

國立政治大學資訊科學系  
Department of Computer Science  
National Chengchi University

碩士論文

Master's Thesis

整合社群關係的 OLAP 操作推薦機制

A Recommendation Mechanism on OLAP Operations  
Based on Social Network

研究生：陳信固

指導老師：李蔡彥 教授

中華民國一百零一年七月

July 2012

整合社群關係的 OLAP 操作推薦機制

A Recommendation Mechanism on OLAP Operations  
Based on Social Network

研究生：陳信固

Student：Hsin-Ku Chen

指導教授：李蔡彥

Advisor：Tsai-Yen Li



國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

in

Computer Science

中華民國一百零一年七月

July 2012

# 整合社群關係的 OLAP 操作推薦機制

## 摘要

近幾年在金融風暴及全球競爭等影響下，企業紛紛導入商業智慧平台，提供管理階層可簡易且快速的分析各種可量化管理的關鍵指標。但在後續的推廣上，經常會因商業智慧系統提供的資訊過於豐富，造成使用者在學習階段無法有效的取得所需資訊，導致商業智慧無法發揮預期效果。本論文以使用者在商業智慧平台上的操作相似度進行分析，建立相對於實體部門的凝聚子群，且用中心性計算各節點的關聯加權，整合至所設計的推薦機制，用以提升商業智慧平台成功導入的機率。經模擬實驗的證實，在推薦機制中考慮此因素會較原始的推薦機制擁有更高的精確度。

**關鍵詞：**社群網路分析、推薦機制、社群偵測、商業智慧、網絡中心性

# **A Recommendation Mechanism on OLAP Operations Based on Social Network**

## **Abstract**

In recent years, enterprises are facing financial turmoil, global competition, and shortened business cycle. Under these influences, enterprises usually implement the Business Intelligence platform to help managers get the key indicators of business management quickly and easily. In the promotion stage of such Business Intelligence platforms, users usually give up using the system due to huge amount of information provided by the BI platform. They cannot intuitively obtain the required information in the early stage when they use the system. In this study, we analyze the similarity of users' operations on the BI platform and try to establish cohesive subgroups in the corresponding organization. In addition, we also integrate the associated weighting factor calculated from the centrality measures into the recommendation mechanism to increase the probability of successful uses of BI platform. From our simulation experiments, we find that the recommendation accuracies are higher when we add the clustering result and the associated weighting factor into the recommendation mechanism.

**Keywords** : Social Network Analysis, Recommendation Mechanism, Community Detection, Business Intelligence, Network Centrality

## 目錄

第一章 緒論.....	1
1.1 前言.....	1
1.2 研究動機.....	2
1.3 研究目的.....	3
1.4 研究方式.....	4
1.5 論文架構.....	6
第二章 相關研究.....	7
2.1 推薦機制.....	7
2.1.1 OLAP 推薦機制.....	8
2.1.2 網頁推薦機制.....	10
2.1.3 Page Rank 演算法.....	10
2.2 社會網絡分析.....	11
第三章 研究方法.....	13
3.1 研究假設.....	13
3.2 系統架構.....	14
3.3 使用者操作紀錄收集.....	15
3.4 多維度操作紀錄正規化.....	16
第四章 社群網絡分析.....	19
4.1 凝聚子群.....	19
4.1.1 Modularity Q 的計算.....	20
4.1.2 社群切割(Subgroups).....	22
4.2 老手程度的判斷.....	23
第五章 推薦機制.....	27
5.1 相似度判斷.....	27
5.2 候選項目篩選.....	29
5.3 推薦機制：最大信心度選擇.....	31
5.4 推薦機制：最大使用人次選擇.....	32
5.5 推薦機制與參考關聯加權.....	32
第六章 系統實作與實驗.....	34
6.1 程式語言、資料來源.....	34
6.2 多維度分析推薦輔助系統功能模組說明.....	34
6.2.1 操作紀錄正規化.....	35
6.2.2 操作紀錄之相似度判斷.....	36
6.2.3 產生推薦候選項目集合.....	37
6.2.4 使用者分群.....	38

6.2.5 老手程度加權計算 .....	41
6.2.6 推薦項目產出 .....	42
6.2.7 使用者回饋機制 .....	42
6.3 多維度分析推薦輔助系統操作介面介紹 .....	43
6.3 模擬實驗 .....	46
6.3.1 推薦機制：最大信心度選擇的模擬結果分析 .....	47
6.3.2 推薦機制：最大使用人次選擇的模擬結果分析 .....	48
6.3.3 推薦機制：最大信心度選擇加入參考關聯加權的模擬結果分析 .....	49
6.3.4 推薦機制：最大使用人次選擇加入參考關聯加權的模擬結果分析 .....	51
6.4 使用者實際操作回饋 .....	52
6.4.1 問卷的回饋與分析 .....	53
6.4.2 推薦機制的正確率分析 .....	54
<b>第七章 結論與未來發展 .....</b>	<b>58</b>
7.1 研究結論 .....	58
7.2 未來發展 .....	59
<b>參考文獻 .....</b>	<b>61</b>
<b>附錄一：多維度分析平台推薦輔助系統實驗問卷 .....</b>	<b>64</b>

## 圖目錄

圖 1-1、商業智慧系統發展程序.....	2
圖 1-2、研究流程圖.....	4
圖 3-1、系統架構圖.....	13
圖 3-2、多維度分析平台操作介面.....	14
圖 4-1、社會網絡圖形.....	18
圖 4-2、Factions 演算法分群結果.....	21
圖 4-3、Girvan-Newman 演算法分群結果.....	22
圖 4-4、Degree 的 Centrality measures 結果.....	24
圖 4-5、Closeness 的 Centrality measures 結果.....	25
圖 4-6、Betweenness 的 Centrality measures 結果.....	26
圖 6-1、多維度分析推薦輔助系統流程圖.....	35
圖 6-2、操作紀錄正規化處理流程圖.....	35
圖 6-3、相似度判斷處理流程圖.....	36
圖 6-4、產生推薦候選項目集合流程圖.....	37
圖 6-5、使用者分群的流程圖.....	38
圖 6-6、以節點 A 為起始點的最短路徑處理範例.....	40
圖 6-7、老手程度加權的流程圖.....	41
圖 6-8、推薦項目產出的流程圖.....	40
圖 6-9、使用者回饋的流程圖.....	43
圖 6-10、推薦系統參數設定介面.....	44
圖 6-11、推薦輔助系統登入畫面.....	44
圖 6-12、查詢操作正規化結果.....	45
圖 6-13、推薦結果.....	45
圖 6-14、更多推薦的結果.....	46

## 表目錄

表 1.1、資訊系統專案導入成功比例圖.....	2
表 3-1、使用者於多維度分析系統的操作紀錄範例.....	14
表 4-1、虛擬關聯分佈矩陣.....	19
表 6-1、操作紀錄分析與正規化的輸入與輸出.....	36
表 6-2、產生推薦候選項目集合的輸入與輸出.....	37
表 6-3、使用者分群的輸入與輸出.....	41
表 6-4、推薦機制：最大信心度選擇的模擬結果.....	48
表 6-5、推薦機制：最大使用人次選擇的模擬結果.....	48
表 6-6、推薦機制：最大信心度選擇加入參考關聯的模擬結果.....	49
表 6-7、最大信心度選擇的推薦機制加入參考關聯加權後的正確推薦順位比較...	50
表 6-8、推薦機制：最大使用人次選擇加入參考關聯的模擬結果.....	51
表 6-9、最大使用人次選擇加入參考關聯加權後的正確推薦順位比較.....	51
表 6-10、實驗人員的背景資料.....	53
表 6-11、OLAP 操作輔助系統的滿意度回饋統計.....	53
表 6-12、實驗的推薦正確性測試結果.....	55
表 6-13、三種類型使用者的測試統計.....	55



# 第一章 緒論

本章主要在說明本研究的動機與目的，透過背景說明闡述本研究所要討論之主要目標，並提出整個研究的流程步驟和架構。

## 1.1 前言

依據 IDC 的商業智慧市場調查報告，亞太地區(不含日本) 2003 年在於商業智慧(Business Intelligence ,BI)的市場規模為 5.3 億美元；但在 2003 年到 2008 年的階段以 18.6%的比率呈現一個複合性的成長，並在 2008 年突破 12 億美元。顯示企業在資訊科技發展從早期的自動化與電子化的推動，逐漸將重心轉移到思考如何整合企業內部資源與資訊。尤其近幾年企業面臨著金融風暴、全球競爭、景氣循環縮短等影響下，企業經理人除了藉由以往的「經驗法則」作為決策的依據外，還必須能即時獲得所需的參考資訊才能夠為企業有效率的作出關鍵決策。但以早期資訊系統發展的情況，往往資訊的取得都是費時費力，經由各系統負責人將局部資料以報表形式提供後，最後仍需進行彙總的整理後才能獲得所需的關鍵指標。因此企業管理階層如何在營運過程中快速取得所需的資訊，已是企業在資訊科技發展的一大重心，而商業智慧的技術即可提供企業作為一個解決方案。

管理大師 Robert Kaplan 及 David Norton 說過：『如果你無法衡量它，你就無法有效的管理它』。對於一個管理者來說，面對著企業以倍數在成長的資料，如何將雜亂的資料轉換為企業經營上的實質效益，而商業智慧系統即是目前企業積極用來達成此目標的一個解決平台。圖 1.1 表示是一般我們在導入商業智慧系統的三個過程，包含商業智慧系統分析、商業智慧系統建置與商業智慧系統運用。一個商業智慧系統的導入，首先需要透過需求訪談協助找出企業內部可量化管理的關鍵指標(KPI)，並且定義該指標的計算公式與該指標在管理層面所需要分析的面向(維度)。接下來依據定義的指標找尋其相對的來源資料，並且利用資料轉換(ETL)的工具轉至倉儲系統內，最後在平台上建立各主題的多維度分析模組，提供每個人依據職務的需求取得所需的相關資訊，讓企業能快速

反應外在的變化與企業的難題。

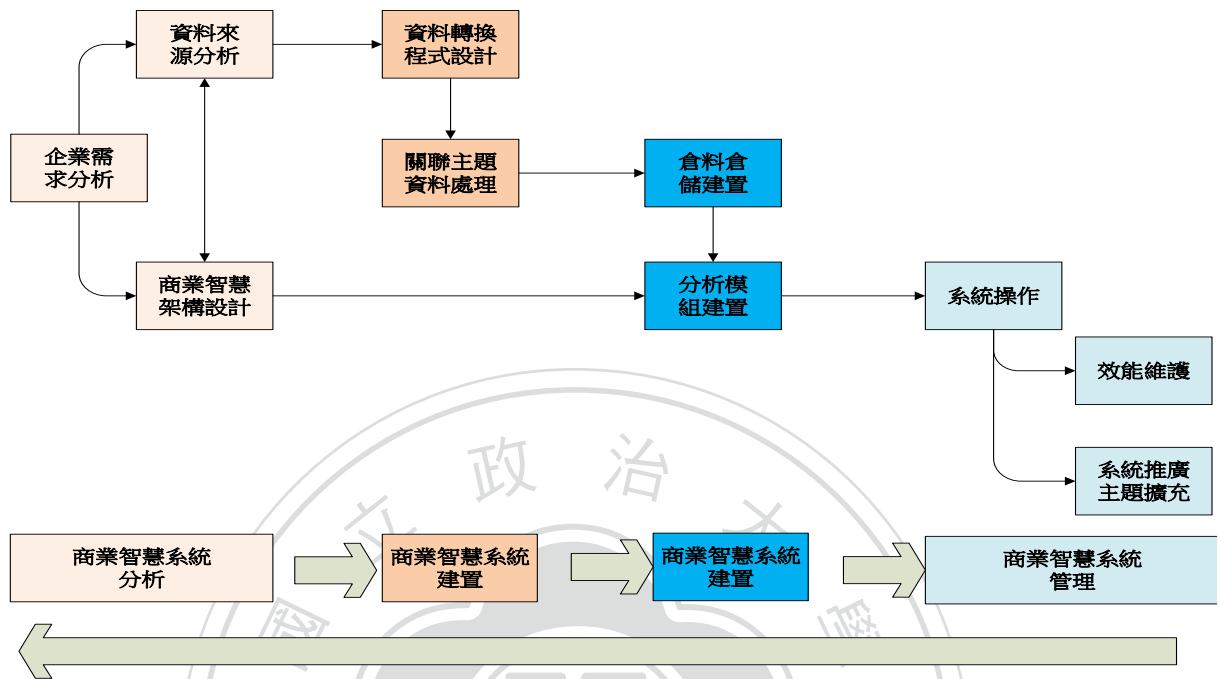


圖 1-1、商業智慧系統發展程序

資料來源：資策會 2012

## 1.2 研究動機

根據 CHAOS Summary 2009 報告，企業在導入資訊系統時的執行狀況如下

專案導入結果	定義	比例
成功(succeeding)	如期，如質，如預算	32%
不如預期 (challenged)	雖然完成了，但時程不如預期，或超支、或刪減需求、或品質不佳等	44%
失敗(failed)	取消專案，或雖然做完了，卻從不使用。	24%

表 1.1、資訊系統專案導入成功比例圖

由表 1.1 中我們可以發現由於資訊系統專案的特性，無法成功完成專案目標的機率高達 76%，而商業智慧的專案又因為使用者的認知不足與系統的複雜度，更造成專案不成功的比率高達 85%。在我們為客戶導入商業智慧專案的過程中，發現通常商業智慧專案失

敗的類型分為兩種，第一種我們歸類為建置不良的商業智慧系統，此種類型的專案因為導入人員在需求訪談階段無法有效引導使用者，造成設計階段無法規劃出正確的模組架構，而導致最後無法建置符合期望功能的系統；這樣的專案通常都是以建置大量的報表來結束專案，而無法提供多維度分析平台應有的效益。第二種我們歸類為商業智慧系統推廣不良，因為商業智慧的專案為了提供使用者一個具即時性與彈性的資料操作平台，在建置的過程中盡可能會將以後會用到的指標與潛在指標內容都整合至模組內，但在導入初期也常因為提供的資訊過多，因而導致對於模組不熟悉的使用者會望之怯步，慢慢的造成此系統的僅侷限在特定的人員使用。目前商業智慧大部分的研究都在探討倉儲系統建置[8]、多維度系統開發方法論與資料採礦的運用以及演算法的研究，也有人研究怎麼提供商業智慧系統需求的分析模型，都是為了提升商業智慧專案成功的機率。在本研究中我們希望能從商業智慧系統的推廣方向著手，探討如何藉由輔助決策系統的方式提升商業智慧在企業內推廣的成效。

### 1.3 研究目的

基於商業智慧專案在於企業推行的成功機率偏低，本研究探討如何建置即時的多維度分析模組的操作推薦系統。嘗試對商業智慧平台上使用者所建立的社群網絡進行分析，依據其操作的相似性對使用者作分群，並且將此結果運用在後續建立的多維度分析平台操作推薦的輔助系統中，降低企業因為商業智慧系統推廣不佳而造成無法達到預期目標的情況。本研究主要的研究目的如下：

- 一. 降低使用者在商業智慧系統上學習的時間：在實務的 OLAP 專案上，我們建置 Cube 的類型主要會分為兩種，一種是給高階管理者使用，包含的內容會是企業經營的主要指標，這種類型的 Cube 通常包含的量值與維度都不會超過 20 個，因此使用者在使用上不會有太大的困難；但第二種類型是提供給專門提供分析報告的人員使用，為了因應他們對資料的即時與有效性，往往我們在建置的過程中會將所有可能分析到的重要面向與指標都整合在單一 Cube 中，以本次實驗

的對象，我們在建置的卡友消費分析的 Cube 中包含的維度就高達一百三十二個，對於系統的老手而言，他們可以很快速的從系統中取得所需的分析資料，但是對於一般的使用者而言，他們可能需要花費大量的時間在找尋所需的維度或量值。因此將藉由推薦系統的輔助，適時在操作的過程中提供建議，經由參考其他相關使用者的操作紀錄，降低使用者在商業智慧系統上學習的時間。

二. 提昇群體間知識分享的成效：在 OLAP 系統中進行資料分析的過程中，我們通常都是利用嘗試性的方法，對系統的所提供的維度與量值作測試性的操作後方能找到所需的資訊，最後在透過報表的討論才能夠達到知識分享的效果；但是這樣的模式會造成報表製作的效率不佳，因此將藉由此推薦系統的建置，讓使用者在操作的過程中能檢視同群體的人員在分析相似問題時關心的面向，加速使用者資料檢索的效率。

三. OLAP 推薦機制在實務上的驗證：以往研究的對象都是利用範例 CUBE 或者是小型的 Cube 進行問卷式的實驗，在本次的實驗中我們將會以實務上已經導入多維度平台的企業做為研究對象，面對更複雜結構與包含更多維度的 Cube 來說，實際能夠成功從一兩百個維度中找出適合的項目做為推薦的機率是非常低的，因此將藉由本次的實驗將會利用實務上的操作紀錄做為實驗資料，用以驗證這樣的 OLAP 推薦機制在實務上的幫助是否如預期。

## 1.4 研究方式

基於上述研究目的，本研究首先會收集使用者在於多維度分析平台上的操作紀錄，並藉由正規化方式將資料整理成可做分析使用的集合。我們將會利用社會網絡分析(Social Network Analysis, SNA)技術，依據使用者在商業智慧平台上資料檢索內容的相似度，將使用者進行分群並且判斷子群內哪些人是生手哪些人是老手。最後，我們將把社群網絡分析的結果運用在推薦輔助系統上。在使用者操作系統的過程中，系統會自動分析該

使用者目前的操作內容，與歷史資料進行相似度的比較，以找出可能的候選項目，最後依據最大使用人次選擇方式作為候選項目選擇的判斷，實驗的過程中也將會將最大使用人次選擇推薦機制與一般常用的依據支持度(Support)的推薦機制相比較。本研究之研究方法步驟如圖 1-2 所示

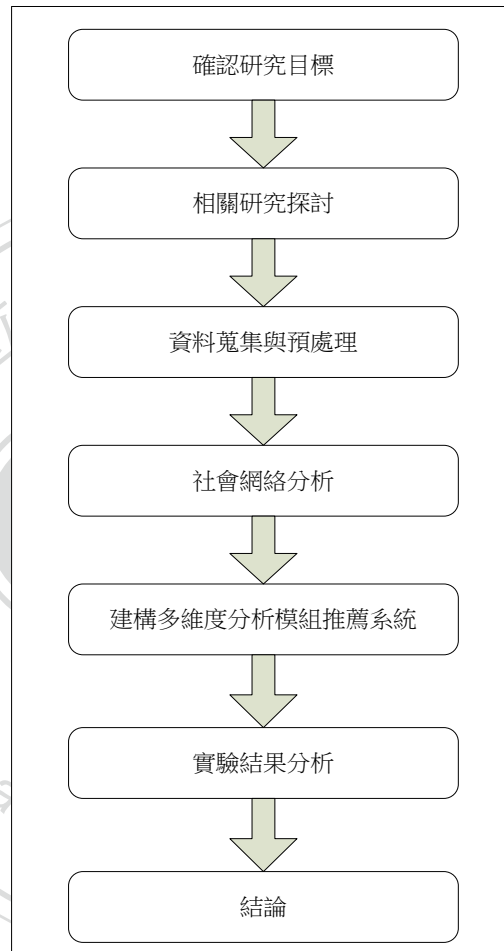


圖 1-2、研究流程圖

## 1.5 論文架構

本研究共分七個章節，本章為研究概論，在說明研究的動機與目的；第二章將對相關研究作介紹，主要是對社會網絡分析(Social Network Analysis)、網頁推薦機制與 OLAP 推薦機制等相關研究作探討；第三章主要在介紹本研究問題的假設與資料收集的方法，首先我們會先依據以往在實務上的經驗來定義此次研究的假設，接著會說明本次研究的系統架構，然後會對資料的收集及資料正規化的處理作進一步的說明。第四章為社群網絡分析，將說明如何對使用者進行分群，並且定義系統中使用者的老手程度。第五章，為推薦機制說明，將對操作相似度判斷、候選項目的篩選與推薦機制分別作說明；第六章為系統的實作與實驗結果的分析與討論；最後第七章介紹本論文的結論與未來的發展方向。





