜好來改變搜尋排序，以增進搜尋品質；或分析大部份使用者使用關鍵字之習慣，推薦使用者相關的搜尋關鍵字，以此幫助不擅於下查詢的使用者能更精準的使用關鍵字，快速的找到資料，減少使用錯誤關鍵字而導致搜尋失敗的機會。因此，分析使用者搜尋意向及習慣是改善搜尋技術重要的一環。

如何瞭解使用者在網路上的搜尋意向及習慣有很多種方式，過去的研究主要可以分為兩類。早期的研究是請使用者明確的提供個人檔案，以個人檔案代表使用者的習慣與喜好，這方式的優點在於能較精準的掌握喜用者的真實喜好，而缺點就在於使用者的喜好及意向隨時都有可能改變，要使用者自己時時刻刻更新有實施上的困難，再加上所有資料都必須使用者自己手動輸入，增加使用負擔，大部分的使用者並不會願意配合。而近幾年的研究則是以搜尋引擎自動儲存使用者所有搜尋動作的查詢日誌（Query log）和點擊日誌（Click log）為主，使用者不須付出維護資訊的負擔，更可以及時掌握使用者最新的搜尋行為，是目前研究方向的主流。

為了刺激這方面的研究進步，美國微軟公司（Microsoft）在 2006 年特別釋出該公司的搜尋日誌供學術界使用，這組資料是由微軟 2006 年 5 月當月的查詢日誌和點擊日誌隨機抽樣而來的，可以反應真實使用者的查詢與點擊習慣，許多相關研究也在這份資料的釋出後陸續發展。

本研究的主要目的是想要分析一般大眾在網路上搜尋和使用關鍵字的習慣，從中找出共同的規則，以此幫助未來搜尋的相關技術，並為這些分析而得的結果建立一個查詢系統，供人查詢。如：企業可以查詢一般使用者最常查詢該公司的哪種產品，哪些產品最少被詢問，如此一來可以決定更好的廣告策略，達到最大的收益；搜尋引擎則可以藉由此系統查詢針對某特定查詢使用者的點擊習性，以此增進排序品質。因此，此查詢系統一旦建立，受惠的將不只是學術界，而是對全部有興趣瞭解使用者網路搜尋習慣的人都將有所幫助。

研修方法及過程

我們的目的是要建構一個查詢日誌的查詢系統，讓大家可以查詢我們的分析結果。而這樣一個系統所要支援的查詢理當相當複雜也相當多樣，如：目前最熱門的查詢是什麼？大家比較常查詢 gmail 還是 YAHOO mail？這些都可以讓我們瞭解使用者的查詢習慣，更可以拿來幫助改善搜尋的排序，提升搜尋系統的品質。

然而，因為可以支援的查詢太多樣，所以我們決定要先從支援簡單的查詢開始，而目前選定的查詢就是『找出一組物件的“共同概念（Common Aspects）”跟“相異概念（Differences）”』，也就是在資料量龐大的查詢日誌中找出使用者針對某一組物件進行搜尋時，有什麼概念是他們共同擁有的，而又有什麼不同的概念是專屬於某個特定物件的。舉例來說：Chicago 跟 Boston 的共同概念是“restaurant（餐廳）”、“mlb（職棒大聯盟）”、“travel（旅遊）”...等等，找出這些共同概念可以幫助我們延伸出其他同類的物件，像是 Denver、New York...，因為他們的共同概念也是餐廳、旅遊；而相異概念則是“cubs（職棒小熊隊）”、“red sox（職棒紅襪隊）”...等，這些相異概念可以看成專屬某物件的特徵概念。支援這樣的查詢可以幫助企業公司瞭解自己與其他企業公司的差別在哪，什麼是自己的特徵產品？也用於產生類別標誌（Class label）的用途上，有許多實際的用途。

實作這個查詢首先要遇到的問題就是何謂一個“概念”？概念可以是一個單字，如：travel、restaurant...；概念也可以不只一個字，如：red sox、new york...；概念更可以是一個片語，如：take into account；要準確定義出一個語言的單位概念，並精準的將一個句子切成這樣的單位是一件難事，有許多過去的研究在探討這個問題，有人提出統計式的語言分析法、也有人用內建的字典查詢法切割，而在本研究裡則是先將一個單字定義唯一一個概念，以此來實作。

將句子分割為一個個的概念之後，接下來就是如何找出共同和相異的概念。在這裡我們引用了自然語言中的語言模型（Language Model），將和每個物件一起出現的概念先計算好並存放在語言模型中，下面以 Chicago 跟 Boston 為例：

Chicago

概念	機率
restaurant	0.01
mlb	0.002
travel	0.001
cubs	0.001
...	...

Boston

概念	機率
restaurant	0.04
mlb	0.01
travel	0.009
sox	0.005
...	...

