

以社會網路分析法改良 TOP-K 支援文件專家找尋模型之研究

李永銘 博士
交通大學
資訊管理研究所
ymli@iim.nctu.edu.tw

張宸豪
交通大學
資訊管理研究所
brucech.chang@gmail.com

張婉瑜
交通大學
資訊管理研究所
eleonore.chang@gmail.com

摘要

在組織中，經由專家幫助決策或解決問題是較有效率的作法。傳統的專家找尋方式往往只記錄某單一專家資訊而非由模型自行推理後推薦並且忽略專家間關係，隨著資訊增加與時間推移，原始記錄的專家資訊往往不再適用。本研究嘗試結合文字探勘(text mining)技術與應用社會網路分析方法學改良之專家找尋模型，並透過以前發生之近似案例來萃取出可用之專家清單，並考量關聯網路中封閉域特性與連結關係，結合程度中心性及關係權重進行專家分數評比，藉此找出在該案例領域中最具影響力之專家來進行推薦。

關鍵詞：社會網路、專家找尋、程度中心性、TOP- K

以社會網路分析法改良 TOP-K 支援文件專家找尋模型之研究

壹、緒論

隨著資訊科技快速的進展，透過各種資訊系統的支援讓各產業得而從早期人工作業模式逐漸轉換為低人力使用率的自動化生產模式。企業為保持其產業競爭力，除致力於產品或服務的創新之外，更必須著眼於製造成本的降低與品質的提升。因此，隨著各項業務需求與品質控制的要求，應用於業界之資訊系統種類越趨繁雜、分工越趨精細、功能亦越趨複雜，相對投入之開發與維護人力、時間與金錢成本亦越形龐大。

然而企業對於資訊人員快速流動、專業知識與異常處理經驗無法進行有效儲存、利用與傳承等問題對企業內部營運已造成極大的傷害。為了避免企業因為資訊系統異常等因素造成其營運損失，如何以有效率的方式來協助解決相關異常，並快速做出正確回應便成為一個企業值得深入探討、研究的議題。

企業內部應用資訊科技所建置之營運、生產、客戶支援系統日趨繁雜，也因此各系統所產生的問題亦日漸成長、複雜。因此，如何從組織已存之知識文件(如異常處理紀錄、專利文件等)找出合適的專家來協助或諮詢便成本研究想探討的課題。

根據 Parin Shah(2009)所整理，專家找尋機制大致上可以分為個人化分析(personal profiling)與文件化分析(document profiling)兩大類。個人化分析著重於利用專屬之描述檔來找出與領域相關之專家，而文件化分析則著重於利用與領域近似之文件來萃取出專家。一般來說，個人化分析較有效率，但常因專家資訊誤植而致準確率降低；文件化分析結果較為正確，但須進行大量文件比對而耗時較久。

而 Ji-Meng Chen 等人(2011)為解決文件化分析效率過慢問題提出「TOP-K 支援文件專家找尋模型」，透過最近似問題本質之前幾份支援文件來進行小範圍搜尋以提高執行效率。然該模型僅探討專家間之權重關係，並未針對其所形成之網路結構進行探討。因此，在某些條件下該模型可能找到一個最貼近問題本質而其關係連結最少的專家。

再者，根據 Marco van der 等人(2006)提出的理論，認為在擁有相同特性或共同作者的小型社會網路中，具有強連結的節點較為重要。因此本研究將嘗試以 TOP-K 支援文件專家找尋模型結合程度中心性進行研究，以期能找出最具影響力且貼近問題本質之專家並能達成下列目標：

- (一) 找到最貼近問題且關係最多之專家。
- (二) 提高專家找尋的準確度與效率。

以下章節陸續介紹此篇研究內容，章節二為文獻探討，章節三為專家找尋模型，章節四為研究流程，章節五為結論及未來展望。

貳、文獻探討

在本節中，我們將針對社會網路、分析指標及專家找尋相關的文獻作回顧及探討。在 2.1 節中針對文件探勘相關技術及觀念進行介紹。2.2 節則介紹社會網路形成與分類。2.3 節中則著重於探討社會網路中分析指標。2.4 節中則針對專家找尋機制相關之文獻進行研究。