## 第二章 相關研究

本研究的目的是在對線上分析處理 (On-Line Analytical Processing, OLAP) 系統的使用者行為進行分析，經由推薦輔助系統提升企業在於商業智慧專案的推廣。研究的過程中，參考的觀念技術包含在 OLAP 上的使用者行為預估[9][13]與推薦技術[17][19][21]、網頁推薦機制[12]與資料探勘[6][14]等；此外，我們也會利用社群網絡分析(Social Network Analysis, SNA)的技術[15][16]，依據使用者在 OLAP 系統上操作的相似度，將一個大網絡分割為多個同質的凝聚子群，並且識別每個人在各社群中的專家程度，並且反應至推薦系統的加權係數中，以提昇推薦的精確度；後續我們將對這些相關研究作進一步的探討。

### 2.1 推薦機制

日常生活中我們在做決定時，通常會參考外部的建議後，才做出最後的選擇；舉例來說，在決定看哪部電影或者是買甚麼車時，我們可能會上搜尋引擎查詢別人的評價或者是參考此領域專家的文章，而做出最適當的決定。不過當能參考的資訊過多的時候，反而不知道哪些資訊是有效，也因此有許多技術與研究用來協助我們在巨量資料中取得所需的資訊，而推薦機制就是其中一個經常被使用的輔助方式。一般的推薦機制主要是以資訊過濾為基礎，利用歷史經驗來分析使用者的行為模式，進行喜好的預測來降低資料過載的情形。在一個推薦系統中，將會依據使用者的操作習慣，協助使用者在眾多物件篩選出候選的物件，然後推薦符合使用者需要的資訊。一般的推薦系統主要分為三種類型：

- 內容式資訊過濾推薦系統      Content-Based Filtering Systems
- 協同式資訊過濾推薦系統      Collaborative Filtering Systems
- 混合式資訊過濾推薦系統      Hybrid approaches

不同的推薦系統著重於解決不同的推薦問題，而相關研究的探討範圍包括：資料取得的方式、推薦系統的應用領域、推薦方法的革新等。下面將個別作說明：

- 內容式資訊過濾(Content-based Filtering)[9][20]：系統首先會收集使用者資訊，像是該使用者曾經瀏覽過的網頁及這些網頁的屬性(例：關鍵字或類型等)，並分析過去使用者的資訊，每個項目皆會有屬於自己的屬性標籤，而所有項目集合起來就是一個使用者的喜好主檔。在進行推薦的時候，我們會依據各項目相對於使用者的主檔資訊作匹配，判斷哪個項目最適合做為推薦。此過濾方式通常運用文件自動分類的方法，並適合用在處理半結構化的文件上，且最常被使用在有限屬性的項目上，像是書籍與電影上。
- 協同式資訊過濾(collaborative Filtering)[5][7]：依據擁有共同經驗之群體的喜好，來推薦使用者感興趣的資訊。利用合作的方式提供個人對於資訊的評分，而評分不一定只能侷限於感興趣的部分，不感興趣資訊的部分也相當重要。此方式在進行推薦時，首先依據使用者對於項目的評比建置使用者的主檔，然後利用統計或者是機器學習的方式來建立各使用者所屬的群集，我們稱之為「最同好群 (Nearest Neighbors)」，最後藉由同好群集內的成員所喜好的項目作推薦，過程中會結合權重的觀念來產生有順序的推薦項目。
- 混合式資訊過濾 (Hybrid approaches)：結合 Content-based Filtering 與 collaborative Filtering 兩種方式，即本研究採行方式。

在本研究中，我們將會利用內容式資訊過濾的方式，找出相似度一致的項目做為推薦的候選項目；此外我們在協同過濾上我們將會利用社群網絡分析的技術，對使用者進行同質資訊需求的分群取代一般的個人背景與使用者喜好的收集，我們將會找出候選項目中哪幾個項目，在相同子群內的使用頻率或使用人次最高做為推薦的標的。下列我們將先介紹幾個在 OLAP 推薦上相關的研究，後續將會對社群網絡分析上的相關研究作說明。

### 2.1.1 OLAP 推薦機制

線上分析處理 (On-Line Analytical Processing, OLAP) 是一種提供使用者存取資料倉儲的前端應用，協助決策者快速且有效率的從巨量資料中取得決策所需的資訊。藉由資料模型(Cube)的建置，提供決策者可以利用多維度的觀點，依據不同的主題與面向來對資料進行分析。在資料操作上，利用維度階層上向下探勘(Drill-Down)與向上彙總(Roll-Up)



的功能，有效率的切換資料的面向且取得一致性的決策資訊。在 OLAP 系統中主要分為兩個部分：

- 維度(Dimension): 檢視一個問題時所用的分析面向，如時間、區住地與學歷等，通常都是描述性的項目。而維度的使用方式又分為報表維度與條件式維度。報表維度指的是我們一般放置於報表的行列資訊，主要是用來呈現此維度各成員的量值彙總資訊，提供的操作方式包含鑽研(Drill-Down 與 Roll-Up)與樞紐(Pivot)等。而條件式維度，主要是用來界限報表呈現的內容，例如檢視 2009 年北美的銷售數字中的『2009 年』與『北美』，主要的操作方式是切面(Slice)。
- 量值(Measure): 檢視一個問題時所用的彙總資訊，如金額、次數與庫存等，通常都是量化的項目。

線上分析處理系統為了能提供系統操作者取得全面性的資料，通常在建置的過程中，都會將分析問題所需的維度與相關屬性以及各主題所需分析的量值全都納入系統內，用以提供決策者可有效率的檢視企業 KPI。但也因為這樣的特性，通常會導致新手使用者發生難以下手的情況。因此，有許多的論文就對於如何在 OLAP 系統上提供使用者協助。下列幾篇論文也是在對於 OLAP 系統上輔助方式進行探討。

在 Carsten Sapia(1999)的研究[9]中，作者利用內容式資訊過濾(Content-based Filtering) (Balabanovic & Shoham, 1997) 方式做使用者行為的預測，Carsten Sapia 認為 OLAP 系統上每一個查詢都代表著一種商業問題，而在相似商業的問題處理時會有相似的模式(Patterns)。因此他將 OLAP 上的操作(MDX 查詢語句)正規化為集合的表示式，做為此查詢的 Query Prototype。作者將正規化後的集合，經由『維度的層級(Level)差異』與『維度的功能(Result 或 Selection Dimension)』兩個面向，計算兩個查詢語句的相似度。最後再以相似度的判斷結果，預測該使用者下一步操作。不過在此研究中，存在因忽略認同度而產生『最相似的查詢等是最適合推薦的項目』的錯誤，單純以資料本身的相似度作為判斷，有可能此查詢本身是錯的，但只因它最相似而誤認為最適項目。因此在本研究中會先以相似度高低找出候選的項目，然後進行認同度進行投票後，找出最適合的推薦項目來改善此問題。

在 Chen & Hsu (2008)的研究[20]中，作者進行生手推薦機制、老手推薦機與分群推薦機制的比較。作者認為 OLAP 系統上的使用者應該有生手與老手間之分，生手在此研

究中將考慮在所有的使用者的操作紀錄，計算出使用頻率最高的三個項目進行推薦。但隨著使用者在系統操作經驗的累積，使用者在系統上的推薦需求反而是以往不曾使用過的項目，以提升使用者對於系統功能的使用率。作者利用 K-Means 分群演算法依據使用者行為的相似性進行分群，提升推薦分析資料的精確度。

不過此研究主要是推薦各個 Cube 使用頻率最高的前三名項目，不過實際上 OLAP 系統的操作是一個序列的操作，需要經由多個相關的項目組合才能取得一個有效的結果。但是在同一個 Cube 上使用頻率最高的三個項目組合起來，絕大部分不會是我們問題解決所需的資訊，因此在本研究中，我們依據使用者當下的操作內容的相似度判斷，並且整合社群關係的推薦機制來找出最適合的推薦項目。

### 2.1.2 網頁推薦機制

隨著網際網路的發展與網路頻寬的普及，人們可以快速獲取所需資訊的管道日益擴增。但伴隨著網站數目大幅度的成長，如何在巨量資源中找到有價值的資訊就變成最迫切的議題；也因有搜尋引擎的發明，藉由關鍵字的查詢有效的縮小資料範圍。搜尋的過程中，使用者會輸入所需資訊的關鍵字，但由於網路資源實在太多，且單一關鍵字通無法明確描述出真的資訊需求。因此搜尋引擎查詢的過程中，通常會藉由演算法計算各頁面符合的機率並且排序後呈現結果，避免查詢的結果發生方向偏離的情況。

使用者評分 (User Rating) 的方式是一種經常見的推薦方式，是以使用者的角度來進行推薦。藉由使用者上傳對於各項目的評分後，經由公式的計算統計評分的結果。像是雅虎拍賣網站，在每次交易後讓買家與賣家互相評分，協助後續其他人在交易時很容易知道對方的信任評分。但是此種方式僅限於在於票選項目很少的時候，才可以讓使用者對於全部項目進行評分，然後再依據評分結果進行推薦。但在網際網路這樣可能的候選項目是幾千幾萬個的情況下，要讓使用者對於每個項目進行評分幾乎是不可能。因此，就有許多論文研究各種演算法，依據使用者的歷史操作紀錄，模擬每個人對於各網頁的喜好進行評分，最後再依據統計的結果進行推薦。

### 2.1.3 Page Rank 演算法

由 Google 創始人 Brin 和 Page (1998) 所提出的 Page Rank [11] 便是其中最為著名的一個評分演算法，此方法將查詢結果依據實用性作為排名的一個主要因素，用以表現網頁的相關性與重要性。Page Rank 演算法中具有兩個特性：

- 頁面連結越多代表此網頁越重要：利用網頁間連結的特性，每一個頁面的連結都代表使用者對於此頁面投一認同票，因此連結越多代表此頁面重要性越高。
- 如果此網頁被重要的網頁連結時，則此網頁也是很重要的網頁。

此研究利用投票方式取代傳統以資料出現頻率判斷項目的好壞，以往在判斷推薦項目的適合程度時，都只利用『點選的次數』作分析，這樣的方式卻很容易造成只以『點』的方式判斷推薦項目的好壞。而在 OLAP 系統上的操作與網頁上的瀏覽行為相似，都存在以 Session 為導向 [10] 的特性 (Newman, 2006)；因此我們在設計 OLAP 推薦輔助機制上亦可參考 Page Rank 演算法中的兩個特性，可以透過連結關係進行認同投票，將同情境下操作人次最多的項目視為最適合推薦的標的，此外我們亦會依據每個人的老手程度，對使用者的認同票進行加權，用以取代以往一人一票的方式。

## 2.2 社會網絡分析

社會網絡分析 (Social Network Analysis, SNA) 的研究，主要是在探討人與人之間互動關係 (Newman, 2003; Wasserman & Faust, 1994) [15][16] 所建立的社會結構對於特定個體所產生的影響，以及關聯間存在的『潛在結構』 [3] (Latent Structure)。在社會網絡中，主要是由『節點 (Node)』、『關聯紐帶 (Relation)』、『連結 (Link)』三個要素組成。

- 節點：亦稱為社會個體 (Social Entities)，它是社會網絡中的最小單位，例如：個人、部門、公司、國家..等；而節點在不同的關係型態下，可以同時屬於不同的網絡。在本研究中，我們會將在 OLAP 系統上的每一個使用者視為此網絡中的節點。
- 連結：亦稱為社會整合程度，說明的是社會個體與網絡關聯的互動程度與類型，通常包含關係的強度、方向等。在本次研究中，我們會利用推薦採納的結果來建立使用者間的連結，當今日使用者 A 接受我們的推薦項目時，我們會針對在候選項目篩選時關聯的使用者與使用者 A 間建立一個正向的連結，反之我們也會建立一個表示為此使用者參考這些使用者的一個反向連結。

- 關聯紐帶：社會個體間因為某種連結而產生的互動，並且因同質性的特徵造成影響；而同質性一般可從內容來定義。例如：因在同一個部門一起工作所構成的同事關係與常見因血緣或是婚姻所造成的情感關係。在本次的研究中，我們將會利用資訊需求的相似程度建立起一個虛擬部門的關聯。

在社會網絡中，我們通常會用密度(Density)來描述網絡中節點彼此之間的互動狀況。在 W.D. Nooy (2005)的凝聚子群分析研究[16]中，密度越高的群體代表節點關係越密切，也表示著此子群內的節點因為某種的關聯而形成一個社群。對於社群的詮釋，可能會因不同研究目的而有所不同。參考 M. E. J. Newman (2006)提出的研究[18]，我們認為在 OLAP 系統上操作行為相仿者，他們在實際的企業組織屬同一部門的機率比較大，相對其操作對於推薦上的幫助比較有效益，因此在本次的實驗上我們會將行為相似的人歸屬於同一社群。在大部分社群網絡的研究中[4]，都會利用網絡中心性[2](Brass & Burkhardt, 1992)來判斷節點在『潛在結構』的網絡重要性，節點在群組中的中心性越高相對代表具有越高的影響力。Freeman(1979)的研究[1]則指出代表中心性概念的三種形式為程度中心性、中介中心性及接近中心性。程度的中心性主要是在觀察網絡中節點間活絡的程度，存在越多關聯的節點表示其非正式權力與影響力相對也高；接近中心性主要是要衡量出社會網路的全域中心性，並藉此判斷此節點與其他節點的接近程度，與其他節點的總距離越短表示此節點的接近中心性愈高；在中介中心性中，主要是在觀察節點間的關聯間是否都存在以某個節點作為中介的橋樑，該節點在社會網絡中通常扮演著資訊串聯的角色，掌握著資訊流通的關鍵位置。在後續的研究中，我們將會利用社會網絡分析的分支度(Degree)中心性，利用操作紀錄被其他人參考的次數來找出哪些人在多維度分析平台中屬於老手，並且在進行推薦的時候進行關聯加權。



## 第三章 研究方法

本章為問題定義與研究方法的說明，一共分為三節。第一節，將會就在此次研究過程中所提出的假設作說明；第二節，將會對本研究所提出的推薦系統架構作介紹；第三節與第四節，我們將會針對此次研究過程來源資料的收集與預處理作說明。後續我們會將正規化後的資料做為後續社群網絡分析與操作推薦機制使用，最後，我們會在第六章中對提供的推薦系統進行實驗分析與驗證。

### 3.1 研究假設

首先，本研究將先針對多維度分析系統之特性進行說明，以作為本研究所提操作推薦機制之前提，而本研究在建立線上分析操作推薦模型時，也將以下列假設作為模型資料建立的基礎。

1. 操作的順序不影響推薦的結果：在 OLAP 系統上的操作，雖然每筆記錄都是一個有順序性的結果，但在本次研究中在進行相似度的判斷時將會忽略此順序性。因為在 OLAP 的系統中，他對於『維度-性別』、『維度-年齡』、『量值-平均消費金額』三物件進行下列的操作都可取得各年齡層的男女消費能力的分析的結果：

- 選擇『維度-性別』、選擇『維度-年齡』、最後選擇『量值-平均消費金額』
- 選擇『維度-年齡』、選擇『維度-性別』、最後選擇『量值-平均消費金額』
- 選擇『量值-平均消費金額』、選擇『維度-年齡』、最後選擇『維度-性別』
- 選擇『維度-性別』、選擇『量值-平均消費金額』、最後選擇『維度-年齡』

2. 在線上分析系統上的使用者是有老手與生手之分：在線上分析系統的使用者中，我們認為同領域的族群所作的操作紀錄，應該要比不同族群所作的操作在進行推薦時候來的重要。而在相同的族群內又區分為老手與生手，通常一個族群中老手的比率是比較低的，但是其操作紀錄對於推薦機制上的影響比重卻應該比生手來的更重

要，此因素更應該被凸顯於推薦的結果上。

### 3.2 系統架構

本研究根據多維度分析系統使用者的操作行為，提出一套基於查詢相似度判斷與使用者關聯度判斷為基礎的 OLAP 系統推薦機制。圖 3-1 為本次研究的個人化多維度分析平台推薦的模型架構，此一模型可視為使用者在前端操作介面與資料庫資料中的一個中間層 (Middleware)。

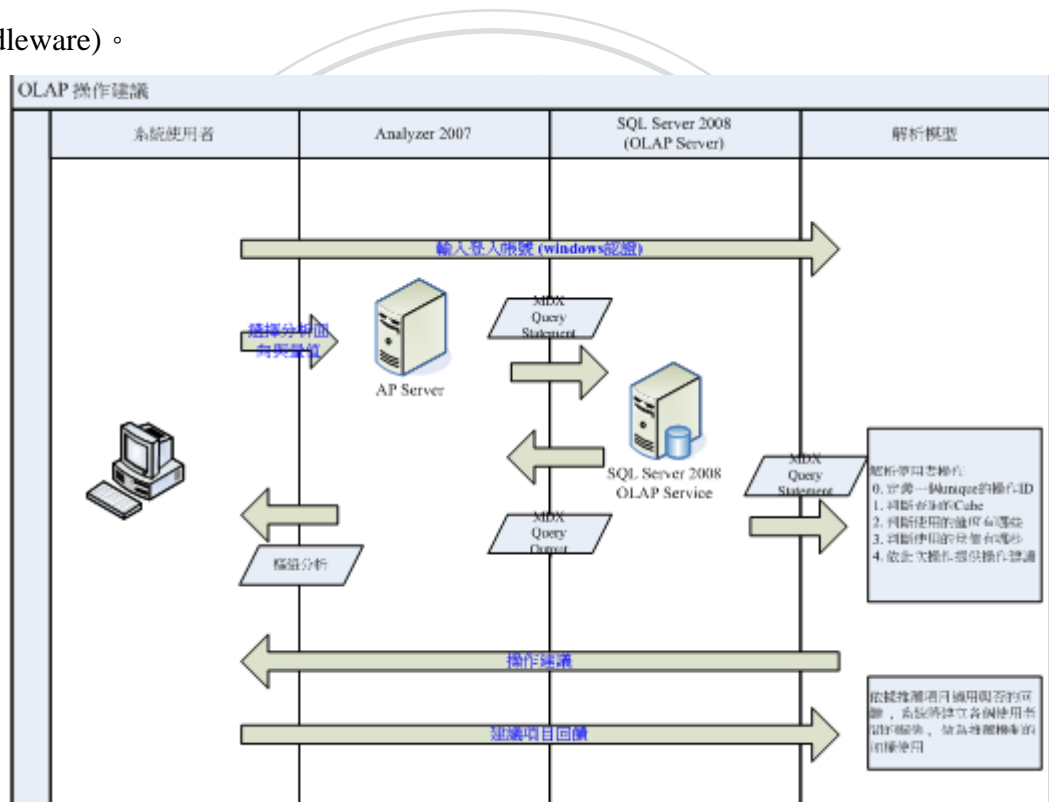


圖 3-1、系統架構圖

系統在使用者對多維度分析系統進行操作的過程中，主動記錄使用者的所有查詢紀錄，並且解析個別的查詢描述式用以進行正規化，並且將結果保存於操作紀錄的資料庫中；系統依據記錄下來的資料，經由相似性與使用者分群與老手程度的判斷，最後將推薦的項目以清單的方式提供使用者作為參考。使用者在取得推薦項目清單後，將會對提供的項目進行系統反饋，用以提供系統進行人員關聯強度的學習，提升後續的推薦準確率。在本章節我們將說明在研究中如何收集分使用者的操作紀錄，並且將每筆記錄依據特定

