KM-14

# 協同過濾推薦系統多樣性之研究

**張東勝* 曾國城** 黃謙順*****

*中國文化大學資訊管理研究所 研究生 shengger.chang@gmail.com
**亞太創意技術學院多媒體與電腦娛樂科學系 教授 tikic@ms.apic.edu.tw
***中國文化大學資訊管理研究所 教授 cshwang@faculty.pccu.edu.tw

## 摘要

經營電子商務網站的企業提供推薦系統的目的主要在於增加銷售的機會。過往推薦系統的研究方向多著墨在單一推薦商品的精準度上，這樣的情形無法滿足一些具有多興趣範圍的使用者的使用滿意度。本研究於傳統協同過濾推薦程序中加入多樣性參數，希能增加推薦清單整體的多樣性以滿足使用者對推薦清單的滿意度和接受度
本研究的目的：1.推薦系統多樣性形成因素分析：增加多樣性因素使推薦商品清單具高多樣性特性的可能性分析。2.推薦系統多樣性方法的比較分析：本研究架構提出2個多樣性影響因素，分別於形成鄰居和產生推薦清單時置入多樣性影響因素與只於形成鄰居或產生推薦清單置入多樣性影響因素的方法的多樣性比較分析。

**關鍵字：協同過濾推薦系統、相似度、多樣性**

# 1. 緒論

## 1.1 研究背景

全球資訊網的爆炸性成長及新興電子商務的興起促使推薦系統的發展[1]，例如最著名的電子商務推薦系統應屬亞馬遜網絡書店（Amazon.com），當顧客選擇一本自己感興趣的書籍，馬上會在底下看到一行「Customer Who Bought This Item Also Bought」，各網絡書店也跟進做這樣的推薦服務如台灣的博客來網絡書店[8]，推薦系統確實在企業發展電子商務相關服務上扮演重要角色。

然而，經營電子商務網站的企業提供推薦系統的目的主要在於增加銷售的機會，希望經由顧客對推薦系統的推薦品質感到滿意提昇顧客對公司的忠誠度。當顧客沉浸於收到有相關聯的推薦項目時，同時也會發生如果推薦項目內容太相近會很快的對推薦服務失去興趣[3]。

如何有效的增加推薦清單整體多樣性變成是一個可解決推薦服務使用者滿意度的重要課題，推薦系統多樣性的目標是確認出具有不相似的推薦項目且推薦項目同時能涵蓋顧客的興趣[3]，其代表的意義就是在維持一定的商品推薦精準度之上仍能具有滿足顧客多樣化興趣的推薦商品清單。

## 1.2 研究動機

想像你正在使用一個電影推薦系統，假設所有的推薦電影項目都是同一個導演所導的喜劇電影，這樣的推薦結果對客戶的用處是很小的，你會放棄繼續使用這樣一個推薦系統[2]。

過往推薦系統的研究方向多著墨在單一推薦商品的精準度上，這樣的情形無法滿足一些具有多興趣範圍的使用者的使用滿意度。然而多樣性的問題不僅存在於內容式推薦系統中，也存在於協同過濾式推薦系統中。本研究於傳統協同過濾推薦程序中加入多樣性參數，希能增加推薦清單整體的多樣性以滿足使用者對推薦清單的滿意度和接受度，增加再次使用系統的機會。

## 1.3 研究目的

就我們所知，至今很少研究著重於改善推薦系統的多樣性。本研究於「形成鄰居」和「產生推薦清單」時，加入多樣性影響因素。本研究的目的整理如下：

- 推薦系統多樣性形成因素分析：

  增加多樣性因素使推薦商品清單具高多樣性特性的可能性分析。

- 推薦系統多樣性方法的比較分析：

  本研究架構提出2個多樣性影響因素，分別於形成鄰居和產生推薦清單時置入多樣性影響因素與只於形成鄰居或產生推薦清單置入多樣性影響因素的方法的多樣性比較分析。

## 1.4 研究範圍和限制

本研究範圍以MovieLens電影評分為例，資料來源為www.grouplens.org，共計100,000

筆有效資料，包含943位顧客對1682部電影的評分資料。

### 1.5 論文架構

本論文總共分為四個章節，第一章緒論將引入本研究的研究背景與動機，進而提出研究的主要目的，並說明研究範圍與限制。第二章為文獻探討，將針對推薦系統、多樣性等國內外參考文獻進行彙整與探討。第三章提出本研究的主要系統架構、模組。第四章為初步實驗結果，詳細說明本研究初步實驗成果。

## 2. 相關研究探討

### 2.1 推薦系統

在個人化推薦方法的研究領域中，學者Balabanovic和Shoham(1997)將推薦系統技術分為三個類型：1.內容式過濾方法(Content-based Filtering,CBF)；2.協同式過濾方法(Collaborative Filtering,CF)；3.混合式(Hybrid-based Filtering)等推薦技術。