一、文件探勘

傳統資料探勘技術主要針對結構化的資料進行分析、推論，以得到有用的規則來幫助決策或形成知識。但近年來隨著資訊科技與網路的快速發展，使得文件的儲存數量、傳播的速度與使用者取得的方式皆產生劇烈的改變，而造成資料超載等狀況。因此，現今所面臨的問題在於該如何從各種類型文件取得我們有興趣的目標。

也正因此，文件探勘在近年隨著電子媒體的擴張亦成為一門重要的學問，依照 Dan Sullivan(2001)定義為「一種編輯、組織及分析大量文件的過程，為了要提供使用者特定的資訊，以及發現某些特徵及其間之關聯」。其中所使用技術包括資訊檢索(information retrieval, IR)、資訊萃取(information extraction, IE)、自然語言處理(natural language processing, NLP)等。

本研究於此部分則使用自然語言處理相關技術來取得關鍵詞集合，並透過支援向量機(SVM)方法將文件轉換為特徵向量，最後使用餘弦相似度(cosine similarity)來計算文件間之相似度。其中餘弦相似度計算方式如公式(1)。

$$Sim(d_i, d_j) = \frac{\bar{d}_i \cdot \bar{d}_j}{|d_i| |d_j|} \quad (1)$$

其中 d_i, d_j 代表兩個文件之特徵向量， $\bar{d}_i \cdot \bar{d}_j$ 則為兩向量間之內積， $|d_i|, |d_j|$ 分別定義為， $|d_i| = \sqrt{\sum a_k^2}$ ， $|d_j| = \sqrt{\sum b_k^2}$ ， a_k, b_k 分別代表兩文件向量內之關鍵字，而 k 則為文件中關鍵字之個數。

若兩文件向量間夾角越小，則代表兩份文件之間相似度越高。如當夾角為 0 度時，其值為 1，相似度最高。而夾角為 90 度時，則值為 0，代表文件極度不相似。

二、社會網路

社會網路通常是指獨立個體間所形成的的某種鏈結關係。而獨立個體可以是人、群體或組織。而組成網路的兩個最基本元素分別為節點(node)與連結(tie)，換句話說，即節點與節點之間相互連結所形成的結構，就稱之為網路。在社會網路的研究當中，節點就代表著研究的獨立個體，而連結則用來表示個體與個體或個體與群體之間的某種社交關係，如朋友、會員、員工等。

社會網路的研究大略可分為二個類型，第一種是自我中心社會網路，第二種是整體社會網路(Bian 1994)。

(一) 自我中心社會網路

較為強調人際關係間之內涵所帶來的影響。一般稱兩個人有關係，泛指兩個人之間可能在血緣上(例如親子)或者在組織中(例如員工、師生、朋友)有直接的關聯。

(二) 整體社會網路

著重在網路結構所造成的影響。網路結構的分析則多採用圖形理論的觀念，如整體網路結構的緊密程度、節點在圖形網路中所代表的中心性或中介性等問題。若節點具有較高之中心性，在社會學的研究中，因為累積有較多的社會資本，所以通常伴隨擁有較高的權力、影響力與資源。而中介性則專指節點扮演溝通橋梁之重要程度，若中介性高的節點試圖壟斷、或拒絕協助溝通的時，其他節點之間訊息流通的溝通成本將會因此而提高。

三、 社會網路分析指標

衡量社會網路的分析指標大致可分為三種：社會網路規模(size)、社會網路密度(density)、網路中心性(centrality)(Faust, K. 1997, Scott, J. 2000)。社會網路規模是指特定社會網路中所有節點的數量，社會網路的規模是影響節點間關係的重要因素，因為節點彼此間關係的建立會受到資源多寡的限制，若是節點越多表示可用的資源可能也越多。社會網路密度是用來表示節點與節點間是否緊密或表示社會網路中節點之間的連結程度，通常密度越高代表節點與節點之間的關係越緊密。