模式正規化成可用於分析的查詢物件集合。在第四章，我們將依據整理出來的集合作社群網絡的分析，將使用者依據業務的同質性作分群，並且利用中心性作為每個使用者在個別社群中的老手程度進行判斷。最後在第五章，我們將會把利用正規化的結果進行相似度的判斷找出推薦的候選項目，並且利用設計的推薦機制進行各候選項目推薦適合度的判斷。

### 3.3 使用者操作紀錄收集

本研究是以建置於微軟的 Microsoft SQL Server 2008 Analysis Services 服務上的商務智慧方案進行使用者行為分析。在多維度分析的系統上，使用者操作的是如圖 3-2 所示的視覺化介面。



圖 3-2、多維度分析平台操作介面

使用者在多維度分析系統上的所有操作，系統都會轉換成一筆可用於 Cube 上作查詢的 MDX 語句。MDX(Multidimensional Expressions)，是一種對於 OLAP 系統中建立的分析模型進行查詢的語法。它的結構類似我們在一般關聯式資料查詢資料時所使用 T-SQL，但是不同的地方是 T-SQL 對資料的操作是一種只包含行列的二維查詢，但是 MDX 則是利用切片(Slicing)方式對資料做操作。因此在研究過程中我們將會保存每筆操作紀錄的 MDX 查詢語句，以做為分析過程中所需的大量實驗數據。在操作紀錄的收集上，本研

究將利用 Microsoft SQL Server 2008 所提供的 Profiler 功能，記錄每筆使用者在於多維度分析平台上所進行的查詢內容。表 3-1 的資料就是我們所記錄下來的所有使用者在多維度分析系統上所做的操作紀錄，後續我們將會藉由內容的觀察，將每筆資料解析為可做為分析使用的資料。

表 3-1、使用者於多維度分析系統的操作紀錄範例

TextData	ConnectionID	NTUserName	StartTime
SELECT SUBSET([EC_商品資訊].[提報日期].[(All)].ALLMEMBERS, 0, 1) ...	32	GaryChen	2010-04-05 11:54:28.000
SELECT SUBSET([EC_商品資訊].[提報日期].[提報日期_年].ALLMEMBE...	32	GaryChen	2010-04-05 11:54:28.000
SELECT { [EC_產品銷售館別].[銷售館別].[全公司].&[000] }{[EC_產品銷...	34	GaryChen	2010-04-05 11:54:28.000
SELECT NON EMPTY { { { { AddCalculatedMembers([EC_產品銷售館...	27	GaryChen	2010-04-05 11:54:07.000
SELECT SUBSET([EC_產品銷售館別].[銷售館別].[(All)].ALLMEMBERS, 0...	18	GaryChen	2010-04-05 11:53:54.000
SELECT SUBSET([EC_產品銷售館別].[銷售館別].[全公司].ALLMEMBER...	18	GaryChen	2010-04-05 11:53:54.000
WITH MEMBER [Measures].[__PseudoMember] AS '0' SELECT NON EMP...	20	GaryChen	2010-04-05 11:53:54.000
WITH MEMBER [Measures].[業績_訂單數 Pct] AS '([EC_供應商 供應商]...	2526	amiutseng	2009-06-01 21:48:14.000
WITH MEMBER [Measures].[業績_訂單數 Pct] AS '([EC_供應商 供應商]...	2522	amiutseng	2009-06-01 21:48:11.000
WITH MEMBER [Measures].[__CurrentChildCount] AS 'COUNT([EC_產品...	2523	amiutseng	2009-06-01 21:48:08.000
WITH MEMBER [Measures].[__CurrentMemberName] AS '[EC_產品銷售...	2523	amiutseng	2009-06-01 21:48:08.000
WITH MEMBER [Measures].[__CurrentChildCount] AS 'COUNT([EC_產品...	2523	amiutseng	2009-06-01 21:48:08.000
WITH MEMBER [Measures].[__CurrentMemberName] AS '[EC_產品銷售...	2523	amiutseng	2009-06-01 21:48:08.000
WITH MEMBER [Measures].[業績_訂單數 Pct] AS '([EC_供應商 供應商]...	2521	amiutseng	2009-06-01 21:47:50.000

### 3.4 多維度操作紀錄正規化

經由 Profiler 所記錄下來的操作都是一連串複雜的 MDX 語法，無法運用在研究中進行相似度的判斷。因此在本研究中，參考 Carsten Sapi [9] 的作法，將每筆操作記錄的正規化成為一個可運用的項目集合。在本研究中，我們藉由觀察資料庫內的資料屬性，撰寫一隻適當的解析程式來剖析每筆操作結果，將每次的操作紀錄拆分為四個部分：Cube 名稱、維度-報表、維度-過濾條件與量值四個部分，最後再經由簡化的方式將每筆的操作紀錄正規化為一個有限的維度集合(={df,..dr, M})。在此表示式中我們把量值也視為一個特殊的維度。下面是一個 MDX 範例，在後續我們將會依據正規化的方式將此查詢 MDX 進行拆解：



1. 步驟一：分析此次操作的 Cube：由於多維度分析有一個特性，他一次只能對單一的 Cube 操作，而不像是一般的 SQL 的語法可以利用 JOIN 的方法串連多個資料表作查詢，因此我們只要利用 FROM 字串協助定位出此次操作 CUBE 的位置，然後我們可以利用 [ 與 ] 作為識別 CUBE 字串的起始與結束座標點。因此在此範例中，首先我們知道 FROM 的位置是第 95 個字元，那我們將找到的位置加上 FROM 字串的長度 5 後，第 100 個字元就是 CUBE 字串的起始位置，另外我們再從 100 作為起始點找出最接近的 ] 的位置視為結束點(107)，由此兩點的定位我們就可以很明確取得此次操作的 CUBE 名稱為[卡友消費分析]。
2. 步驟二：分析此次操作的量值：觀察多維度分析操作的記錄中，我們發現量值一定是以 Measure 為開頭的字串，因此我們可以利用[Measures].找出每個量值的起始位置，接下來我們可以利用 ] 找出各量值得結束位置，以此範例來說[消費人數]的起始位置為 20 與 25。
3. 步驟三：分析此次操作的過濾維度：在多維度分析系統的查詢語句中，過濾的條件式與 T-SQL 相似，會出現於 WHERE 的描述式中，但是它是對維度以 SLICE 方式進行查詢過濾。因此在解析的過程中，我們將會利用 WHERE 字串協助定位出過濾維度的起始位置，接下來我們可以利用維度的形式 [Dimsnsion].[Level].&[Member] 的特性，利用 [、.&[ 與 ] 等字串的位置找出每個過濾維度的起始與結束的位置。以此範例為例，我們利用 WHERE 與 [ 可以找出過濾維度([特約商].[主產業別].&[TT])的起始位置為 119，然後我們可以利用&[與最靠近的] 找出結束點位置為 136，因此我們就可以取得此次操作的過濾維度為[特約商].[主產業別].&[TT]。
4. 步驟四：分析此次操作的 報表維度：在多維度分析系統中，報表維度通常是以下列兩種形式出現
  - [Dimsnsion].[Level].&[Member]
  - [Dimsnsion].[Level].[All].ALLMEMBERS且報表維度也一定會出現於 WHERE 前面，因此我們可以利用此兩個特徵，分析得到此操作中的報表維度為[清算日].[清算日 年].&[2008 年]與[清算日].[清算日 年].&[2008 年]。
5. 經過系統的解析後將會拆分為下列四類型的物件
  - Cube 名稱：[卡友消費分析]

- 報表維度名稱：
  - [清算日].[清算日 年].&[2008 年]
  - [清算日].[清算日 年].&[2009 年]
- 過濾維度名稱
  - [特約商].[主產業別].&[TT]
- 量值名稱：[消費人數]

但是在推薦機制的相似度判斷中，所有的操作內容都是百分之百相同的情況是非常的低的。因此系統將會再進一步的此結果進行簡化，我們將維度中特定成員(member)的資訊給刪除後，做為正規化後的最小單位，並且將每個查詢式以下列表示式呈現

$$p = (\text{Cube}, Ds = \{s_1, \dots, s_i\}, Dr = \{r_1, \dots, r_j\}, Mm = \{m_1, \dots, m_k\})$$

$Ds$  表示此次操作所包含的過濾維度集合

$Dr$  表示此次操作所包含的報表維度集合

$Mm$  表示此次操作所包含的量值集合

因此此次的解析結果正規化後將如下

- Cube 名稱：[卡友消費分析]
- 報表維度名稱：[清算日].[清算日 年]
- 過濾維度名稱：[特約商].[主產業別]
- 量值名稱：[消費人數]

$p = ([\text{卡友消費分析}], \{[\text{清算日}].[ \text{清算日 年}]\}, \{[\text{特約商}].[ \text{主產業別}]\}, \{[\text{消費人數}]\})$ ，這樣的項目集合也將會作為後續相似度判斷計算時所使用。

## 第四章 社群網絡分析

在使用商業智慧平台的社會網絡中，使用者就是此社會網絡中的節點，而網絡中各種關聯所建立出來的連結將會影響到推薦的方式與準確性。在商業智慧平台系統中，相同部門的使用者存在相同的背景與資訊需求，產生的推薦項目也相對較能符合期望。在本研究中，為了增加此因素在於 OLAP 推薦機制的通用性，我們依據推薦反饋所建立的連結作為網絡中的連結，建立取代原本組織關係的虛擬人際關係。圖 4-1 即是我們依據此連結關聯所產生的社會網絡圖形。在此圖中連結上的數字表示的是兩節點之間的關聯程度，程度越高的表示的是兩者之間的同質性越高，其屬於相同部門的可能性越高。

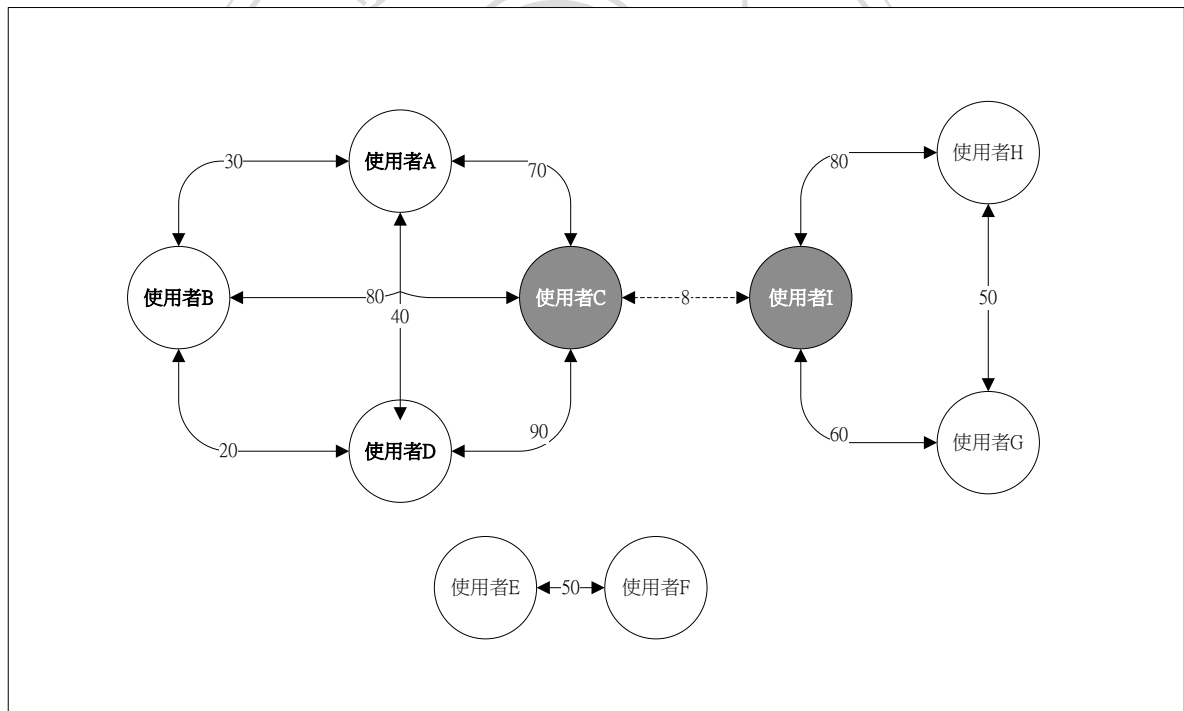


圖 4-1、社會網絡圖形

### 4.1 凝聚子群

表 4-1 中，我們對每個使用者依據在集團中相對應的實際組織單位進行編碼，例如 EC 開頭的表示為電子商務部門；然後我們從錄製的資料中找出兩千筆資料，逐筆拿出來跟另外的一萬四千筆操作紀錄作比較，當有完全相似操作的人時，系統會自動將兩個使用

間增加一條連結，而所有人之間建立的連結結果統計於關聯分佈矩陣表上，我們可發現  
在相同部門的使用者之間，於推薦上的採納程度會比不同部門間的推薦高。

表 4-1、虛擬關聯分佈矩陣

加總 - RELAT	EC_PU_IV	EC_PU_V	EC_PU_VI	EC_PU_VIII	EC_PU_XIII	FA_PU_I	Merchant_PU_I	Merchant_PU_II	Merchant_PU_III	Merchant_PU_IV
操作者										
E_PD_PU_I							5	4	20	8
E_PD_PU_III										
EC_PU_A	11	17	13			9	3			
EC_PU_C1	179	43	83	70		34				
EC_PU_C2	22	16	10	8		1				
EC_PU_C3	65	9	38	33		5				
EC_PU_C4	60	26	53	85		8				1
EC_PU_C5	30	10	3	27		10				
FA_PU_I		1				96	1		3	
Merchant_PU_I							14	12	8	
Merchant_PU_II							1	31	1	
Merchant_PU_III									1	
Merchant_PU_IV										6

在本研究中，我們利用 UCINET 對系統使用者進行虛擬組織的分群，並且在社群中找出相對的老手。本研究執行的步驟如下：

- 計算網絡中最適合的社群數：依據 Newsman and Girvan 的方法，計算在符合最小連結數(Ties)情況下各社群數的 Modularity Q 值，找出 Modularity Q 最大的做為最適合的分割社群數。
- 社群切割(Subgroups)：我們依據 Girvan and Newsman 演算法，利用邊中性的結果區分出 N 個社群。
- 識別社群中的老手：依據中心性的分析結果，我們採用分支度中心度〈In-degree〉來判斷各節點在於社群中相對的重要性與影響力。

上述步驟的進一步說明如下：

#### 4.1.1 Modularity Q 的計算

Newman and Girvan(2006)提出來的模組性(Modularity)[2318]是社群網絡中用來表達的一種結構特徵，可用來衡量網路結構品質的一個指標。其核心概念為社群內的包含大量的關係連結，而在社群間存在少數的連結。因此假設一個社群模組性(Modularity)越高表示社群內的節點存在高關係強度，相對是好的分割。Modularity Q 的基本定義如下：

- 今日有一個 n 個節點的社會網絡， $A_{vw}$ 表示的是此相鄰矩陣內的元素，

$$A_{vw} = \begin{cases} 1; & \text{如果節點 } v \text{ 與節點 } w \text{ 之間有連結，且節點 } v \text{ 不為節點 } w \\ 0; & \text{其他} \end{cases}$$

- 在社群網路分群過程中，各個節點都有一個所屬的社群，我們以 $C_v$ 、 $C_w$ 分別表示節點 $v$ 與節點 $w$ 所屬的社群，因此兩個有連結的節點會落在同一個社群的機率如下：

$$\frac{\sum_{vw} A_{vw} \delta(C_v, C_w)}{\sum_{vw} A_{vw}} \quad \text{其中 } \delta(C_v, C_w) = \begin{cases} 1; C_v = C_w \\ 0; \text{其他} \end{cases}$$

- 假設我們以 $m$ 表示網路中所有關聯的個數，則 $m = \frac{1}{2} * \sum_{vw} A_{vw}$ 。因此我們可以將原本的公式調整如下

$$\frac{\sum_{vw} A_{vw} \delta(C_v, C_w)}{\sum_{vw} A_{vw}} = \frac{1}{2m} * \sum_{vw} A_{vw} \delta(C_v, C_w)$$

- 假設 $K_v = \sum_n A_{vw}$ 表示為節點 $v$ 的分支度(Degree)中心性，也就是說節點 $v$ 有 $K_v$ 個連結。假設網路是隨機的情況下節點 $v$ 與節點 $w$ 的在同一個網路的機率為 $\frac{K_v * K_w}{2m}$ ，因此在隨機網路中的Modularity為：

$$\frac{1}{2m} * \sum_{vw} \frac{K_v * K_w}{2m} \delta(C_v, C_w)$$

- 通常我們會將分割社群的Modularity與隨機網路的Modularity比較，判斷在此社群組合下結構特性的程度。因此我們將原本Modularity的定義調整後如下：

$$Q = \frac{1}{2m} * \sum_{vw} [A_{vw} - \frac{K_v * K_w}{2m}] \delta(C_v, C_w) \quad (1)$$

如果 $Q$ 為0的時候，表示社群內的關聯數與隨機網路一樣，相對的其結構特性不強。根據下列的計算結果，我們可以發現在此網路中最適合組織社群叢集數目為3。

Partition w/ 3 clusters:  $Q = 0.216$

Partition w/ 4 clusters:  $Q = 0.202$

Partition w/ 8 clusters:  $Q = 0.099$

Partition w/ 9 clusters:  $Q = 0.066$

### 4.1.2 社群切割(Subgroups)

在此研究中，我們對 UCINET 提供的子群分割演算法 Factions 與 Girvan-Newman 演算法的作比較：

- Factions 演算法：是使用 Tabu 搜尋方法進行社群的辨識，所依循的原則是尋求集合內各個行動者之截面(profile)的最小「組內平方差和」(Moody,2001 )，圖 4-2 即是 Factions 演算法的分群結果。