語言模型就是把各個概念和出現的機率儲存為一個字典式的表格以供查詢，如此一來當查詢來時，只需比較兩個語言模型的相同和差異即可得到結果。

在實作的過程中我們遇到了一些困難，像是如何快速的回傳結果，目前採取的方式是先將整個查詢日誌中的語言模型都先計算好並儲存起來，如此就可以縮短計算的時間，只需花費比較兩個語言模型的時間。還有一個困難是如何將得到的結果做適度的合併以減少重覆，這個問題的產生是因為目前的作法忽略英文的單複數問題，因此 restaurant 跟 restaurants 都有可能出現在結果中，導致可能會有許多結果重複，我們目前尚未找到合適的方法來快速解決這個問題。最後一個困難就是在如何將結果做排序，如：restaurant 跟 mlb 都是共同概念之一，但應該有程度的差別，如何分辨每個概念的分數並做好排序也是我們面臨到的問題之一。

以上簡短的描述就是我在美國期間進行的研究及其問題的探討，但是六個月的時間並不長，我只完成了這一部分，無法繼續參予整個系統的架構其他工作，是我此行最大的遺憾。

研修心得

這次有機會到美國參加張振川教授的研究團隊六個月，我感到相當榮幸與開心。在這六個月當中，我覺得自己學習並見識到許多東西，成長了很多，以下我將自己的經驗與感想分項列出：

1) 學術上的成長

我在去美國之前所專攻的研究領域就是網路搜尋相關的技術，因此我到美國所加入的研究團隊是伊利諾大學香檳校區資訊工程系的 The Database and Information Systems Laboratory，這個研究團隊很龐大，有幾位著名的大師在領導，讓我學習到許多目前在世界上處於領先前瞻地位的先進搜尋技術。除了技術面的學習之外，更有態度方面的衝擊，我深刻的體會到美國開放自由的學風，大家樂於討論、發現問題、動手研究，這種主動積極的研究態度是我覺得台灣學生較缺乏的，能親自感受到這樣的學風，瞭解自己的不足是我此行一項很大的收穫。

2) 國際的團隊合作

在美國和在台灣求學最大的不同就在於美國有許多國際合作的機會，因為大部分的學生都來自不同的國家，這是我在台灣較少遇到的機會。國際合作的第一大重點就在於溝通，在合作的過程當中，我發現一般平均來說，台灣的學生實作能力都不差，但卻較不擅長於以口頭表達自己的想法，也因此往往在一個團隊合作中變成無聲的角色，學習如何勇敢的表達自己的想法、流暢的用英文溝通，是國際合作中很重要的一步。

除了口語溝通的差別之外，國際合作的運作方式跟我以往在台灣所習慣的方式也不太相同。以往在台灣團隊合作大家講求分工確實、效率第一，所以會在合作的一開始進行幾次討論之後，將分工分派下去，之後大家就以快速準確的動作完成自己份內的工作；然而在美國我經歷了幾次合作經驗，發現他們的團隊合作很注重團隊討論的時間，整個過程討論的時間不比實作的時間少，大家定時開會討論，掌握每個人目前狀態，積極營造所有成員一體是

一整個團隊的感覺。我覺得兩種方式各有優缺點，我希望可以融合雙方的優點，替自己在未來的團隊合作機會裡為自己營造愉悅成功的團隊合作經驗。

3) 業界的實習經驗

我的指導教授張振川老師在美國業界有屬於自己的公司，在張老師的安排之下，在過去六個月中我和張教授的公司一直都有積極的接觸與互動。在和張老師的公司合作之前，其實我對業界的接觸並不多，只有在一年半前的暑假曾到台積電實習兩個月的經驗，當時實習下來的感覺是業界跟學術界似乎並沒有非常密切的合作，學術界研究學術界的，而業界實作業界的。

然而到了張老師公司之後打破了我的成見，在張老師的帶領之下，他的實驗室和公司相當密切的合作，實驗室發展公司需要的技術，著重在問題的探討與解決；當技術一成熟就由公司接手實作，並負責行銷技術，兩方相輔相成，跟我以往的印象不大相同。

經過這段時間的經驗之後，我深深覺得業界開放讓學生實習的機會真的是一件雙贏的事。學生可以實際瞭解業界是如何和學術界合作，並早點先體會上班工作的感覺；而企業也可以趁機塑造自己的企業形象，吸引社會新鮮人。為了達到這樣的雙贏局面，首先業界要先願意釋出合適的研發工作吸引學生來，目前台灣的業界似乎這兩三年來才正開始醞釀這樣的風潮，應繼續保持；而學校也要鼓勵學生積極把握這樣的機會，利用假期充實自己，如此一來會讓雙方都達到最大的效益。

整體來說，此行雖然只有短短六個月，但我跟許多其他國家的學生接觸，受到了另一種文化的洗禮，學到很多東西。我瞭解了自己缺乏的東西，我缺乏口頭表達的能力，也缺乏國外學生的批判思考，這都是我需要鍛鍊自己、好好培養的；但同時我更瞭解到自己的長處在哪，該如何在未來好好發揮自己的長處，彌補自己的短處，將是我接下來很重要的課題。

建議

很感謝校方及政府提供菁英獎學金讓我到美國六個月研修，我很感謝也很珍惜有這樣的機會，而針對菁英獎學金的辦理方式我有以下兩點小建議：

1) 機票費的補助

有許多同學都是想要靠自己的力量獨立出國研修才會申請這份獎學金，然而菁英獎學金只涵蓋了生活費的部份，大家都要自己額外負擔機票費，如果可以連同機票費一起補助，相信會有許多優秀的同學來爭取這份出國研修的機會。

2) 延長研修時間

我覺得六個月的時間以做研究來說其實有點短，希望可以將研修時間加長至一年左右，我覺得這樣的時間比較剛好。而剛好去年公佈新一屆菁英獎學金的實施辦法已經明白的把研修時間加長至一年，期待新一屆的受獎同學可以珍惜這麼好的機會。

以上兩點是我對菁英獎學金小小的建議。有這樣的獎學金提供學生短期出國研修真的是一個很棒的機會，希望校方跟政府可以把菁英獎學金好好的繼續辦理下去，讓更多學生可以受惠。