網路中心性則可用來衡量節點在網路中影響力的大小，而一般而言分析網路中心性指標可分為三種：程度中心性(degree centrality)、親近中心性(closeness centrality)、中介中心性(betweenness centrality)(Freeman, L.C. 1979)。以下針對這三種測量網路影響力的計算做相關的說明：

(一) 程度中心性(degree centrality)：

利用相鄰節點的數量來衡量社會網路的區域中心性，節點程度中心性值越高表示這個節點的影響力可能越大。計算方式如公式(2)。

$$C_D(n_i) = \frac{d(n_i)}{g-1} \quad (2)$$

其中， n_i 為欲計算之節點， $d(n_i)$ 為計算與其他節點關係量， g 為網路所有節點量。

(二) 親近中心性(closeness centrality)：

主要是測量節點與其他節點的接近緊密程度，接近中心性值越高表示該節點影響其他節點的速度快也很強烈。計算方式如公式(3)。

$$C_C(n_i) = \frac{g-1}{\left[\sum_{j=1}^g d(n_i, n_j) \right]} \quad (3)$$

其中， n_i 為欲計算之節點， $d(n_i, n_j)$ 為兩點間距離， g 為網路所有節點量。

(三) 中介中心性(betweenness centrality)：

衡量某一節點存在於其他任兩點路徑上的重要程度。當中介中心性值越高時，表示該節點是位於溝通與橋梁的重要地位。計算方式如公式(4)。

$$C_B(n_i) = \frac{\sum_{j < k} g_{jk}(n_i)}{[(g-1)(g-2)]} \quad (4)$$

其中， $g_{jk}(n_i)$ 為通過 i 連接 j 及 k 最短路徑數， g_{jk} 則為連接 j 及 k 最短路徑數。

根據Mitchell, J. C.(1969)與Scott, J.(2002)提到程度中心性是利用相鄰的個體數量，來衡量社會網絡的區域中心性，可以看出個體控制範圍大小；而接近中心性與中介中心性會計算網路所有節點相關距離。本研究在此部分主要利用支援文件與專家之直接關係所形成的關聯網路，因此主要探討程度中心性對模型造成之影響。

四、 專家找尋機制

近幾年來，專家找尋(Expert Finding)成為一個新興的課題，有許多學者亦針對該領域付出相當心力並提出各種不同的方法，並進而協助使用者或企業能從內部網路與線上論壇找到特定主題之專家。

根據 Parin Shah(2009)所整理，專家找尋機制大致上可以分為個人化分析(personal profiling)與文件化分析(document profiling)兩大類。

(一) 個人化分析(personal profiling)

透過各種包含專家描述的文件來擷取專家相關資訊，並進而形成專家個人描述檔。當個人描述檔完備後，針對所輸入之特定關鍵字，搜尋引擎則查詢透過專家描述比對，取得相關之專家列表。如圖 1 所示：

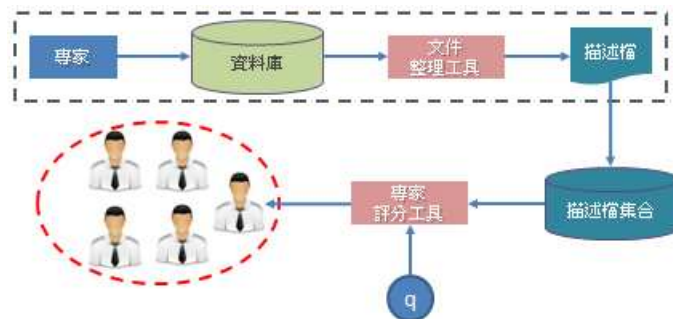


圖 1 個人化分析概念圖

(二) 文件化分析(document profiling)