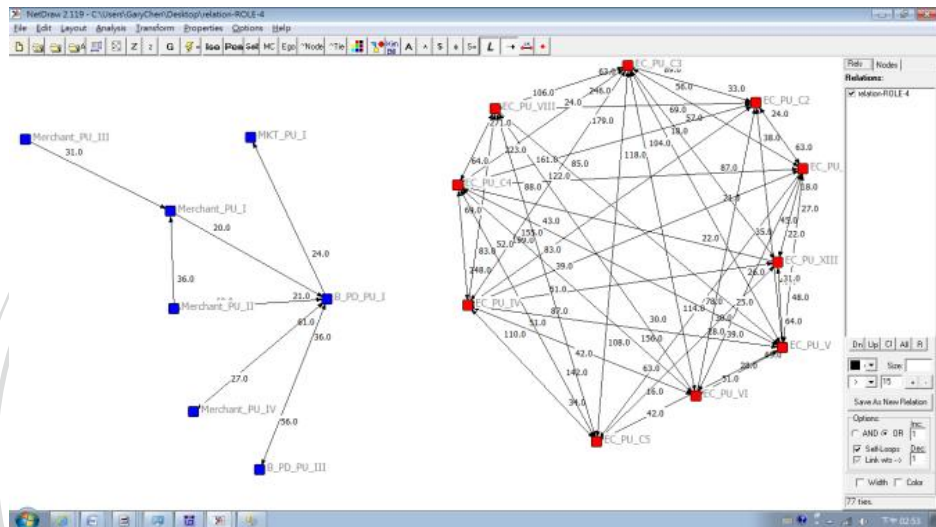


圖 4-2、Factions 演算法分群結果

- Girvan-Newman 演算法：利用邊中介性(edge betweenness centrality)的特性，在一個網絡中假設節點之間的最短路徑都要經過某一個邊才能到達時，移除此邊即可很容易取得一個有效的分割結果，圖 4-3 即是 Girvan-Newman 演算法的分群結果。

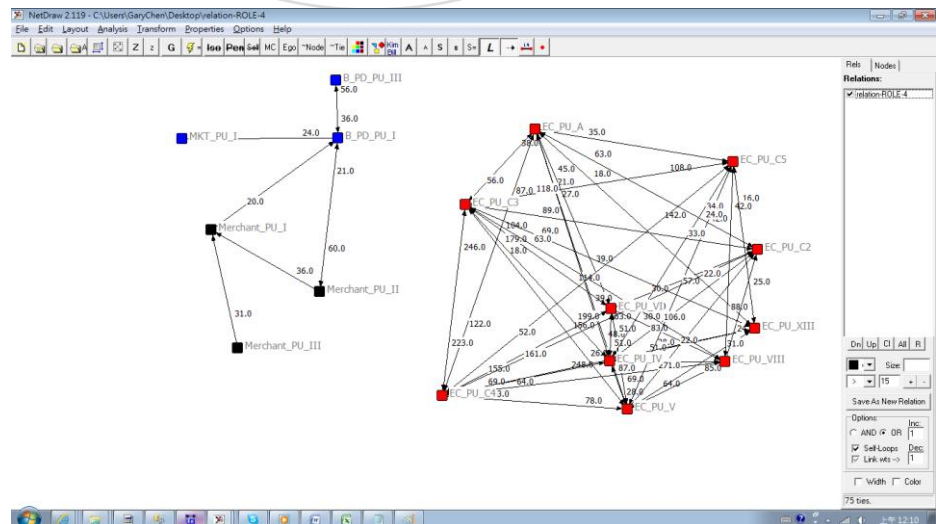




圖 4-3、Girvan-Newman 演算法分群結果

由結果的比較，我們可以發現利用 Factions 作組織社群的分派上，只能夠有效的識別電子商務與實體通路的這種比較大分類的使用者類型，但是 Girvan-Newman 演算法中，我們在實體通路上，進一步能有效區分特約商管理部門與經營管理部門的使用者，因此在此次研究，我們將會利用 Girvan-Newman 演算法作為分群依據。

下列是邊  $e$  中介性的計算公式：

$$C_b(e) = \sum_{s \neq v \in G} \frac{\sigma_{st}(e)}{\sigma_{st}} \quad ; e = \{x, y\} \quad (2)$$

$\sigma_{st}(e)$  表示的網絡中所有節點  $s$  到節點  $t$  的最短路徑中，經過邊  $e$  的加總個數，因此計算過程中，

$$\sigma_{st}(e) = \begin{cases} 0 & , \text{if } d_{st} < d_{sx} + 1 + d_{yt} \\ \sigma_{sx} \times \sigma_{yt} & , \text{otherwise} \end{cases} \quad ; d_{sx} \leq d_{sy} \quad (3)$$

$d_{st}$ ：表示節點  $s$  到節點  $t$  的最短路長度

$d_{sx}$ ：表示節點  $s$  到節點  $x$  的最短路長度

$d_{yt}$ ：表示節點  $y$  到節點  $t$  的最短路長度

$\sigma_{sx}$ ：表示節點  $s$  到節點  $x$  的最短路徑數

$\sigma_{sx}$ ：表示節點  $s$  到節點  $x$  的最短路徑數

因此在分群的過程中，將依據下列的步驟進行：

1. 步驟一：計算網絡  $G$  中所有邊的中介性
2. 步驟二：移除中介性最高的邊，將網絡  $G$  拆分為  $G'$  與  $G''$
3. 步驟三：重複步驟 1 與步驟 2，直到滿足 Modularity  $Q$  計算出的最適合社群數

## 4.2 老手程度的判斷

在商業智慧平台的使用上，有一些人在部門中相對重要且具影響力，通常這樣的人都是部門中對於系統使用比較久且比較了解系統功能的人。因此這樣的人在於推薦的判斷上，其操作相對於部門內新人應該來的更重要。而在網絡中通常我們都利用分支度 (Degree) 中心性、緊密度 (Closeness) 中心性、中介 (Betweenness) 中心性三種指標來評估一

個節點的重要性。

- 分支度(Degree)中心性：主要是利用每個節點的連結數目作為中心性的判斷，假設有一個節點與其他節點的連結越多的時候，表示該節點在網絡中相對於其他節點更為活絡，相對其擁有較高的中心性。假設一個社會網絡的節點個數為  $N$  時，我們計算節點  $v$  分支中心性的公式如下：

$$C_d(v_j) = \frac{\sum_{i=1}^N \alpha(v_i, v_j)}{(N-1)} \quad \forall i, 1 \leq i \leq N \quad (4)$$

$$\alpha(v_i, v_j) = \begin{cases} 1, & \text{節點 } i \text{ 與節點 } j \text{ 之間有連結} \\ 0, & \text{節點 } i \text{ 與節點 } j \text{ 之間無連結} \end{cases}$$

Degree 可依據連結的方向分為 In-Degree 與 Out-Degree，In-Degree 代表的是推薦時節點 A 被其他節點參考到的連結數，而 Out-Degree 表示的是推薦時節點 A 參考其他節點的連結數。In-Degree 越高的節點所表示該節點在社群中的認可度相對高，該節點可能屬於部門中的老手，因此在系統進行推薦時經常會被其他人採納。而 Out-Degree 越高且分散在許多節點上時，表示的是該節點可能是系統的生手，由於對系統功能的不了解，因此常常會嘗試去參考不特定的人操作作為參考，圖 4-4 即是我們在 UCINET 上以 Degree 找出來的中心性結果。

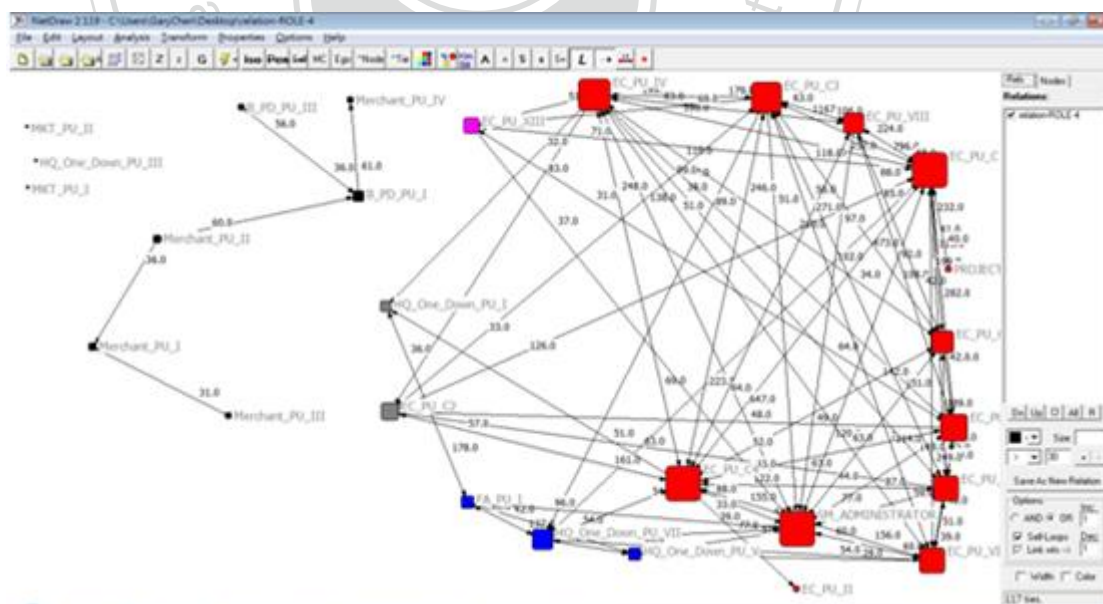


圖 4-4、Degree 的 Centrality measures 結果

- 緊密度(Closeness)中心性：Closeness 著重的是該節點與其他節點的靠近程度，



值介於 0 到 1 之間，值越高表示該節點與越容易到達其他節點，相對表示此節點與其他節點之間的密度高。我們在計算節點  $v$  的緊密度中心性公式如下：

$$C_c(v_j) = \frac{(N-1)}{\sum_{i=1}^N d(v_i, v_j)} \quad \forall i, 1 \leq i \leq N \quad (5)$$

$d(v_i, v_j)$ ：表示節點  $i$  與節點  $j$  最短路徑之連結數

圖 4-5 是我們在 UCINET 上以 Degree 找出來的中心性結果，在結果中發現此指標雖然能區分出哪些人可能會是老手，但是在老手之間的程度卻無法有效的區分出來。尤其當子群內的節點數比較少的時候，彼此間互動所建立的連結數是很平均的，因此會容易將整個子群的人都視為程度相當的老手或生手。

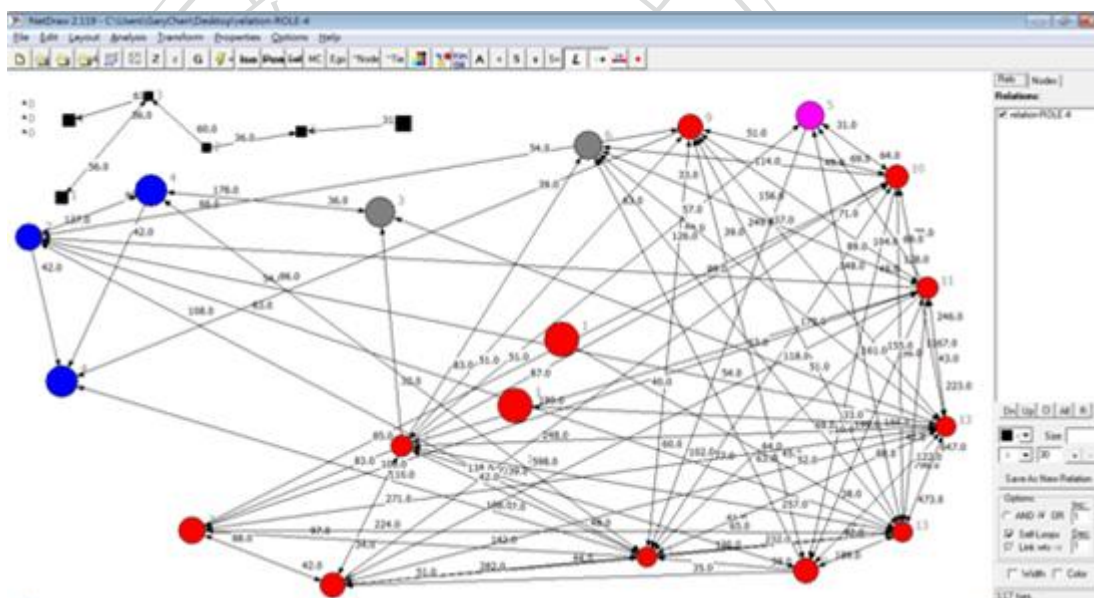


圖 4-5、Closeness 的 Centrality measures 結果

- 中介(Betweenness)中心性：在網絡中任兩個節點間一定存在一個最短路徑 (geodesics)，假設一個節點經常在最短路徑上面表示越多節點需要經由他才能到達其他節點。通常這樣的節點扮演的是橋樑的角色，其 In-Degree 與 Out-Degree 的值會很平均。計算節點  $v$  的中介中心性公式如下：

$$C_b(v_j) = \frac{\sum \beta(v_i, v_k) / \delta_{ik}}{(N-1)(N-2)} \quad (6)$$

$\beta(v_i, v_k)$ ：表示任兩個節點間的最短路徑中經過  $j$  的次數

$\delta_{ik}$ ：表示任兩個節點的最短路徑的數量

N：表示網絡中節點的個數

圖 4-6 是我們在 UCINET 上以 Betweenness 找出來的中心性結果。在商業智慧平台上，老手通常對於系統的操作與功能上都很了解，相對 Out-Degree 的數值通常都會偏低。如果採用此指標通常會讓老手的中心性降低，因此在此研究中將不考慮此指標。

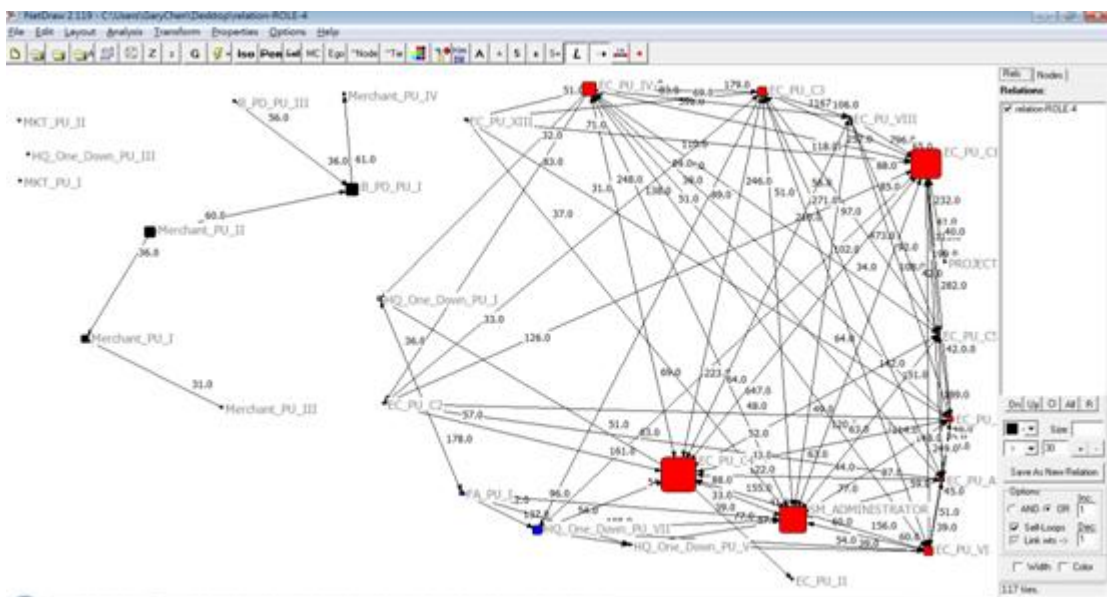


圖 4-6、Betweenness 的 Centrality measures 結果

最後我們將三種中心性的結果與利用同部門使用者互評的結果比對後，發現 In-degree 的結果是最符合我們在商業智慧平台上做為老手程度判斷的指標，他在每個子群中可以有效的識別使用者對於系統的熟悉度。因此在後續的推薦機制上我們將會以此方式作為投票加權的判斷。

## 第五章 推薦機制

在第三章，我們將蒐集回來的資料正規化處理整理成可供分析用的集合，在本章我們將藉由相似度的判斷與推薦機制設計提供使用者在系統操作過程所需的建議。多維度分析系統上的推薦機制通常分為兩種類型：內容相似度推薦與同好推薦。我們將同時考慮內容過濾與協同過濾兩種技術，提供混合式的推薦機制。在內容過濾方面，我們將會利用相似度的判斷縮小候選項目的範圍，後續我們將會依據支持度(Support)與最大使用人次選擇兩種推薦機制挑選出機率越高的項目作為推薦；最後我們也將會把社群網絡分析的結果反應至兩個推薦機制上。而在第六章中，我們也將會以兩個推薦機制的結果進行比較。

### 5.1 相似度判斷

如果我們每次在作推薦時，都把所有的操作紀錄都納入考慮時，將可能因為資料量大造成系統的效能不彰。因此在資料的過濾上，我們會將正規化後的操作項目集合，利用相似度的計算公式，取出與此次操作相同且操作項目數大於此次操作的資料作為推薦用的候選集合。由於相似度主要是利用涵蓋度 (inclusion ratio) 的概念作判斷，依據 MDX 查詢語句  $\rho_1$  與  $\rho_2$  在於 Cube、維度與量值上差異作為相似度的判斷，因此計算出來的相似度是不可逆的。舉例來說當查詢  $\rho_1$  是查詢  $\rho_2$  的子集合時，查詢  $\rho_2$  相對於查詢  $\rho_1$  是完全相似的，但是查詢  $\rho_1$  相對於查詢  $\rho_2$  是部分相似的。Cube 操作的差異 ( $\Delta C_t$ )：在多維度分析平台上我們一次只能對一個 Cube 進行資料的操作，因此假設使用者存取的目標 Cube 不同時，此兩筆的操作是無相關的，因此操作的 Cube 是否相同是決定兩個查詢是否相似的最大要素。

$$\Delta C_t = \begin{cases} 1, \rho_1 \text{ 與 } \rho_2 \text{ 操作相同的 Cube} \\ 0, \rho_1 \text{ 與 } \rho_2 \text{ 操作不同的 Cube} \end{cases}$$

查詢 $\rho_2$ 相對於查詢 $\rho_1$ 在於操作內容的涵蓋率判斷上，會先計算查詢 $\rho_1$ 中包含的操作項目個數做為分母；分子部分我將依據兩查詢在過濾維度( $D_s$ )、報表維度( $D_r$ )與量值( $M_m$ )三個部分的操作相同個數，因此假設查詢 $\rho_2$ 包含查詢 $\rho_1$ 中所有的操作項目時，我們會認為查詢 $\rho_2$ 完全相似於查詢 $\rho_1$ 。下列為系統用以判斷查詢 $\rho_1$ 對於查詢 $\rho_2$ 之間涵蓋率的公式：

$$|\rho_2 - \rho_1| = \Delta C_t * (\Delta D_s + \Delta D_r + \Delta M_m)$$

$$\Delta D_s = |(D_{s1} \cap D_{s2})| / (|D_{s1}| + |D_{r1}| + |M_{m1}|)$$

表示的是兩個查詢在過濾維度( $D_s$ )間的不同個數

$$\Delta D_r = \frac{|D_{s1} \cap D_{r2}|}{(|D_{s1}| + |D_{r1}| + |M_{m1}|)}$$

表示的是兩個查詢在報表維度( $D_r$ )間的不同個數

$$\Delta M_m = |(M_{m1} \cap M_{m2})| / (|D_{s1}| + |D_{r1}| + |M_{m1}|)$$

表示的是兩個查詢在量值的相同個數

下列是我們利用三個正規化後查詢，計算相對於查詢( $\rho_1$ )相似度的例子

$\rho_1 = ([卡友消費分析],$

$\{[清算日],[清算日 年],[供應商].[區域]\},$

$\{[產品別].[產品類別],[卡友維度].[年收入級距]\},$

$\{[消費人數]\})$

$\rho_2 = ([卡友消費分析],$

$\{[清算日],[清算日 年],[供應商].[區域],[供應商].[區域]\},$

$\{[產品別].[產品類別],[卡友維度].[年收入級距]\},$

$\{[消費人數],[消費金額]\})$

$\rho_2$ 相對於 $\rho_1$ 的相似度計算如下：

$$|\rho_2 - \rho_1| = \Delta C_t * (\Delta D_s + \Delta D_r + \Delta M_m)$$

$$= 1 * ((2/5) + (2/5) + (1/5))$$

$$= 1 \quad \text{百分之百相似}$$

$p_3 = ([\text{供應商分析}],$   
 $\{[\text{清算日}], [\text{清算日 年}], [\text{供應商}], [\text{區域}]\},$   
 $\{[\text{產品別}], [\text{產品類別}]\},$   
 $\{[\text{消費金額}]\})$