針對所輸入之特定關鍵字，查詢資料庫所有文件並取得與查詢相關之文件。而後根據這些相關文件之內容，如電子郵件、姓名等資訊來萃取出可能之專家列表。如圖 2 所示：

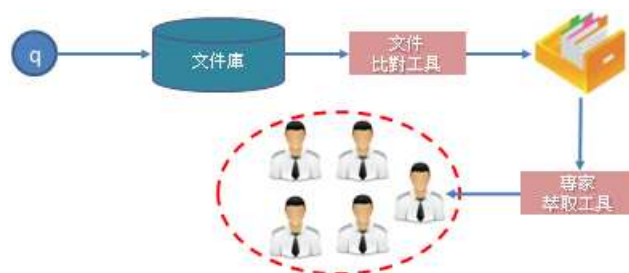


圖 2 文件化分析概念圖

而 Ji-Meng Chen 等人(2011)所提出之前 K 筆支援文件專家找尋模型即是屬於文件化分析的專家找尋模型。該模型首先利用查詢條件從文件庫找出相關性較高之文件後，接著從文件中萃取出現在文件中之專家來形成專家列表，而後再深入探究專家間之共同出現(co-occurrence)之關係來建置成一個小型網路並利用其間關係權重來找出得分最高之專家進行推薦。

參、專家找尋模型

在此章節，本研究將介紹一個支援文件專家找尋機制之改良模型，並且針對基本想法與模型各參數意義進行說明。

一、基本想法

根據 Ji-Meng Chen 等人(2011)所提出之前 K 筆支援文件專家找尋模型，認為透過問題查詢來取得最近似之前 K 筆支援文件，並從這些文件萃取出專家清單，那麼透過探討支援文件與問題的相似度、共同出現在支援文件之關係(co-occurrence)與關係權重值(relation weight)則能進行評分來取得最高得分之專家，如圖 3 所示。然而，此模型雖已加入專家間關係來進行評分加值，仍可能因支援文件高相似度而找到一個評分最高，而其關係最薄弱、專業年資最少之專家。

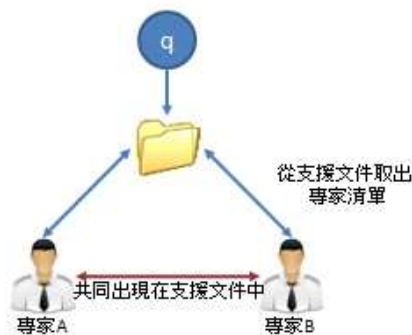


圖 3 專家關係圖

因此，若將原模型關係圖擴展來看，如圖 4 所示。可以發現透過支援案例與專家因共同撰寫或共同出現等關係會形成一個關聯網路(affiliation network)，若僅探討支援文件相似度及專家權重貢獻度將可能導致誤差；反之若於原模型外再加入網路架構之探討，則能找到一個文件相似度高、與其他專家關係最多之較具影響力專家，進而能有效提高專家找尋之正確性。也就是說，當模型加入網路結構探討時即可能找到一個年資較長、經驗較豐富且社會資產(social capital)最豐富的專家。

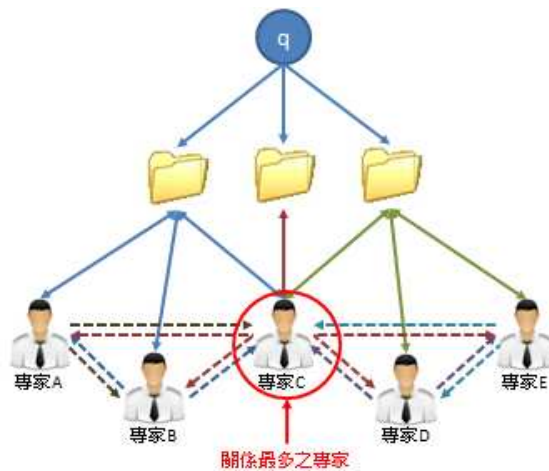


圖 4 專家網路圖

二、 支援文件專家找尋改良模型

本研究之專家找尋改良模型認為當某些人頻繁出現在支援文件中，那麼這些人可能就是這個主題的專家，而透過文件相似度、關係權重與程度中心性(degree centrality)總合評分則能找出其中最貼近主題、連結關係最多且最具影響力的主要專家。

(一) 專家初始分數(initial expert score)

首先，我們利用文章相似度、專家出現機率與專家之程度中心性計算專家的原始分數，如公式(5)所示：

$$S_1(e, q) = \frac{pf^2}{tf} \sum_{d \in SD} p(d|q)p(e|d, q) \quad (5)$$

其中， pf 為專家出現在文件 d 之次數， tf 為專家出現在所有支援文件 SD 之總次數。而 $p(d|q)$ 為透過餘弦相似度(cosine similarity)所計算之文件相似度， $p(e|d, q)$ 為專家出現在文件 d 的機率。

(二) 關係權重(association weight using co-occurrence)

緊接著，我們必須利用專家共同出現(co-occurrence)在某個支援文件中之關係來計算他們之間的權重，以便後續進行某特定專家綜合評分時能將其相關之其他專家貢獻度大小以量化的方式來表示。權重計算方式如公式(6)所示：

$$W_{i,j} = \frac{cr(e_i, e_j)}{|E_i| \cdot |E_j|} \quad (6)$$

其中， $|E_i|$ 指的是與專家 e_i 有關係的專家數， $cr(e_i, e_j)$ 是計算兩專家共同出現之支援文件個數，如公式(7)所示

$$cr(e_i, e_j) = \sum_{d \in SD} a(e_i, e_j, d) \quad (7)$$

當兩專家都出現在文件則為 1；反之則為 0。在此部分，本研究透過計算專家間共同出現於支援文件之交集機率方式來表達兩專家間關係緊密程度。當計算所得之值越高，代表專家間關係益加密切。

(三) 專家找尋模型

最後，完成專家初始分數與關係權重計算後，本研究依據 Ji-Meng Chen 等人(2011)所提出之遞迴計算模型加入程度中心性提出改良模型，如公式(8)所示：

$$S_{n+1}(e, q) = \gamma C_D(e) S_n(e, q) + (1 - \gamma) \sum_{(e, e') \in E} W_{e, e'} S_n(e', q) \quad (8)$$