$p_3$  相對於  $p_1$  的相似度計算如下：

$$\begin{aligned}
 |p_3 - p_1| &= \Delta Dc * (\Delta Ds + \Delta Dr + \Delta Mm) \\
 &= 0 * (2/5 + 1/5 + 1/5) \\
 &= 0 \quad \text{相似度零}
 \end{aligned}$$

$p_4 = ([\text{卡友消費分析}],$   
 $\{[\text{清算日}], [\text{清算日 年}]\},$   
 $\{[\text{產品別}], [\text{產品類別}], [\text{卡友維度}], [\text{年收入級距}]\},$   
 $\{[\text{消費人數}], [\text{消費金額}]\})$

$p_4$  相對於  $p_1$  的相似度計算如下：

$$\begin{aligned}
 |p_4 - p_1| &= \Delta Dc + \Delta Ds + \Delta Dr + \Delta Mm \\
 &= 1 * (1/5 + 2/5 + 1/5) \\
 &= 0.8 \quad \text{相似度百分之八十}
 \end{aligned}$$

## 5.2 候選項目篩選

在進行完查詢相似度的判斷後，我們可以從歷史操作紀錄的資料庫中取得許多筆完全相似的紀錄。在第三章中，我們認為順序在於 OLAP 系統上的操作是不會影響最後資訊取得的結果，因此假設今日有一個操作的查詢包含 A、B 與 C 三個項目時，另外有一個操作的查詢包含了 A、B、C、D 與 E 五個項目時，我們會認為包含 A、B、C 與 D 四個項目的操作跟包含 A、B、C 與 E 四個項目的操作都可能會是最適合推薦的下一步，也因此我們會認為項目 D 與項目 E 會是此次推薦的候選項目。因此在實做上，我們會將所有符合的操作項目集合作聯集，並且扣除此次的操作項目後，產生推薦的項目的候選清單。下面是一個候選項目清單的產生範例。假設使用者有一個  $q_1 = ([\text{卡友消費分析}],$



{[清算日].[清算日 年],[供應商].[區域]},{ [產品別].[產品類別],[卡友維度].[年收入級距]},{[消費人數]})，假設在歷史資料庫中相對於  $q_1$  完全相似的操作紀錄有下列三筆：

$q_1 = ([卡友消費分析],$   
 {[清算日].[清算日 年],[供應商].[區域],[卡友維度].[教育程度],  
 [卡片維度].[發卡來源][卡友維度].[年齡級距] },  
 {[產品別].[產品類別],[卡友維度].[年收入級距]},{  
 [消費人數]})

$q_3 = ([卡友消費分析],$   
 {[清算日].[清算日 年],[供應商].[區域],[卡友維度].[教育程度],  
 {[產品別].[產品類別],[卡友維度].[年收入級距],  
 [產品別].[產品金額級距]},{  
 [消費人數]})

$q_4 = ([卡友消費分析],$   
 {[清算日].[清算日 年],[供應商].[區域],[卡友維度].[教育程度],  
 [卡友維度].[教育程度]},{  
 {[產品別].[產品類別],[卡友維度].[年收入級距]},{  
 [消費人數],[消費金額],[消費次數]})

聯集此三個查詢的項目集合為

$q_{all} = ([卡友消費分析], \{[清算日].[清算日 年],[供應商].[區域],[卡友維度].[教育程度],[卡片維度].[發卡來源],[卡友維度].[年齡級距],[產品別].[產品類別],[產品別].[產品金額級距],[卡友維度].[年收入級距],[消費人數],[消費金額],[消費次數]\})$

排除本次操作的項目後，即為本次推薦的所有候選項目。以上述例子中，下列項目[卡友維度].[教育程度]，[卡片維度].[發卡來源]，[卡友維度].[年齡級距]，[產品別].[產品金額級距]，[消費金額]，[消費次數] 都是後續可能的推薦項目。

但是在實際的商業運作上，找出來的可能候選項目的量可能是幾百幾千個，這樣反而會造成使用者因自行過濾有效項目，而造成提供的建議內容無法達到預期的效果。為了解決此問題，後續我們將對過濾出來的項目依據最大信心度與最大使用人次選擇兩種推薦機制，對於每個項目計算一個分數，依據分數的高低決定哪些項目最適合作為推薦。

### 5.3 推薦機制：最大信心度選擇

在 Agrawal 關聯法則[27]中，主要是用來找尋資料庫中資料之間的關係，依據最小支持度(Support)與最小信賴度(C Confidence)的門檻，對資料進行掃描找出符合門檻限制的高頻率項目集合做為關聯的規則，而掃描的過程中將從兩個項目的集合開始，持續的找尋下一級的候選項目集合，直到不再出現新的候選項目集合為止。此推薦機制參考關聯法則中的最大信心度選擇(Most-Confident Selection)觀念而來，主要是以”相同情境下各候選項目被操作過的次數”作為判斷依據，主要是找出使用頻率最高的操作項目作為推薦標的。在此機制中，系統會統計符合下列兩個限制的資料筆數作為分母：

- 操作內容與本次操作內容百分之百相似的
- 候選項目符合最小支持度(Minimum Support)

接下來我們會逐一計算在在此次的操作內容前提下，各個候選項目出現的次數(Support)作為分母，計算每個候選項目的適合推薦的分數。因此，當一個項目信心水準越高的情況，表示此項目的使用頻率高，相對推薦成功的機率會比較高的。下列為最大信心度選擇機制在計算候選項目 b 在操作項目集合為 A 時分數的公式：

$$\begin{aligned} \text{Score}(A, b) &= \text{Confidence}(A \rightarrow b) \\ &= P(b|A) = \frac{\text{Support}(A \cap b)}{\text{Support}(A)} \end{aligned} \quad (8)$$

A：在此次操作項目的集合

b：候選推薦項目

Support(A ∩ b)：表示在包含 A 操作項目集合時，同時也操作過項目 b 的次數

Support(A)：表示在包含 A 操作項目集合的操作次數

## 5.4 推薦機制：最大使用人次選擇

在最大信心度選擇的推薦機制上，我們是利用每個項目的頻率做為推薦的判斷依據，但當某個候選項目的操作頻率是很高不過卻是集中少數人的操作上時，此時推薦的成效是非常不好。因此在最大使用者人次選擇的推薦機制上，我們會認為在推薦上同部門中存在比較多的人對此候選項目的認同會比候選項目被操作的次數來的重要。此機制主要是以”相同情境下各候選項目操作過的人次”作為判斷依據，找出最多人使用的操作項目作為推薦標的。在此推薦方式中，系統會先統計所有候選項目曾經操作過的人數作為分數計算時的分母，接下來系統會將之前所列出來的候選項目逐一去判斷，計算每個候選項目曾經使用過的人數作為分子。因此，當一個項目使用過的人數越多的情況，表示此項目是越多人關注的，相對推薦成功機率會是比較高的。下列為最大使用人次選擇機制在計算各候選項目分數時的公式：

$$\text{Score}(A, b) = \frac{\alpha(A \cap b)}{\alpha(A)} \quad (9)$$

A：在此次操作項目的集合

b：候選推薦項目

$\alpha(A \cap b)$ ：操作過集合 A 與項目 b 的人次

$\alpha(A)$ ：操作過集合 A 的人次

## 5.5 推薦機制與參考關聯加權

在上述兩種推薦機制中，我們認為每個人操作所佔的重要性是一樣的；但實際上，是否相同部門與人員對於系統的熟悉程度都應該影響其在候選項目重要性的判斷。舉例來說，有四個人甲、乙、丙、丁，甲是採購部門的新人，而丁是甲在採購部門的同事，乙與丙是另外一個行銷單位的同事，採購單位會比較著重於銷售商品的成本，但是行銷單位會著重於產品銷售的數量與金額。但由於行銷單位的同仁數相對於採購單位來的多，因此資料庫的操作紀錄也會相對的多；而這樣的情況很容易造成推薦的結果錯誤。而相同的在同部門中，通常會有部分的人對於系統的功能是比較了解的，因此該人員的操作紀錄也應該提高其加權，讓他的操作能有效的反應至推薦結果上。因此，我們將會把凝聚子



群的社群結果與參考關係加權加入原有推薦機制，並且作了下列的改善：

- 相同社群內的使用者，其操作紀錄在於候選項目判斷上才會有效。
- 使用者在該社群中分支度(In-Degree)越高，候選項目判斷上的加權越高。

下列為各使用者加權係數的計算公式：

$$W_i = \begin{cases} 0; \text{使用者 } i \text{ 與推薦對象不屬於同一社群} \\ C_d(v_i) = \frac{\sum_{j=1}^n \alpha(v_i, v_j)}{(N-1)} ; \text{其他} \end{cases} \quad (10)$$

因此，我們將兩個推薦機制做下列的調整：

- 推薦機制 最大信心度選擇調整後公式如下：

$$\begin{aligned} \text{Score}'_s(A, b) &= \text{Confidence}(A \rightarrow b, W_i) \\ &= \frac{\sum_{i=1}^n \text{Support}(A \cap b, i) * W_i}{\sum_{i=1}^n \text{Support}(A, i) * W_i} \end{aligned} \quad (11)$$

$n$ ：表示為曾經同時操作過項目集合  $A$  與候選項目  $b$  的使用者集合

$\text{Support}(A, i)$ ：使用者  $i$  操作過項目集合  $A$  的次數

$\text{Support}(A \cap b, i)$ ：使用者  $i$  同時操作過項目集合  $A$  與候選項目的次數

$W_i$ ：使用者  $i$  的老手加權係數

- 推薦機制最大使用人次選擇調整後公式如下：

$$\text{Score}'_r(A, b) = \frac{\sum_{i=1}^n \alpha'(A \cap b, W_i)}{\sum_{i=1}^n \alpha'(A, W_i)} \quad (12)$$

$n$ ：表示為曾經同時操作過項目集合  $A$  與候選項目  $b$  的使用者集合

$\alpha'(A, W_i)$ ：使用者  $i$  操作過項目集合  $A$  的加權人次

$\alpha'(A \cap b, W_i)$ ：使用者  $i$  同時操作過項目集合  $A$  與候選項目  $b$  的加權人次

$W_i$ ：使用者  $i$  的老手加權係數

## 第六章 系統實作與實驗

本章節將簡略說明實作推薦機制之資料來源、開發環境及程式語言，在 6.2 節裡則概略的介紹多維度分析推薦輔助系統的操作介面，並在 6.3 節我們會將收集的操作資料分成十組進行推薦機制的驗證以及兩種推薦機制的實驗結果。最後在 6.4 節上，我們會將此次設計的推薦輔助系統提供給實際的使用者作使用，藉由問卷與系統操作紀錄的分析來驗證此推薦機制的效果，與收集使用者對於推薦系統的回饋，以做為後續研究使用。

### 6.1 程式語言、資料來源

在多維度分析推薦輔助系統的實作上，所使用的程式語言為 Visual C#與 T-SQL，並使用 Visual Studio 2010 整合式開發環境 (Integrated Development Environment, IDE)與 SQL Server Management 作開發。開發與實驗的作業系統為 Windows 2003 R2 64 位元，CPU 為 Intel® Core™ i3 CPU M 330 2.13GHz，記憶體為 8GB。

在本次實驗中，我們將以某家已經導入多維度平台一年半的公司為研究對象。該公司在導入商業智慧專案的初期，我們採用對部門種子員工進行需求訪談與後續小團體的教育訓練，並且經由半年的推廣後才能在實體通路經營部門達到預期的導入效益。在第二階段的導入目標為電子商務單位，由於建置的 Cube 所包含的維度更為複雜與系統使用者的數量更多，第一階段的指導模式已經無法適用。在本研究中，我們收集此公司使用者在商業智慧平台上五十六未使用者一個月的操作紀錄，以做為實驗分析的資料，範圍包含電子商務、供應商管理與經營管理三個單位的所有操作，總計一萬八千六十九筆。後續我們將會依據第三章的方法，將所有操作紀錄轉換成可分析的結果集合，以做為後續實驗使用。在下一節我們將就多維度分析推薦輔助系統的幾個主要功能模組作說明。

### 6.2 多維度分析推薦輔助系統功能模組說明

圖 6-1 為多維度分析推薦輔助系統的處理流程圖，此推薦輔助系統主要包含操作紀錄正規化、操作相似度判斷、產生推薦候選項目集合、使用者分群、老手程度加權計算與推薦項目產出六個功能模組，在後續章節中我們將會對每一個功能作詳細的說明。

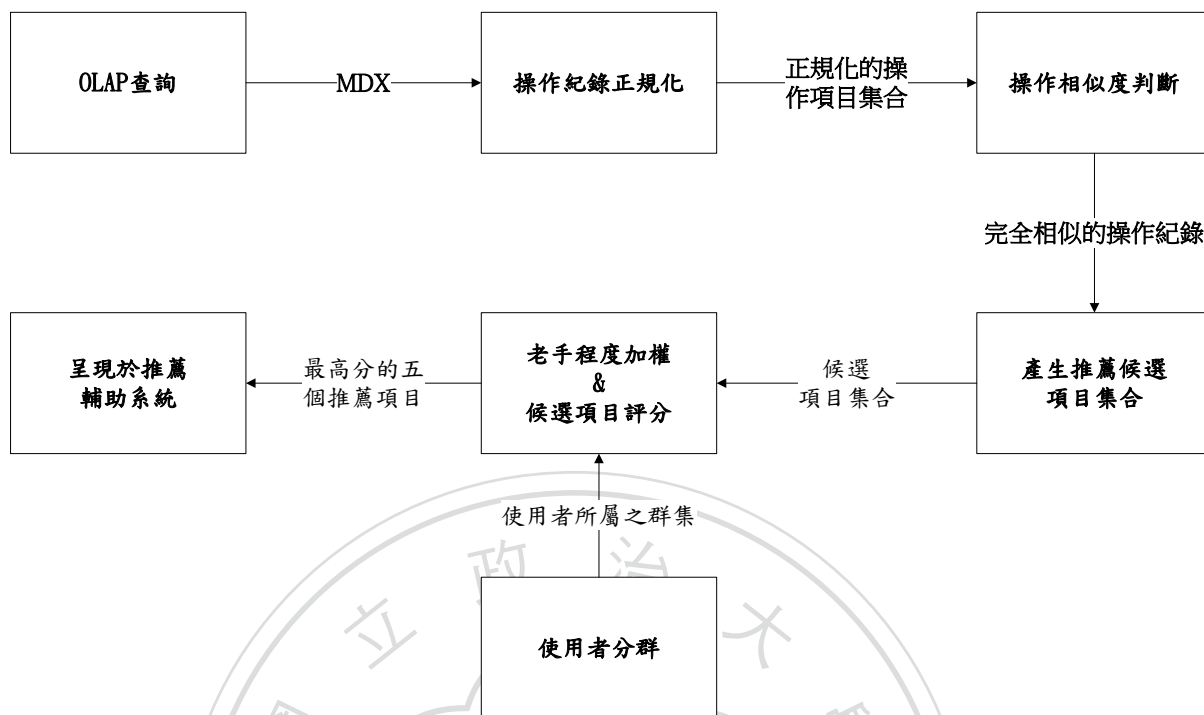


圖 6-1、多維度分析推薦輔助系統流程圖

## 6.2.1 操作紀錄正規化

圖 6-2 為此資料正規化處理的流程圖，對 SQL Server Profiler 所錄製的操作紀錄進行分析，將收集到的每一筆操作記錄，轉換成一組記錄著操作時間、操作使用者與操作項目的集合。分析的過程中，我們將會依據第三章的資料正規化的方式，將原始複雜的 MDX 轉換成過濾維度、報表維度與量值三種類型的項目集合。

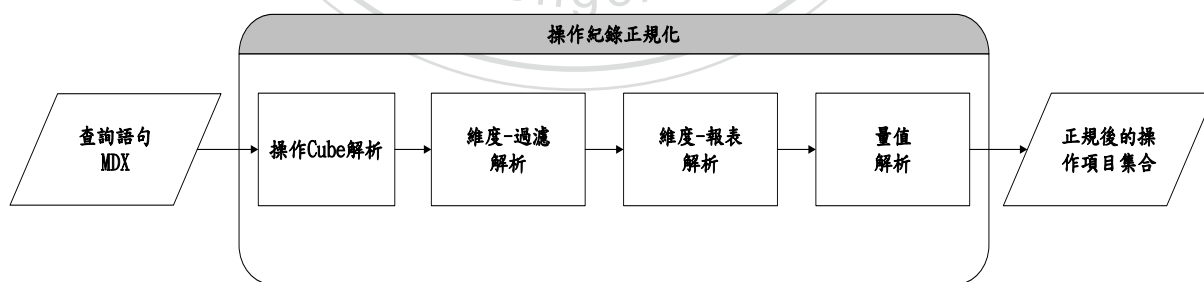


圖 6-2、操作紀錄正規化處理流程圖

根據分析得到的結果，我們可以了解此使用者目前的操作內容，用以跟歷史資料庫比較用。表 6-1 是此功能模組的輸入與輸出的結果。

表 6-1、操作紀錄分析與正規化的輸入與輸出

Input	Output
操作時間: 20120302205643	操作時間: 20120302205643
操作者: GaryChen	操作者: GaryChen
SELECT ([Measures].[消費人數]) ON 0, ( {[清算日].[清算日 年].&[2008 年],[清算日].[清算日 年].&[2009 年]} ) ON 1 FROM [卡友消費分析] WHERE ([特約商].[主產業別].&[TT])	Cube 名稱: [卡友消費分析] 報表維度名稱: [清算日].[清算日 年] 過濾維度名稱: [特約商].[主產業別] 量值名稱: [消費人數]

## 6.2.2 操作紀錄之相似度判斷

圖 6-3 為操作紀錄相似度判斷的流程圖，我們將上一節解析出的項目集合與儲存在資料庫中正規化後的歷史操作紀錄作比較，比較的過程中我們將依據操作的 Cube、過濾維度、報表維度與量值的差異作比較，在相似度判斷的結果上，我們首先會找出涵蓋查詢的操作紀錄，但如果系統無法找到完全相似的候選項目時，系統會找尋相似度次之的操作紀錄取代，以作為後續的推薦候選項目。舉例來說，使用者此次的操作內容為（[卡友消費分析]，[[清算日].[清算日 年]]，[[特約商].[主產業別]]，[[消費人數]]）時，系統會嘗試找尋是否有完全相似的操作紀錄做為後續篩選用，假設今日歷史操作資料庫並未存在完全相似的資料時，系統會自動找出相似度次之的操作紀錄做為候選項目的篩選用，以本例來說我們就會找出與報表維度、過濾條件維度與兩值中兩個相同的操作紀錄。