其中， n 為遞迴次數， γ 為專家影響力。而與原始模型不同在於加入程度中心性可以協助我們找到貼近問題且關係最多的專家。

肆、研究流程

在此章節，本研究將針對改良模型研究流程進行介紹並利用範例計算來協助大家能快速了解模型運作方式與實作方式。研究流程如圖 5 所示：

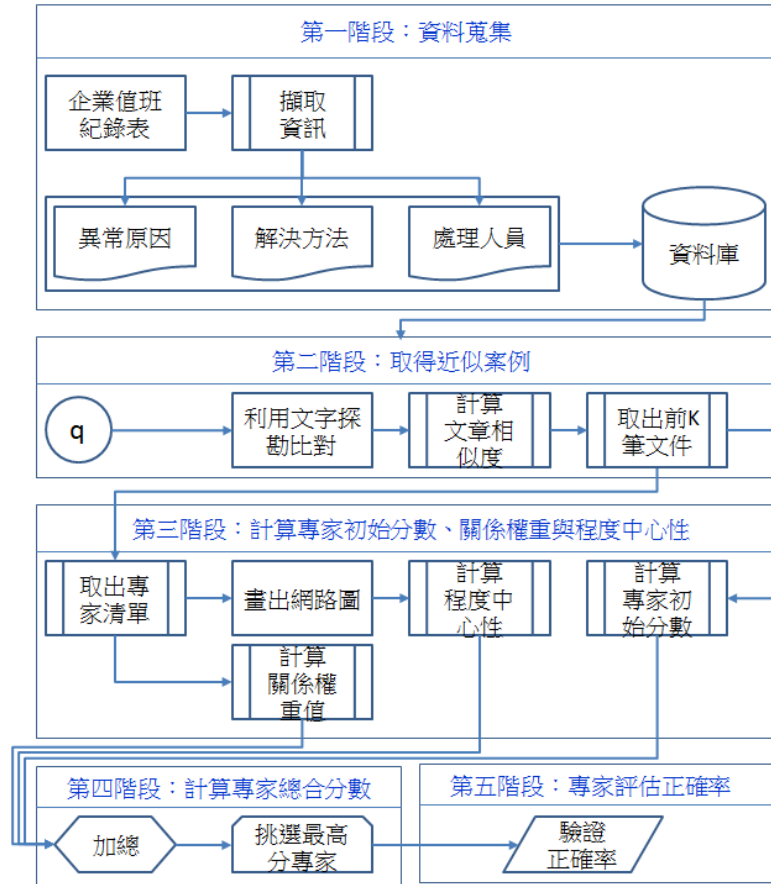


圖 5 研究流程圖

一、 第一階段：資料蒐集

為實驗專家找尋模型是否能有利益於企業找尋內部專家，本研究以科學園區某半導體公司資訊系統值班紀錄表為資料來源，該表格式如圖 6 所示。

	E	G	H	I	K
	Date Time	Problem	Cause	Action	Recorder
1	2006/12/16 0:00	Cannot reserve (told by TCS); User couldn't reserve FOUF	Error msg: Lot has not been reserved. FOUF in stocker, just waiting for transportation. User/TCS did not check transfer job on site and deleted reservation but transfer command still exists in MCS/OPI	Ask AMHS member to delete the transfer job. EQP EGA01(FULL-A) then reserved foup again.	
2	2007/1/4 15:00	User got ECMS alarm continuously.	It's ECMS Alarm, which ask user to call 81 when got the alarm continuously.	Notify user to check the engineering Lots she asked. If they(the engineering Lots she asked) had changed and got the alarm again. Please called us again. With other questions, please call	
3	2007/1/2 9:13	SCHDB Lock.	User disconnect SCH system abnormal.	Connect to SCH DB, and delete lock record.	
4	2007/1/2 10:20	BackupNGHold	SpecOverHold occur first. Cause BackupNGHold occur.	Release SpecOverHold and BackupNGHold. System will auto execute Backup step.	
5					

圖 6 資訊系統值班紀錄表

根據其固定格式，實作一個資料讀取程式依序讀入資訊系統異常原因、解決方法與處理人員，而後再行寫入資料庫以利後續各階段使用。而因企業內部系統繁雜，本實驗選定特定類別並蒐集其相關處理案例共 137 件以供實驗使用。

二、 第二階段：取得近似案例

於此階段，本研究首先利用工具分別對取得之 137 份案例進行斷詞來取得關鍵詞組，並藉由詞性過濾留下動詞與名詞之重要詞性。緊接著利用支援向量機方式，將各案例轉換為對應之案例特徵向量。最後，利用餘弦相似度(cosine similarity)計算查詢關鍵字集與已存案例相似度，在完成計算並依大小排序後則依序取出前 K 筆支援案例以供擷取專家之用。例如，輸入「FOUP 註冊 SEO 後一直沒更換」後，系統找到兩筆支援案例如表 1 所示：

表 1 支援案例資訊

案例一(C1)			
類別	OPI	功能	SEO
錯誤訊息	*	錯誤碼	*
異常描述	註冊 SEO 後，一直沒上機台進行更換 FOUP 動作		
維護歷史	EXPA, EXPC	相似度	0.93
案例二(C2)			
類別	OPI	功能	SEO
錯誤訊息	*	錯誤碼	*
異常描述	FOUP 無法註冊 SEO 以進行 SPLIT 動作		
維護歷史	EXPA, EXPB, EXPA	相似度	0.62