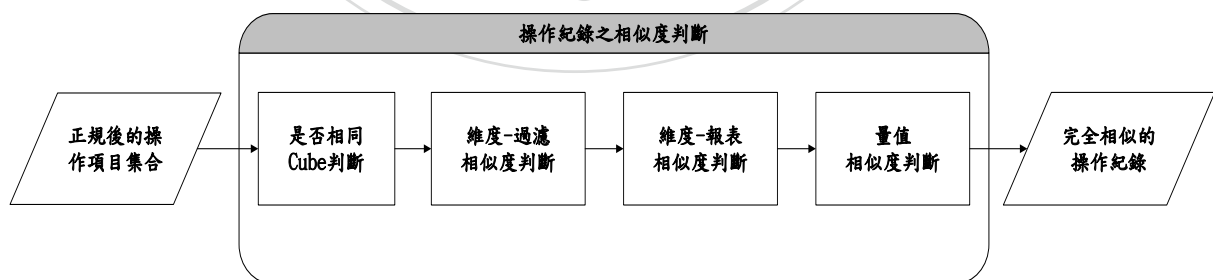


圖 6-3、相似度判斷處理流程圖

### 6.2.3 產生推薦候選項目集合

圖 6-4 為產生推薦候選項目集合的處理流程圖，在上一節我們將歷史資料庫中與此次完全相似的操作紀錄篩選出來後，我們會將篩選出來的操作紀錄所包含的集合，過濾掉此次操作所包含的項目後，將剩餘的操作項目去除重複的動作，以找出所有可能的推薦候選項目，而表 6-2 是此功能模組的輸入與輸出的結果。

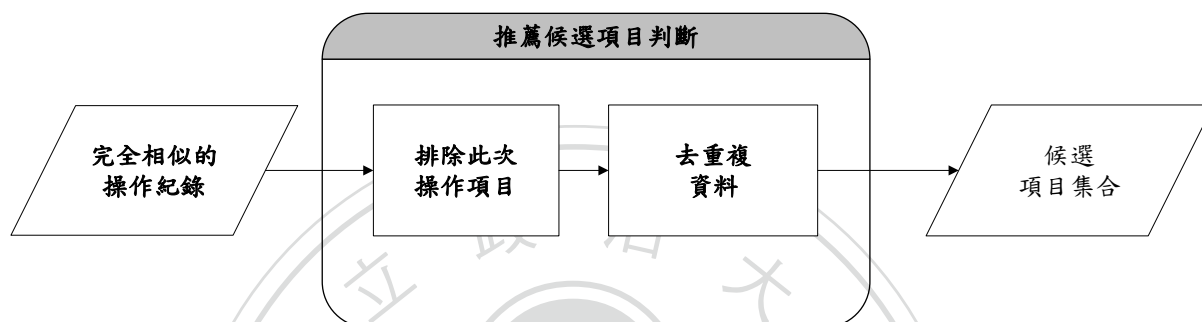


圖 6-4、產生推薦候選項目集合流程圖

表 6-2、產生推薦候選項目集合的輸入與輸出

Input	Output
此次操作紀錄	推薦候選項目
$q_1 = ([\text{卡友消費分析}], \{\{\text{清算日}, [\text{清算日 年}], [\text{供應商}], [\text{區域}], [\text{卡友維度}], [\text{教育程度}], [\text{卡片維度}], [\text{發卡來源}], [\text{卡友維度}], [\text{年齡級距}]\}, \{\{\text{產品別}], [\text{產品類別}], [\text{卡友維度}], [\text{年收入級距}]\}, \{\{\text{消費人數}\}\})$	$\{\{\text{產品別}], [\text{產品金額級距}], [\text{消費金額}], [\text{消費次數}]\}$
歷史操作紀錄	
$q_3 = ([\text{卡友消費分析}], \{\{\text{清算日}, [\text{清算日 年}], [\text{供應商}], [\text{區域}], [\text{卡友維度}], [\text{教育程度}]\}, \{\{\text{產品別}], [\text{產品類別}], [\text{卡友維度}], [\text{年收入級距}], [\text{產品別}], [\text{產品金額級距}]\}, \{\{\text{消費人數}\}\})$	
$q_4 = ([\text{卡友消費分析}], \{\{\text{清算日}, [\text{清算日 年}], [\text{供應商}], [\text{區域}], [\text{卡友維度}], [\text{教育程度}]\}, \{\{\text{產品別}], [\text{產品類別}], [\text{卡友維度}], [\text{年收入級距}], [\text{產品別}], [\text{產品金額級距}]\}, \{\{\text{消費人數}\}\})$	

應商].[區域],[卡友維度].[教育程度],[卡友維度].[教育程度]},{[產品別].[產品類別],[卡友維度].[年收入級距]},{[消費人數],[消費金額],[消費次數]})	
---	--

## 6.2.4 使用者分群

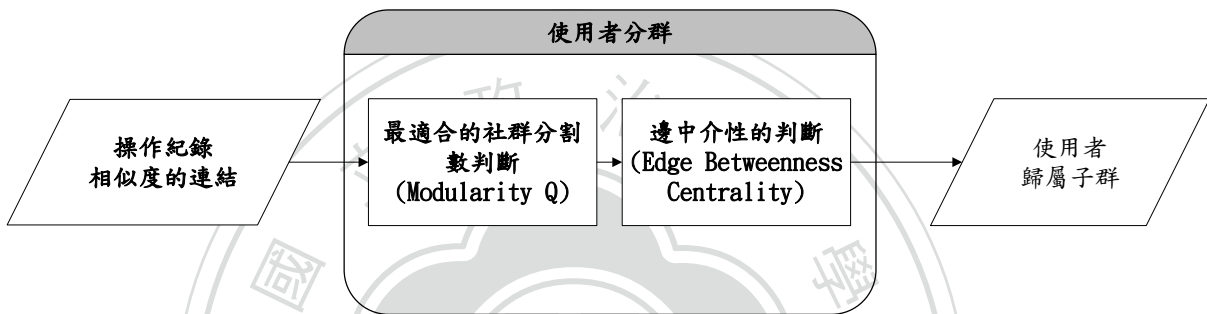


圖 6-5、使用者分群的流程圖

圖 6-5 為使用者分群的流程圖，在使用者分群上系統首先會依據使用對於推薦結果的回饋建立使用者之間的連結，當使用者點選推薦結果中感興趣的項目時，系統會從歷史操作紀錄的資料庫中找出曾經操作過相同內容的使用者，並且將兩個使用者間增加一條連結。接下來我們會過濾使用間連結總數大於 5 的，並且依據 Modularity Q 的計算得到最適合的社群分割數，接下來我們會利用邊中性的判斷，定義出各使用者歸屬的分群，做為後續推薦項目篩選時，用來過濾同質使用者的操作紀錄使用，用以提昇推薦機制的正確率。在邊中介性的計算上，我們首先利用 Dijkstra 演算法找出每個節點至其他節點的最短路徑，首先我們會先將起始節點丟至『已判斷節點的佇列』，並且將剩餘的節點放置『未判斷節點的佇列』，接下來我們計算『未判斷節點的佇列』中各節點至『已判斷節點的佇列』的距離，找出路徑長度最少的並且將此節點從『未判斷節點的佇列』移至『已判斷節點的佇列』，重複計算直到『未判斷節點的佇列』為空時。最後我們將每條邊出現在最短路徑上的次數做為邊中介的值。下列是我們在處理各節點至其他節點的最短路徑與取得邊中介值的程式碼，而圖 6-6 是一個我們以節點 A 為起始點的最短路徑處理範例：



```

-- 逐一處理每個節點至其他節點的最短路徑
FOR EACHENODE IN GRAPH
    CALL GET_THE_SHORTEST_PATH(@NODE)
-- 取得位於最短路徑次數最多的邊(NODE_1, NODE_2)
SELECT NODE_1, NODE_2 , COUNT(1)
FROM SHORTEST_PATH_RESULT
GROUP BY NODE_1, NODE_2
ORDER BY COUNT(1) DESC

PROCEDURE GET_THE_SHORTEST_PATH (@START_NODE)
DECLARE
    @NODE_QUEUE_1 AS dbo.NODE_LIST , -- 已判斷節點佇列
    @NODE_QUEUE_2 AS dbo.NODE_LIST , -- 未判斷節點佇列
    @NODE_MID AS VARCHAR(10) , -- 中間點
    @NODE_END AS VARCHAR(10) , -- 終點
    @DISTANCE AS INT -- 距離
-- 將起始點放置已判斷節點佇列
INSERT @NODE_QUEUE_1
SELECT @START_NODE , 0
-- 將起始點放置未判斷節點佇列
INSERT @NODE_QUEUE_2
SELECT *, NULL FROM dbo. [FC_GET_NEW_QUEUE](@START_NODE)
-- 當未判斷節點佇列不為空時
WHILE ((SELECT COUNT(1) FROM @NODE_QUEUE_2)>0)
BEGIN
    -- 取出長度最短的節點
    SELECT @NODE_MID=NODE1, @NODE_END=NODE2, @distance=distance
    FROM (
        SELECT TOP 1 node1, NODE2 , DISTANCE AS DISTANCE
        FROM dbo. [NODE_LINK_Avw] A
            INNER JOIN @NODE_QUEUE_1 B ON A.NODE1=B.NODE
            --INNER JOIN @NODE_QUEUE_2 C ON A.NODE2=C.NODE
        WHERE node1 IN (SELECT NODE FROM @NODE_QUEUE_1) AND NODE2 IN (SELECT NODE FROM
@NODE_QUEUE_2)
        ORDER BY B.ACC_DISTANCE+distance ASC
    ) A
    -- 抓出之前走過的點

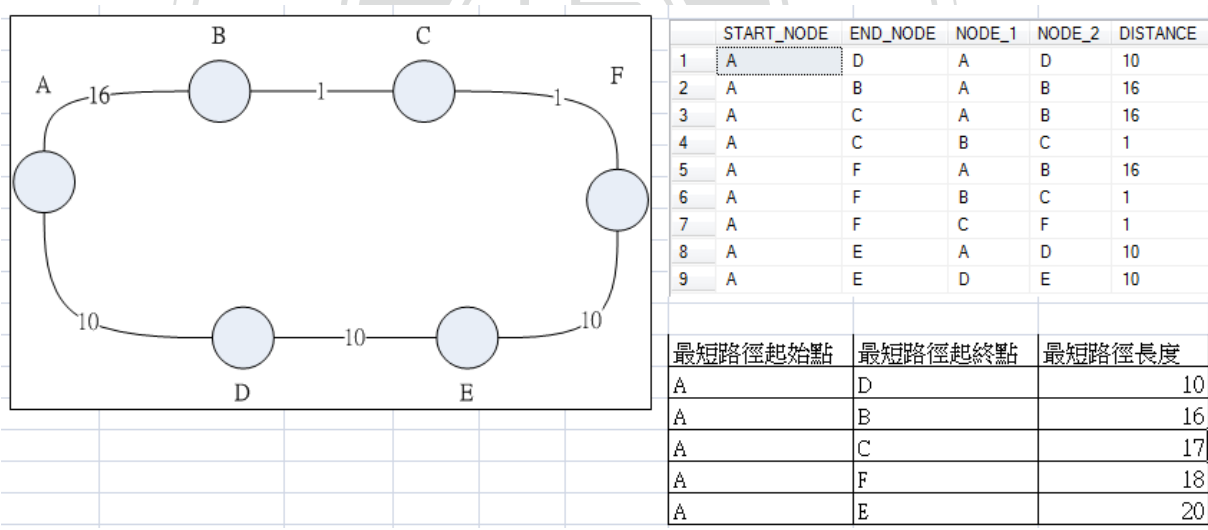
```

```

INSERT dbo.SHORTEST_PATH_RESULT
SELECT @START_NODE , @NODE_END, NODE_1, NODE_2, DISTANCE
FROM SHORTEST_PATH_RESULT
WHERE START_NODE = @START_NODE AND END_NODE = @NODE_MID
-- 將此筆資料丟入結果
INSERT dbo.SHORTEST_PATH_RESULT
SELECT @START_NODE, @NODE_END, @NODE_MID, @NODE_END, @distance
-- 將此次抓出來的點放置QUEUE1
INSERT @NODE_QUEUE_1
SELECT @NODE_END, ISNULL(SUM(DISTANCE), 0)
FROM SHORTEST_PATH_RESULT
WHERE START_NODE = @START_NODE AND END_NODE = @NODE_END
-- 將此次抓出來的點從QUEUE2移除
DELETE FROM @NODE_QUEUE_2 WHERE NODE = @NODE_END
END

```

圖 6-6、以節點 A 為起始點的最短路徑處理範例



而在本次的實驗樣本上，將依據操作的時間找出最近的四千筆操作資料與之前的一萬四千筆的操作紀錄作比對，當有其他使用者操作包含與樣本操作紀錄的相似時，系統會自動將兩個使用者間增加一條連結，並且利用這樣的結果我們將可以建立第一次的分群結果，系統後續在實務上的運作時，將採用批次的方式於每日的夜間進行分群的計算。

表 6-3 是一個使用者分群輸入與輸出的範例，以前四筆資料為例子使用者 justinyang 在 2011 年 05 月 27 日 17 點 22 分的操作，而被採納的推薦項目在候選項目的篩選時曾

經有 Amandachang、Bentsao、Kennychu 與 grace3389 四個人曾經操作相同的操作內容，因此在社群網絡的連結中各別建立一筆此四人與 justinyang 間的連結，後續系統就會依據各節點間的社群網絡進行凝聚子群的分割，產出各使用者歸屬的分群資訊做為輸出。

表 6-3、使用者分群的輸入與輸出

輸入			
使用者	操作時間	被參考使用者	
justinyang	20110527172247	Amandachang	
justinyang	20110527172247	Bentsao	
justinyang	20110527172247	Kennychu	
justinyang	20110527172247	grace3389	
janasher	20110527101708	Janylo	
janylo	20110527172336	Kennychu	
LEOLIN	20110527172428	ConnieLai	
justinyang	20110527172447	Kennychu	
Jliang	20110527172608	Janylo	
輸出			
Janasher	Group_1	Bentsao	Group_2
Janylo	Group_1	Kennychu	Group_2
Jliang	Group_1	grace3389	Group_2
Justinyang	Group_2	ConnieLai	Group_3
Amandachang	Group_2	LEOLIN	Group_3

### 6.2.5 老手程度加權計算

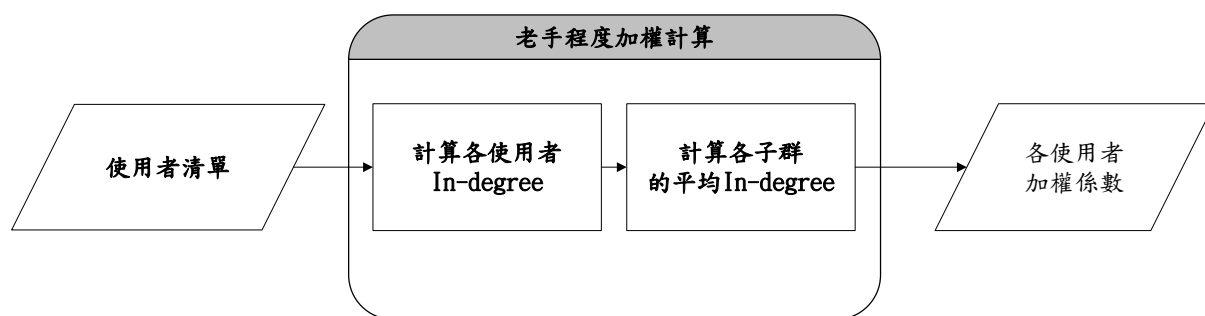


圖 6-7、老手程度加權的流程圖

圖 6-7 為老手程度加權的流程圖，第五章我們有提到在 OLAP 系統上的使用者，是否相

同部門對於推薦的正確性會有很大的影響，此外對於系統的熟悉程度對於操作的建議也是很重要的因素。在此次研究的實作上，我們會依據同分群內各使用者的 In-degree 資訊作為老手程度加權的判斷；我們在之前的實驗中發現如果一個使用者的 In-degree 數目越大者，表示他在該分群中資訊獲得越多的認同。因此，我們在推薦候選項目的評分上，我們將每日晚上計算的加權結果反應在系統權重的設定上。有關此部分的處理方式，我們首先會先計算所有使用者的 In-degree 數目，接下來我們會計算各子群內的平均 In-degree 數，最後將各個使用者的 In-degree 數目除以子群內的平均 In-degree 數就是該使用者的加權。

### 6.2.6 推薦項目產出

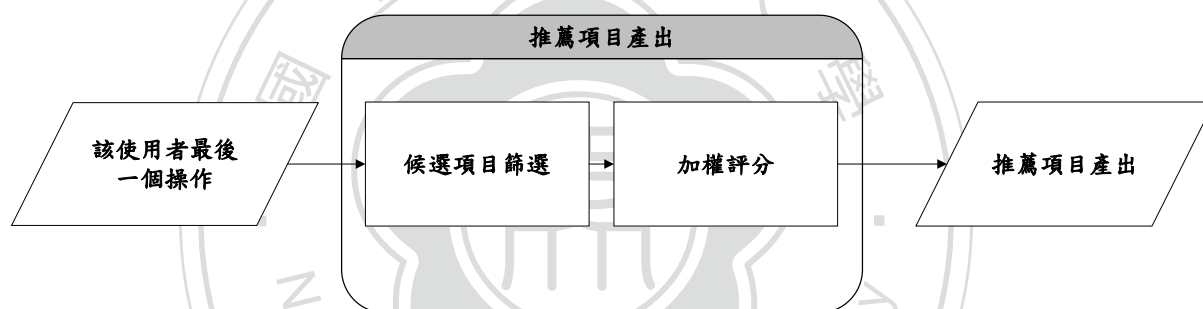


圖 6-8、推薦項目產出的流程圖

圖 6-8 為推薦項目產出的流程圖，系統在計算老手程度的加權與使用者分群後，推薦機制將會依據推薦機制類型、最小出現筆數(Minimum Support)、每頁推薦項目個數與是否啟用關聯加權的設定，決定產出的內容。在處理的過程中，首先我們會將該使用者最後的操作內容作分析後輸入推薦機制內，系統會找出完全相似、大於最小出現筆數且操作者歸屬同分群的資料作為分析的對象；接下來在候選項目的篩選中，我們會依據加權後的結果作評分，最後依據使用者設定的每頁推薦項目個數做為最後推薦的產出。

### 6.2.7 使用者回饋機制

圖 6-9 為使用者回饋的流程圖，在系統提供的推薦項目時，假設使用者甲對於推薦的項目 A 與項目 B 都認為有效時，使用者會在系統中點選此兩項目並且回饋至系統，系統會依據此回饋的資訊，找出此次推薦過程中包含此兩項目操作的使用者，在系統中增加

一筆該使用者與這些使用者間的連結；此外我們也針對使用者在操作過程中，我們也會將此次操作與前次操作進行比較，當差異的項目與最後一次採納的推薦項目一致時，我們也將會建立一筆負向的連結，用以真實反應推薦被採納的狀況，並且將統計的結果作為後續凝聚子群與老手程度的更新時使用。

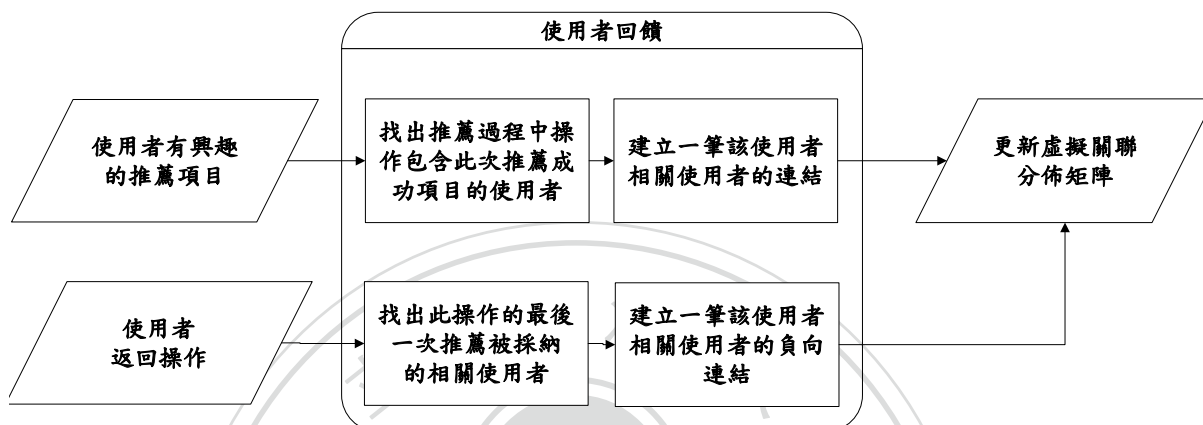


圖 6-9、使用者回饋的流程圖

### 6.3 多維度分析推薦輔助系統操作介面介紹

在本次研究所設計的推薦輔助系統，主要分為兩個部分，系統參數設定的介面與使用者的操作介面。圖 6-10 中就是此次設計的推薦系統的參數設定介面，我們可以在此設定下列推薦機制的判斷設定：

- 推薦機制類型：我們可以在此設定推薦時系統所使用的演算法，目前的選項包含最大信心度或者是最大使用人次選擇。
- 推薦項目個數：我們可以此設定每次系統推薦時，每個頁面顯示的推薦項目個數，預設值為 5 個。
- 最小出現筆數(Minimum Support)：
  - 最大信心度：當一個候選項目符合 Minimum Support 限制，才會進行推薦。
  - 最大使用人次選擇：當一個人對候選項目操作的筆數大於 Minimum Support，其對此項目的連結才被認為是有效的連結。
- 是否啟用關係加權處理：推薦判斷時，是否將凝聚子群與老手程度加權加入判

斷。當設定為是時

- 同群使用者的操作才會納入候選項目與認同機制判斷
- 使用者的操作會有加權效果，依據其 In-degree 中心性的程度作加權。
- 當無候選項目時是否回推：當使用者的操作進行到無任何符合的推薦項目時，系統是否需要回推至上一操作步驟，重新進行推薦的判斷。

推薦機制類型	最大信心度選擇
推薦項目個數	10
最小出現筆數 (Minimum Support)	0 (PS: 出現次數大於此設定才會顯示於系統中)
是否啟用關係加權處理	Y (PS: 是否將參考關係係數套用於各候選項目分數的計算中)
當無候選項目時是否回推	N (PS: 假設找不到任何相符合的候選項目時 是否往前一步驟推)
保存 回推薦系統	

圖 6-10、推薦系統參數設定介面

使用者操作介面上，在使用者登入多維度分析平台後，系統會在右邊區塊呈現此次提供的推薦輔助機制，如圖 6-11 所示。



圖 6-11、推薦輔助系統登入畫面

首先需要輸入需要使用者的帳號，確認追蹤的帳號後系統會即時的紀錄該使用者的操作，並且進行正規化的分析。在 OLAP 的操作過程中，當使用者需要推薦輔助系統的協助時，只需在點選『操作資料分析』（如圖 6-12 所示），系統將會帶出該使用者最後的操作內



容，系統將會依據最後的操作內容與歷史操作資料庫進行匹配，提供依據演算法計算出最有可能的五個項目，做為進行下一個操作的參考，如圖 6-13 所示。

The screenshot shows the Strategy Companion Analyzer CPM interface. On the left, there is a list of metrics such as '業績\_出貨延遲訂單比例' and '業績\_加價購訂單金額'. The main area displays a '樞紐分析表1' (Pivot Table 1) with columns for '銷售館別\_館別' and '業績\_合計(已付款+取消+銷貨退回+快樂購卡點數折抵金額)'. The right side features a recommendation panel with instructions: 'A. 如果正確請點選推薦' and 'B. 如果不正確請點選刷新'. Below these instructions is a table with columns 'Object\_TYPE' and 'Object\_DESC', and buttons for '操作資料分析' and '推薦項目'.

Object_TYPE	Object_DESC
Cube	EC_訂單分析
Dimension (Column or Row : I)	EC_產品銷售館別_銷售館別_館別
Measure	業績_合計(已付款+取消+銷貨退回+快樂購卡點數折抵金額)

圖 6-12、查詢操作正規化結果

The screenshot shows the Strategy Companion Analyzer CPM interface. On the left, there is a list of metrics such as '平均單筆銷售金額(營收)' and '業績\_使用抵用券(已付款+取消+銷貨退回)'. The main area displays a '樞紐分析表1' (Pivot Table 1) with columns for '銷售館別\_館別' and '業績\_合計(已付款+取消+銷貨退回+快樂購卡點數折抵金額)'. The right side features a recommendation panel with instructions: '推薦機制一: SUPPORT 推薦結果' and '下列清單是推薦項目 如果有存在期項目請勿選後送出'. Below these instructions is a table with columns 'Object\_DESC' and buttons for '送出項目', '返回推薦頁面', and '更多的推薦'.

Object_DESC
<input type="checkbox"/> 業績_合計(已付款+取消+銷貨退回)訂單淨毛利
<input type="checkbox"/> 業績_合計(已付款+取消+銷貨退回)訂單淨毛利率
<input type="checkbox"/> 業績_成立已付款訂單金額
<input type="checkbox"/> 業績_成立已付款訂單淨毛利
<input type="checkbox"/> 業績_成立已付款訂單淨毛利率

圖 6-13、推薦結果

假設此次的推薦結果中包含使用者有興趣的資料時，在系統中勾選有興趣的項目後點選送出項目，系統將會自動更新該使用者與其他使用者間的連結。假設此次推薦結果中不存在正確的項目時，使用者可以點選『更多的推薦』，系統將會排除此次推薦項目後重新進行推薦，如圖 6-14 所示。

推薦機制一: SUPPORT 推薦結果  
下列清單是推薦項目 如果有存在期望項目請勾選後送出

Object	DESC
<input type="checkbox"/>	業績_合計_(已付款+取消+銷貨退回)訂單淨毛利
<input checked="" type="checkbox"/>	業績_合計_(已付款+取消+銷貨退回)訂單淨毛利率
<input checked="" type="checkbox"/>	業績_成立已付款訂單金額
<input type="checkbox"/>	業績_成立已付款訂單淨毛利
<input type="checkbox"/>	業績_成立已付款訂單淨毛利率

圖 6-14、更多推薦的結果

### 6.3 模擬實驗

在本次的實驗中，我們將正規化後的資料分為十組。實驗的過程中，每次我們會選擇其中一組做為測試組，剩下的九組做為訓練組來做為實驗。最後將模擬驗證的結果以精確度(Precision)作為評斷推薦機制的優劣，精確度的計算公式如下：

$$\text{Precision}(R_i, G_j) = \frac{|R \cap A|}{|R|} \quad (13)$$

R: 使用推薦機制的次數

A: 推薦並且被接受的次數

在本次的模擬實驗中，我們利用 SQL Server 撰寫一隻 Stored Procedure 進行驗證，在此 Stored Procedure 中，我們會將每組資料隨機取兩百資料做註記，後續將以相同的資料對兩種推薦機制與加權後的推薦機制進行驗證。在下面的虛擬碼中，首先我們將以迴圈的方式對每組資料進行實驗，在實驗過程中我們將會對隨機一筆未實驗過的資料標註，並且對此資料同時做最大信心度與最大使用人次選擇兩種推薦機制的實驗。

```

WHILE @GROUP_ID <= 10
BEGIN
    WHILE @CNT <= 200
    BEGIN
        -- 取得此次實驗用的資料列編號
    
```

```

EXEC [GET_RAND_SEED] @GROUP_ID , @RAND_SEED OUTPUT
-- 判斷此次取得的資料是否已被選取過
SELECT @CNT= CASE WHEN SEED_IND IS NULL
      THEN @CNT+1 ELSE @CNT END
FROM REF_RAND_SEED_ALL
WHERE [SID] = @RAND_SEED
-- 將此筆資料標註為實驗用紀錄
UPDATE A
      SET SEED_IND =1
FROM REF_RAND_SEED_ALL A
WHERE [SID] = @RAND_SEED
WHERE [SID] = @RAND_SEED
-- 推薦機制：最大信心度選擇的模擬測試
EXEC dbo.SP_CHECK_RECOMMENDATION_SUPPORT @RAND_SEED
-- 推薦機制：最大使用人次選擇的模擬測試
EXEC dbo.SP_CHECK_RECOMMENDATION_PR_I @RAND_SEED
END
SELECT @GROUP_ID = @GROUP_ID +1 , @CNT=1
END

```

實驗的過程，我們會依據操作的順序作判斷，移除最後一個操作項目，並且傳入兩種推薦機制中取得推薦項目。當該操作的最後一個項目包含於推薦機制產出的五個項目中時，我們會認為此次的推薦正確。下列我們將分別對最大信心度與最大使用人次選擇兩種推薦機制驗證結果作說明，此外我們也會將凝聚子群與老手程度加權加入至推薦機制作驗證。

### 6.3.1 推薦機制：最大信心度選擇的模擬結果分析

在此次的模擬測試中，我們會將之前篩選出來的紀錄逐筆的傳入至最大信心度選擇的推薦機制中，系統將會依據測試組的資料找出完全相似的資料並且找出可能的候選項目，並且依據每個候選項目被操作的次數做為判斷找出出現頻率最高的五個項目做為推薦，下列表 6-4 是我們在最大信心度選擇推薦機制的模擬結果。在最大信心度選擇的推薦機制的實驗結果，平均來說，此推薦機制的精確率[14]為 57.35%。

表 6-4、推薦機制：最大信心度選擇的模擬結果

GROUP	測試筆數	推薦機制：最大信心度選擇正確筆數	
1	200	93	46.50%
2	200	116	58.00%
3	200	146	73.00%
4	200	86	43.00%
5	200	133	66.50%
6	200	140	70.00%
7	200	123	61.50%
8	200	68	34.00%
9	200	100	50.00%
10	200	142	71.00%
合計	2000	1147	57.35%

### 6.3.2 推薦機制：最大使用人次選擇的模擬結果分析

在此次的模擬測試中，我們會將之前篩選出來的紀錄逐筆的傳入至最大使用人次選擇的推薦機制中，系統將會依據測試組的資料找出完全相似的資料，並且找出可能的候選項目，依據每個候選項目被使用次數做為判斷，取出使用人次前五名的五個項目做為推薦，下列表 6-5 是我們在最大使用人次選擇推薦機制的模擬結果。

表 6-5 推薦機制：最大使用人次選擇的模擬結果

GROUP	測試筆數	推薦機制：最大信心度選擇正確筆數	
1	200	101	50.50%
2	200	137	68.50%
3	200	137	68.50%
4	200	92	46.00%
5	200	117	58.50%
6	200	138	69.00%
7	200	115	57.50%

8	200	75	37.50%
9	200	96	48.00%
10	200	142	71.00%
合計	2000	1150	57.50%

在至最大信心度選擇的推薦機制的實驗結果，平均來說此推薦機制的精確率為 57.50%。在實驗結果上，我們發現在不考慮關係加權的情況下，兩種推薦機制的精確率幾乎差不多是 57 個百分比。在後續兩個章節中，我們將會把分群結果與參考關係加權的加入推薦機制作比較，並且會將推薦結果跟原本的推薦機制結果作比較。

### 6.3.3 推薦機制：最大信心度選擇加入參考關聯加權的模擬結果分析

在此次的模擬測試中，我們將啟用推薦判斷中分群結果與參考關係加權的功能，系統在篩選候選項目的時候，首先會將非同群使用者的操作紀錄排除，此外每個使用者的操作次數會依據老手程度進行加權的效果，然後利用此調整後的方式篩選出同群使用者操作加權次數後的前五個項目最為推薦，下列表 6-6 是我們在最大信心度選擇加入參考關聯加權的模擬結果分析。

表 6-6、推薦機制：最大信心度選擇加入參考關聯的模擬結果

GROUP	測試筆數	推薦機制：最大信心度選擇正確筆數	
1	200	98	49.00%
2	200	126	63.00%
3	200	152	76.00%
4	200	96	48.00%
5	200	149	79.00%
6	200	160	81.50%
7	200	126	63.00%
8	200	110	55.00%
9	200	119	59.50%
10	200	149	74.50%
合計	2000	1285	64.25%

在最大信心度選擇的推薦機制加入參考關聯加權的實驗結果此推薦機制的精確率從原

本的 57.35% 上升到 64.25%。除了正確的提升外，我們在表 6-7 中也發現了，最大信心度選擇的推薦機制中，當我們將參考關聯與分群的結果加入後，推薦項目正確的順位上也往上提升。

表 6-7、最大信心度選擇的推薦機制加入參考關聯加權後的正確推薦順位比較

未考慮關聯加權與分群因素的研究數據							
GROUP_ID	測試筆數	正確筆數	出現在第一順位的筆數	出現在第二順位的筆數	出現在第三順位的筆數	出現在第四順位的筆數	出現在第五順位的筆數
1	200	93	45	24	9	15	0
2	200	116	81	21	6	6	2
3	200	146	111	18	10	0	1
4	200	86	20	63	0	2	1
5	200	133	118	7	4	2	2
6	200	140	107	11	4	12	6
7	200	123	97	14	1	9	2
8	200	68	55	9	1	2	1
9	200	100	80	11	4	2	3
10	200	142	128	7	2	5	0
考慮關聯加權與分群因素的研究數據							
GROUP_ID	測試筆數	正確筆數	出現在第一順位的筆數	出現在第二順位的筆數	出現在第三順位的筆數	出現在第四順位的筆數	出現在第五順位的筆數
1	200	98	53	32	6	4	3
2	200	126	91	22	7	5	1
3	200	152	125	21	10	1	0
4	200	96	73	11	1	11	0
5	200	158	126	9	8	4	2
6	200	163	113	20	10	9	8
7	200	126	100	15	3	6	2
8	200	110	61	12	14	19	4
9	200	119	86	17	10	2	4
10	200	149	130	8	6	5	0



### 6.3.4 推薦機制：最大使用人次選擇加入參考關聯加權的模擬結果分析

在此次模擬測試中，我們將現有的最大使用人次選擇推薦機制加入參考關聯加權，在篩選候選項目的時，首先會將非同群的使用者排除，在認同人次的判斷上會依據老手程度進行加權的效果，然後利用此調整後的方式篩選出同群中人操作加權人次後，認同人次最高的前五個項目最為推薦。下列表 6-8 是我們最大使用人次選擇推薦機制中加入參考關聯加權的模擬結果分析。

表 6-8、推薦機制：最大使用人次選擇加入參考關聯的模擬結果

GROUP	測試筆數	推薦機制：最大信心度選擇正確筆數	
1	200	119	59.50%
2	200	148	74.00%
3	200	155	77.50%
4	200	95	47.50%
5	200	161	80.50%
6	200	163	81.50%
7	200	126	63.00%
8	200	130	65.00%
9	200	120	60.00%
10	200	151	75.50%
合計	2000	1368	68.40%

表 6-9、最大使用人次選擇推薦機制加入參考關聯加權後的正確推薦順位比較

未考慮關聯加權與分群因素的研究數據							
GROUP_ID	測試筆數	正確筆數	出現在第一順位的筆數	出現在第二順位的筆數	出現在第三順位的筆數	出現在第四順位的筆數	出現在第五順位的筆數
1	200	101	73	15	5	2	6
2	200	137	91	21	11	12	2
3	200	137	105	26	4	1	1
4	200	92	14	15	1	62	0
5	200	117	87	24	6	0	0
6	200	138	104	11	9	12	2
7	200	115	86	20	6	3	0

8	200	75	55	14	5	1	0
9	200	96	73	19	3	0	1
10	200	142	120	16	1	5	0
考慮關聯加權與分群因素的研究數據							
GROUP_ID	測試筆數	正確筆數	出現在第一順位的筆數	出現在第二順位的筆數	出現在第三順位的筆數	出現在第四順位的筆數	出現在第五順位的筆數
1	200	98	83	18	9	5	4
2	200	126	95	30	9	12	2
3	200	152	116	27	6	5	1
4	200	96	77	4	13	1	0
5	200	158	142	6	5	5	3
6	200	163	118	21	16	7	1
7	200	126	98	15	6	6	1
8	200	110	72	45	9	4	0
9	200	119	90	16	2	7	5
10	200	149	128	11	4	5	3

在實驗結果上，我們發現在不考慮關係加權的情況下，兩種推薦機制的正確率幾乎差不多是 57 個百分比。但是在將分群結果與參考關係加權的加入推薦機制時，我們發現兩種推薦機制在精確度都提升至少 5 個百分比。而最大使用人次選擇推薦機制的又比最大信心度推薦機制提升多 5 個百分比。另外，在未加權時，最大使用人次選擇推薦機制在某些測試組的推薦結果不如最大信心度推薦機制，但在加入分群結果與關係加權後，幾乎每一組的推薦結果都比最大信心度推薦機制好。此外，我們在推薦順位的比較上，一樣在加入關聯加權與分群結果至最大使用人次選擇推薦機制後，推薦成功的順位也往上提升。

#### 6.4 使用者實際操作回饋

在本次實驗中，除了在上一小節中我們利用資料庫的歷史資料模擬測試與結果分析外，我們也將邀請十位目前多維度分析平台的使用者操作此次實驗所建置的推薦系統，並且請使用者提供對本系統提供關於操作介面與推薦機制正確性的回饋與建議。在上一節在比較了兩種推薦機制後，本次實驗中，我們將以成效比較好的最大使用人次選擇推薦機制做為實驗中使用的推薦方法。本次邀請進行實驗的十位人員中，主要是歸屬於電子商

務部門、供應商管理部門與企業經營部門，下列表 6-10 是本次實驗人員的背景資料：

表 6-10、實驗人員的背景資料

人員編號	部門	OLAP 系統 使用期間	平均每日使用 OLAP 次數	平均多少個操作才能 找到所需的資訊	平均每週多少次的 操作找到所需的資 訊
使用者 A	電子商務 (館長)	一年以上	三次至五次	十至十五個操作	十次以內
使用者 B	電子商務 (館長)	六個月至一 年	五次至十次	十至十五個操作	十至二十五次
使用者 C	電子商務 (組員)	三至六個月	十次以上	十五至二十個操作	二十五次以上
使用者 D	電子商務 (組員)	三個月內 (一個月內)	十次以上	二十個操作以上	經常無法有效找到
使用者 E	電子商務 (組員)	三個月內	五次至十次	經常找不到	十至二十五次
使用者 F	企業經營 (副理)	兩年以上	三次至五次	十個操作內	十次以內
使用者 G	企業經營 (組長)	一年以上	五次至十次	十至十五個操作	十至二十五次
使用者 H	企業經營 (專員)	三個月內	五次至十次	十五至二十個操作	十次以內
使用者 I	供應商管 理(組長)	一年以上	三次至五次	十至十五個操作	十至二十五次
使用者 J	供應商管 理(專員)	三個月內	五次至十次	十五至二十個操作	十至二十五次