三、 第三階段：計算專家初始分數、關係權重與程度中心性

當根據查詢條件取得近似案例後，即能利用其相似度、專家出現頻率與網路關係來計算初始分數、關係權重及程度中心性。在此，我們以第二階段提供之範例進行推演如下所示：

(一) 程度中心性

根據案例維護清單與共同出現之關係整理出關係矩陣，根據如表 2 所示：

表 2 專家關係矩陣

	專家一(EXPA)	專家二(EXPB)	專家三(EXPC)
專家一(EXPA)	1	1	1
專家二(EXPB)	1	1	0
專家三(EXPC)	1	0	1

根據公式(2)，我們就能分別得到 $C_D(EXPA) = 1$ 、 $C_D(EXPB) = 0.5$ 、 $C_D(EXPC) = 0.5$ 。

(二) 專家初始分數

依據案例相似度、出現次數即能依據專家初始分數公式(5)計算各專家初始分數，整理如表 3 所示：

表 3 專家初始分數計算參數暨結果整理表

	專家一(EXPA)	專家二(EXPB)	專家三(EXPC)
C1-出現次數	1	0	1
C2-出現次數	2	1	0
出現總次數	3	1	1
C1-出現機率	0.33	0	1
C2-出現機率	0.67	1	0
C1-案件相似度	0.93	0.93	0.93
C2-案件相似度	0.62	0.62	0.62
專家初始分數	0.65	0.62	0.93

由結果可知，在不考慮任何網路關係與權重值之下，初始分數僅取出最貼近問題本質之專家。

(三) 專家間權重

依據專家共同出現之案例與權重計算公式(6)，並以第二階段範例計算權重，結果如表 4 所示：

表 4 專家關係權重整理表

	專家一(EXPA)	專家二(EXPB)	專家三(EXPC)
專家一(EXPA)		0.5	0.5
專家二(EXPB)	0.5		0
專家三(EXPC)	0.5	0	

四、計算專家總合分數

完成近似案例擷取並透過計算取得專家初始分數與關係間之權重值後，即能利用專家總合計算公式(8)進行計算已取得各專家最後得分。計算結果如表 5 所示：

表 5 專家分數計算整理表

	專家一(EXPA)	專家二(EXPB)	專家三(EXPC)
專家初始分數	0.65	0.62	0.93
程度中心性	1	0.5	0.5
原始分數(n=2)	0.69	0.53	0.75
改良分數(n=2)	0.69	0.32	0.42
原始分數(n=3)	0.68	0.48	0.63
改良分數(n=3)	0.59	0.22	0.25

由推演結果可以發現原始模型於第三次運算後才找到正確專家，且與次高分專家分數差距僅 0.05，易造成系統判斷失效；而新模型於第二次運算即找到正確專家，且與次高分專家分數差距僅 0.27，於第三次運算時差距則擴大至 0.34，充分表現出網路富者越富的現象，且能充分提高系統判斷準確度。

因此專家找尋模型加入程度中心性後，能夠在更少的遞迴次數下更快找出具有較多關係，且貼近問題之正確專家。並且當遞迴次數增多時，正確專家與其他專家分數也形

成一個明顯差距以利系統判斷。因此，相較於原始模型方法，當運用新模型來進行專家找尋則能夠有效提升運算速率與推薦準確率。

五、 評估方式

因專家係從某企業內部之值班紀錄表萃取而得，但因專家們皆是該領域之專家，僅差別在年資、專業程度以及負責系統之差別。所以本研究主要針對專家推薦滿意度為主要驗證標的，透過滿意度高低來評判模型推薦之優劣。

所以本研究自訓練案例中取出 10% 案例當作測試樣本，並利用原模型及改良模型支援案例中取出之專家資訊交予專家針對推薦專家進行滿意度調查。評分標準為 1-5 分，分數由小至大分別為「非常不滿意」、「不滿意」、「尚可」、「滿意」、「非常滿意」。

在此，我們以收集之 137 份案例作為訓練資料，並抽取其中之 14 份案例當作測試案例進行專家推薦並委請專家人工判定滿意程度後所得相關結果如表 6 所示：

表 6 滿意度結果整理表

	樣本數	測試案例數	平均滿意度
原始模型	137	14	4.13
改良模型	137	14	4.57

我們可以發現兩組模型滿意度分數均介於滿意至非常滿意。改良模型平均滿意度為 4.57 分高於原始模型，而其得分較高之原因乃其為其推薦之專家大多為年資較長、經驗較為豐富之員工，在主觀認定下這些專家能夠給予協助或諮詢的程度較高，因而給予較高之滿意度。