#### 6.4.1 問卷的回饋與分析

在實驗的過程中，首先我們將會利用問卷【附錄一】的方式收集使用者對於推薦系統使用的心得與建議，收集每個人對於 OLAP 系統熟悉的程度，以做為後續推薦結果的分析，並且在最後請使用者提供介面與功能上的建議，做為後續改善的目標。表 6-11 是此次實驗的受訪者對於 OLAP 操作輔助系統的滿意度回饋統計，此外我們也將利用五點量表

的方式來量化本次受訪者對於此輔助系統在各面向的滿意程度。

表 6-11、OLAP 操作輔助系統的滿意度回饋統計

	非常滿意	不同意	沒意見	同意	非常同意	滿意程度
	20	40	60	80	100	
推薦輔助系統能正確提供出您所需要的資訊？	0	0	1	7	2	82
推薦項目出現的順位是否符合您的期望？	0	0	2	6	2	80
推薦輔助系統能否加速您的問題處理？	0	0	1	8	1	80
推薦的過程中，是否也能協助您了解相同部門的人最近關心的資訊？	0	0	2	6	2	80
推薦輔助系統是否能提高您使用多維度分析平台的意願？	0	0	0	4	6	94
推薦輔助系統的操作介面與使用的流程的滿意？	0	0	4	6	0	72
整體滿意度						81.23

在回饋的結果中我們可以發現到使用者對於此次設計的推薦輔助機制，在於推薦項目的認同與資訊共享的成效上都達八十分以上，表示使用者對於推薦機制上的肯定；此外我們可以發現所有的人在使用此次設計的 OLAP 操作輔助系統後，是可以提升使用者在於 OLAP 系統的使用意願。在整體的問卷結果中我們可以發現，使用者對於本設計的推薦機制的滿意度達 81.23 分，代表使用者對於本推薦系統的效益是正面的肯定。

#### 6.4.2 推薦機制的正確率分析

在推薦機制的正確率評斷上，我們將針對使用者的操作歷程中點選推薦項目的頻率作推薦的正確率的分析。表 6-12 是本次使用者在於推薦正確性上的測試結果，在本次的實驗的推薦正確率判斷上，我們首先以模擬測試時驗證正確率的方法做分析，我們可以發現在實際操作的驗證效果比模擬測試的正確率來的高很多，進一步的與使用者了解後發現，在歷史資料模擬的測試上，每次的正確項目只有一個，但是在實際上的推薦上我們推薦的內容中可能會對多個項目有興趣，或者是有的是以往未曾操作過，但在系統推薦

後方感興趣的項目，因此在實驗的結果上我們可以發現在推薦的正確率上會比模擬測試

後的結果來得佳。表 6-12、實驗的推薦正確性測試結果

	OLAP 操作次數	使用推薦次數 (A)	接受推薦次數(B)	回一步驟次數(C)	正確性判斷 B/A	正確性判斷 (B-C)/A	使用推薦比率
使用者 A	65	29	25	2	86.21%	79.31%	44.62%
使用者 B	56	36	29	5	80.56%	66.67%	64.29%
使用者 C	53	28	24	6	85.71%	64.29%	52.83%
使用者 D	96	82	70	40	85.37%	36.59%	85.42%
使用者 E	87	76	68	32	89.47%	47.37%	87.36%
使用者 F	72	22	19	3	86.36%	72.73%	30.56%
使用者 G	61	30	24	3	80.00%	70.00%	49.18%
使用者 H	62	55	46	14	83.64%	58.18%	88.71%
使用者 I	49	35	28	4	80.00%	68.57%	71.43%
使用者 J	86	66	48	15	72.73%	50.00%	76.74%

但在進一步的了解使用者的操作內容後，我們可以發現以單一步驟做為正確率的判斷上可能會不正確的；由於使用者在推薦當下或許覺得該項目是適合的，但在使用後卻發現該操作的結果並不是所需的資訊後，使用者就會移除之前推薦的項目，並且選擇額外的項目做為操作。因此我們將原先的正確率做了調整，首先我們統計使用者點選推薦項目次數視為使用者對此推薦輔助系統的認同票，接下來我們會將使用者操作內容與上次的操作比較，當此次的操作項目數小於上次的操作時，我們即認為該次推薦的結果與預期的結果不合，並且將此操作視為使用者對推薦輔助系統投反對票，最後我們將認同票數扣除反對票數後，除以使用者點選推薦的次數，以做為推薦輔助系統正確率的判斷。

表 6-13、三種類型使用者的測試統計

	平均使用推薦的次數	回一上一步驟	正確性判斷 B/A	正確性判斷 (B-C)/A
一年以上	29	3	82.76%	72.41%
三個月至一年	32	5.5	82.81%	65.63%
三個月以內	69.75	25.25	83.15%	46.95%

表 6-13 是我們對三種 OLAP 系統使用期間類型使用者的測試統計結果，在調整後的正



確率上，我們可以發現三個月內的使用者在於正確率會比平均來的低，在與使用者的檢討後發現，由於該類型的使用者對於系統功能的不熟悉，因此在推薦項目是否是適合的判斷上會比較低的，但經由推薦機制的輔助，卻是可以有縮小使用者在瀏覽資料時需要嘗試的維度範圍達到引導的功能，進而可以提升使用者對於 OLAP 系統的使用意願。

在於推薦系統的使用率上，一年以上的使用者在於系統的推薦使用比率是比較低的，但在我們經由觀察該類型使用者的操作過程中發現，該類型的使用者對 OLAP 系統的功能與操作熟悉，因此通常都會在操作六個步驟以上才會開始點選推薦的功能，所以雖然該類型的使用率是比較低的，但也因為其操作目的比較明確了，因此推薦系統會比較容易找到適合的項目做推薦，此外我們也發現使用者對於系統推薦的項目在選擇時即能明確知道是否為適合的項目，因此對於推薦返回的頻率相對會比較低，對於該類型的使用者來說，本實驗所設計的推薦輔助機制是可以有效的提供同部門的同仁感興趣的項目，做為日常分析時的參考用。最後在三個月至一年內的使用者類型上，由於該類型使用者在 OLAP 系統上累積一段時間的操作經驗，對於經常使用的項目是有辦法辨別是否為其所需的內容，但是對於未曾使用過的項目推薦時，他會很感興趣了解別人在看哪些資訊，但有時候也會遇到推薦的內容與其期望的資訊不一致，因此對於該類型的人來說，在返回上一步驟的次數會比一年以上的使用者來的多二點五次，在考慮反對票的影響後推薦的正確率依然保持在六成五以上，因此受訪者依然認為這樣的推薦機制是能達到當初預期的操作輔助與知識分享的效果。

在本次的實驗中，我們除了收集使用者對於目前推薦系統的回饋與實際操作效益的分析外，下面也將針對使用者提出幾個未來 OLAP 推薦機制改善的建議做說明。由於實務上使用都未曾有過 OLAP 系統推薦機制的使用經驗，因此使用者會期望能增加些推薦項目額外的相關資訊用以增加推薦項目的識別能力，進而減少操作過程中返回的機率。主要期望增加的資訊包含下列幾種：

- 推薦的項目中提供曾經操作過的人員中頻率最高的前十名：系統在推薦的過程中會自動的依據使用者的分群結果與老手程度的加權，但在使用者使用的初期會期望能提供此資訊協助使用者確認該項目的操作人員是否真的都是同業務人員；此外，在有些特殊情況使用者會希望只採納部分人員的操作，而非參考同部門內所有人員的操作；因此藉由此資訊的提供能有效提供使用者做為進一步的篩選用。
- 推薦的項目最近一次操作的時間與近三十天內操作過的次數：由於本次的實驗的資



料是一個月內的操作紀錄，因此資料庫內資料的時效都仍為有效，因此推薦的結果是符合期望的。但在企業實際的導入後，使用者會擔心系統會因為無法有效判斷候選項目的時效性，導致在推薦上會經常發生因為過期的知識所產生的錯誤結果，因此在未來的研究上建議增加時間因素於候選項目的篩選。



## 第七章 結論與未來發展

### 7.1 研究結論

企業在導入商業智慧平台的過程中，通常導入失敗的最大原因不會是在工具與技術面上，絕大部分都是因為平台的使用率不如預期，而最大的因素就是因為使用者對於商業智慧提供過於豐富的資訊不知如何使用。傳統在 OLAP 系統的推薦機制上，通常都只考慮到資料的相似度，在使用者的分群上也需要經由背景資訊的收集後才有辦法進行，但這樣的研究結果通常只能適用於特定情境。

因此為了讓企業在商業智慧平台的推廣上能不需要額外作背景資料的收集與分析的作業，此研究將依據使用者在 OLAP 上面的互動與相似度進行分析，提供一個整合社群關係的 OLAP 操作推薦機制。在設計的推薦機制中，我們以同質的凝聚子群來取得實體的部門關係，此外在認同度上我們增加依據 In-degree 中心性計算的關聯加權係數，對於系統老手的操作進行加權的效果，用以提升推薦機制的正確率。我們對於本次實驗的結果上可以歸納為下列幾點：

- (1) 利用凝聚子群的結果可以有效的排除無效的操作：以往的推薦機制經常都會因為將系統內所有使用者的操作都納入推薦的考量時，經常會造成推薦判斷上受到無效資料的影響，而降低推薦的正確率。因此在本次實驗中，我們利用 Girvan-Newman 演算法有效的區分出使用者所屬的子群，提升系統推薦的成效。
- (2) 加入老手程度的加權可以凸顯老手在於推薦機制上的重要性：以往的推薦機制上，每個使用者對於項目的認同上都只代表一票，但對於系統的老手來說，對系統的熟悉程度相對於其他人來的高，因此在本次研究中我們利用 In-Degree 來判斷每個使用者的老手加權，並且有效的提升推薦機制的正確率。
- (3) 有效的提升正確推薦的順位：在加入社群網絡分析的結果於推薦機制中時，我們可以發現推薦項目出現正確的順位也明顯往上提升，當正確推薦的順位越高的時候表示我們可以用最少的推薦即可達到推薦的目的。
- (4) OLAP 推薦機制可以有效的提升使用者使用 OLAP 系統的使用率：在使用者的回

饋上我們可以發現，在 OLAP 系統中提供一個有效的推薦系統給使用者使用，是可以協助降低使用者解決問題與系統學習的時間，另外在知識的分享上也是有很大的成效，進而提升使用者對於 OLAP 使用的意願，最終可以讓 OLAP 系統在企業內部發揮預期的成效。

## 7.2 未來發展

本篇論文旨在 OLAP 系統中，依據使用者的互動狀況與相似度的判斷，篩選出最適合的推薦項目。在本次的實驗結果中，我們可以發現社會網絡分析結果對於在 OLAP 的推薦機制上存在明顯的幫助，在後續的研究上除了可運用關聯互動分析結果與中心性加權在相關的輔助機制上，也建議朝下列兩個方向進行改善：

- (1) 候選項目演算法的改善：在此次的研究中，最大使用人次選擇推薦機制的候選項目評斷上，我們需要逐一取出各候選項目，然後計算各別的連結人次並且作加權效果，但是此方式在候選項目過多時會非常耗時，因此建議可嘗試加入新的演算法改善判斷上的速率。
- (2) 社群網絡分析擴充：在本次的研究上，每個節點間的連結我們是利用操作模式的相似度來表示，並且依據關聯的次數做為每個邊的權重，用以在建立虛擬的部門子群並且做老手程度加權上使用，建議在後面的研究上除了可以在操作模式的相似度上再找一些新的因素外，我們也建議嘗試利用使用者在基本資料或者是喜好資料上的關聯做為節點之間的邊，做為提升推薦精確度使用。
- (3) 老手程度加權係數的補強：有的專家可能是在以往的工作上培養對於 OLAP 系統的能力，但是其操作紀錄並未保存在現有的操作紀錄資料庫內，因此需要藉由一段時間的資料累積才能夠有效的反應此類型的專家，因此在後續的研究上，可以思考如何改善此部分的不足或者是縮短系統學習的時間。
- (4) 本次的研究上，我們是利用使用者在一個月內的操作資料進行分析，因此未考慮

到推薦項目時效性的因素，舉例來說在母親節時候大家經常關心的內容，在於母親節過後該主題就會失去其有效性，因此建議在後續的研究中可以針對推薦項目的最近一次操作時間與當月操作次數作加權，用以降低時效性或者是特殊事件對於推薦正確的影響。此外在資料時的效性上，由於最大使用人次選擇推薦機制在推薦的評比上是利用人次的方式，而最大信心度選擇法推薦機制在推薦的評比上是利用項目出現的次數，理論上前者在於反應時效性上的時間應該會後者來的短，因為後者通常都要是大家在於使用上的次數大於前者時才會顯現出新的議題，但是前者只要是同部門內看的人變多的時候，新的知識就會很容易被挑選出來，這部分可提供為後續研究上的觀察重點。



## 參考文獻

- [1] Freeman, L. C. "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, 1, 215-239, 1979.
- [2] D.J. Brass, and M.E. Burkhardt, "Centrality and Power in Organization," in Noria, N. & Eccles, R. G. (Eds.) *Networks and Organizations: Structure, Form and Action*, Boston, Massachusetts: Harvard Business School Press, pp.191-215, 1992.
- [3] M. S. Granovetter , "Problems of Explanation in Economic Sociology", in N. Nohria and R. Eccles (Eds), *Networks and Organization: Structure, Form, and r Action*, pp.25-62, 1992.
- [4] S. Wasserman, and K. Faust, *Social Network Analysis: Methods and Applications*, New York, Cambridge University Press, 1994.
- [5] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, "Applying Collaborative Filtering to Usenet News," *Communications of the ACM* 40(3), pp. 77-87, 1997.
- [6] J. Han, "OLAP Mining: An Integration of OLAP with Data Mining," in *Conference Tutorial: Integration of Data Mining and Data Warehousing Technologies*. (Microsoft PowerPoint slides), ICDE 1997.
- [7] M. Balabanovic and Y. Shoham, "Fab Content-based, collaborative recommendation," *Communications of the ACM* (40:3), pp.66-72, 1997.
- [8] M. Golfarelli and S. Rizzi, "Conceptual design of data warehouses from E/R schemes," in *Proc. 31st Hawaii International Conference on System Sciences*, 1998.
- [9] C. Sapia, "On Modeling and Predicting Query Behavior in OLAP Systems," in *Proc. of the International Workshop on Design and Management of Data Warehouses (DMDW'99)*, 1999.

- [10] I. Zukerman D. Albrecht and A. Nicholson “Predicting Users’ Requests on the WWW,” in *Proc. of the 7th International Conference on User Modeling*. Springer-Verlag,1999
- [11] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web, 1999.
- [12] A. Ansari, S. Essegaiier, and R. Kohli, “Internet recommendation systems,” *Journal of Marketing Research*, 37(3): 363, 2000.
- [13] C. Sapia, “Predicting Query Behavior to Enable Predictive Caching Strategies for OLAP Systems,” *PROMISE*, in Y. Kambayashi, M. Mohania, A.M. Tjoa, (eds.) *DaWaK 2000, LNCS*, vol. 1874, pp. 224 - 233. Springer, Heidelberg.
- [14] J. Han and M. Kamber, *Data Mining: Concepts and Techniques: Measure the quality of a recommendation*, Morgan Kaufmann., 2000.
- [15] M. Newman, “The Structure and Function of complex Networks,” *SIAM Review*, 45(2), 2003.
- [16] W.D. Nooy, *Exploratory Network Analysis with Pajek*, New York: Cambridge University Press, 2005.
- [17] B. Satzger, M. Endres, and W. Kießling, “A Preference-Based Recommender System,” in Bauknecht, K., Pröll, B., Werthner, H. (eds.) *EC-Web 2006, LNCS*, vol. 4082, pp. 31–40, 2006.
- [18] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577-8582, 2006.
- [19] A. Giacometti, P. Marcel, and E. Negre, “A Framework for Recommending OLAP Queries,” in *International Workshop on Data Warehousing and OLAP*, pp. 73–80. ACM, New York, 2008.
- [20] Y.-T. Chen and P.-Y. Hsu, “Supporting Tools to Query Data in Business Intelligence



Systems,” in *Proc. of International Conference on Business and Information*, (BAI2008), 2008.

- [21] H. Jerbi, F. Rava, O. Teste, and G. Zurfluh, “Applying Recommendation Technology in OLAP Systems,” in *Proc. of International Conference on Enterprise Information Systems* (ICEIS2009), 2009.



## 附錄一：多維度分析平台推薦輔助系統實驗問卷

您好：

這是一份關於『多維度分析平台之推薦輔助系統操作心得』問卷，本問卷主要是收集您在此輔助系統上的使用心得與建議，來調查本次實驗設計的推薦系統在於功能與操作介面上的成效，您的寶貴建議對本研究具有相當大的幫助。非常感謝您撥空參與本階段之研究，謝謝。

實驗日期 \_\_\_\_\_

## 一. 背景資料

A. 所屬的部門？

- 電子商務       企業經營       供應商管理

B. 職稱 \_\_\_\_\_

C. 在接觸本系統前，您是否有接觸過 OLAP 系統？

- 是       否

D. 您加入公司的時間？

- 一個月內       一至三個月       三至六個月  
 六至一年       一年以上

## 二. 多維度分析系統熟悉度

A. 您使用此 OLAP 系統的時間？

- 一個月內       一至三個月       三至六個月  
 六至一年       一年以上

B. 您平均每日使用 OLAP 系統的次數？

- 三次以內       三次至五次       五次至十次  
 十次以上

C. 在多維度分析平台中，通常您在解決一個問題時，平均需要多久次

的操作才可以找到您要的資訊？

- 十個操作內
- 十至十五個操作內
- 十五至二十個操作內
- 二十個操作以上
- 經常找不到

D. 您在操作多維度分析平台上是否通常憑自己就可以進行資料分析，還是總是需要他人的協助才能夠進行？

- 通常我自己就可在此系統上取得所需的資訊
- 大部分我都可以自己解決，但是有的時候還是需要請教別人
- 通常我都會請教該單位比較熟系的人才能找到所需的資訊

E. 在多維度分析平台的操作中，平均每週有多少次的查詢結果是能夠解決您的問題？

- 十次以內
- 十至二十五次
- 二十五以上
- 經常無法有效的找到

### 三. 多維度分析平台之推薦輔助系統滿意度

		非常滿意	不同意	沒意見	同意	非常同意
1.	推薦輔助系統能正確提供出您所需要的資訊?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.	推薦項目出現的順位是否符合您的期望?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	推薦輔助系統能否加速您的問題處理?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4.	推薦的過程中，是否也能協助您了解相同部門的人最近關心的資訊?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	推薦輔助系統是否能提高您使用多維度分析平台的意願?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	推薦輔助系統的操作介面與使用的流程的滿意?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

#### 四. 多維度分析平台之推薦輔助系統建議

請寫下對於系統的介面與功能上的任何建議，做為後續系統改善的使用：