伍、 結論與未來展望

社會網路分析方法是由社會學家結合數學方法、圖論等基礎學說而發展出來的定量分析方法。近年來，亦由於社群力量在情感交流、口碑影響、市場行銷及國際貿易等領域廣泛應用而讓社會網路分析方法成為一種熱潮。

本研究利用社會網路分析法探討專家間關係及相互影響的力量，並進而提出一個 TOP-K 文件專家找尋之改良模型。就研究結果而言，我們可以發現改良後的專家找尋模型加入了程度中心性之參數後，讓專家評分因該參數之影響而造成分數快速變化，進而加大專家得分之區隔程度來達到比原模型更好的判斷準確度，並能有效處理原模型中在特殊情況中所造成之錯誤專家推薦。例如，某專家與其他專家沒有任何關係，但其文件與問題的相關度極高時，那麼該專家可能變成評分最高的專家。

因此，透過改良的專家找尋模型能更有效率的找到特定主題的專家外，亦更容易找到比其他專家更富有關係(如合作夥伴、學術成員或專案成員等)的專家，來協助企業員工進行新問題處理或視為諮詢角色以補案例資訊不足之處。如此即能協助企業有效率的解決問題、獲得更多創新的想法，進而使得企業得而提升其整體競爭力。

參考文獻

1. Faust,K.,“Centrality in affiliation networks.”, Social Networks, vol.19, 1997, pp.157-191.
2. Freeman, L.C.,“Centrality in Networks Conceptual Clarification.”, Social Networks, vol.1, 1979, pp.215-239.
3. Ji-Meng Chen, Jie Liu, Ya-Lou Huang, Min Lu,“Efficient top-k support documents for expert search using relationship in a social network”, ICMLC, vol.4, 2011, pp.1479-1484.
4. Marco van der Leij, Sanjeev Goyal “Strong Ties in a small world” Tinbergen Institute Discussion Paper, TI 2006-008/1.
5. Mitchell, J. C., “Social networks and urban situations. ”, England: Manchester University Press, 1969.
6. Parin Shah,“Expert Finding Using Social Networking”, SJSU, Master Theses, 2009.
7. Scott, J.,“Social Network Analysis : A Handbook”, London, Sage, 2000.
8. Scott, J., “Social Network Analysis: Critical Concepts in Sociology”, New York, Routledge, 2002.
9. Sullivan,D.,“Document Warehousing and Text Mining”,Wiley Computer Publishing, 2001.
10. Y.J Bian,“Guanxi and the allocation of jobs in urban China.”, The China Quarterly, vol.140, 1994, pp.971-999.

Improving TOP-K Support Documents for Expert Search by Using

Social Network Analysis

Dr. Yung-Ming Li

National Chiao Tung University
Institute of Information Management
ymli@iim.nctu.edu.tw

Chen-Hao Chang

National Chiao Tung University
Institute of Information Management
bruceh.chang@gmail.com

Wan-Yu Chang

National Chiao Tung University
Institute of Information Management
eleonore.chang@gmail.com

Abstract

Searching experts is an effective solution to solve problems and help make decisions in an organization. Traditional approaches of expert finding usually only record the expert information without using a self-inference model nor the experts relationship. In the real world, experts will be influenced by other experts, finding the most influence expert can give a good chance to solve the given topic. In this paper we utilize the text mining technique and social network analysis method to develop top-k support documents and expert search model. We use text mining and similarity technique to find the candidate expert list. Using the degree centrality of the expert candidate network and a given topic allows us to find the expert of influence through the experts score rank. In such words, this approach combining the reliability and influence between the experts can improve the accuracy of expert recommendation for the topic-solving in the real world.

Keywords: Social Network; Expert Search; Degree Centrality; TOP-K support documents