協同過濾演算法可分為兩大演算法：1.基於使用者的協同過濾演算法(User-based collaborative filtering algorithm)；2.基於項目的協同過濾演算法(Item_based collaborative filtering algorithm)。

一、基於使用者的協同過濾演算法(User-based collaborative filtering algorithm)

典型的協同過濾演算法是基於使用者間相似度，用於許多電子商務系統的推薦機制。藉由選擇與目標使用者最接近的k位使用者和經由結合這些使用者的喜好形成預測結果。此種協同過濾技術包含兩個階段。

第一個階段：使用者相似度計算。計算目標使用者與其他使用者間的相似度，其中一種方法稱作Pearson 方法，計算公式如2-1

$$sim(x,y)=\frac{\sum_{i\in P_{co\_rated}}(R_{xi}-\overline{R_x})(R_{yi}-\overline{R_y})}{\sqrt{\sum_{i\in P_{co\_rated}}(R_{xi}-\overline{R_x})^2}\sqrt{\sum_{i\in P_{co\_rated}}(R_{yi}-\overline{R_y})^2}} \quad (2\text{-}1)$$

標楷體 $P_{co\_rated}$ 為使用者x與使用者y已經同樣給予評分的項目集合，即 $P_{co\_rated}\in \mathrm{R}_x\cap R_y$ 。 $R_x=\{R_{x1},R_{x2},\ldots,R_{xi}\}$ 與 $R_y=\{R_{y1},R_{y2},\ldots,R_{yi}\}$ 表示使用者x與使用者y的評分記錄，$i$ 代表項目編號，$R_{xi}$ 表示使用者x對項目 $P_i$ 的評分值；$\overline{R_x}$ 表示使用者x對已經評分過的所有項目的平均評分值；$R_{yi}$ 表示使用者y對項目 $P_i$ 的評分值；$\overline{R_y}$ 表示使用者y對已經評分過的所有項目的平均評分值。

第二個階段：預測結果計算。經由計算與顧客x相似的使用者的評分總和來計算顧客x對項目i的評分，計算公式如2-2

$$P_{x,i}=\overline{R_x}+\frac{\sum_{y\in p}sim(x,y)(R_{y,i}-\overline{R_y})}{\sum_{y\in p}|sim(x,y)|} \quad (2\text{-}2)$$

P代表k位最接近鄰居的集合。

最後，經由上述的預測結果產出Top_N推薦清單。

二、基於項目的協同過濾演算法(Item-based collaborative filtering algorithm)

基於使用者的協同過濾演算法會隨著鄰居使用者的增加而線性的增加計算的複雜度。基於項目的協同過濾演算法為避開此瓶頸藉由探索項目間的關聯。因為項目間的關係相對靜態。項目演算法提供相同的品質並需要較少的線上即時計算。

此種協同過濾技術包含兩個階段。

第一個階段：項目間相似度計算。其中一種方法稱作Pearson 方法，計算公式如2-3

$$sim(I_i, I_j) = \frac{\sum_{u \in Uij} (R_{ui} - \overline{R_u})(R_{uj} - \overline{R_u})}{\sqrt{\sum_{u \in Uij} (R_{ui} - \overline{R_u})^2} \sqrt{\sum_{u \in Uij} (R_{uj} - \overline{R_u})^2}} \quad (2\text{-}3)$$

項目i與項目j的相似度是由計算相關係數 $sim(i, j)$ 而得，先區隔出已經同樣給予項目i與項目j評分的使用者集合 $U_{ij}$，$i$ 代表項目編號；$R_{ui}$ 表示使用者對項目 $I_i$ 的評分值；$\overline{R_u}$ 表示u位使用者對項目i與項目j的平均評分值。

第二個階段：預測結果計算。經由計算與項目i相似的項目的評分總和來計算顧客x對項目i的評分，計算公式如2-4

$$P_{x,j} = \frac{\sum_{j \in p} sim(I_i, I_j) R_{x,j}}{\sum_{j \in p} | sim(I_i, I_j)} \quad (2\text{-}4)$$

最後，經由上述的預測結果產出Top_N推薦清單。

## 2.2 多樣性相關文獻

在個人化網頁搜尋領域中，Radlinski and Dumais提出評估多樣性搜尋結果改善使用者滿意度的方法，主要的概念是從搜尋結果中挑選出Top_K多樣性結果項目。

從搜尋服務演變至推薦服務後，在改善推薦系統多樣性的研究發展上，Kaith and Barry提出推薦多樣性是非常重要的而且傳統的推薦系統皆是低多樣性的，一個較典型的改善方法就是同時考量相似度與多樣性的推薦方法。

具備多樣性的推薦方法中，主要有兩種方法，第一種是在形成鄰居的過程中加入多樣性影響因素；第二種方法是在產生推薦清單時加入多樣性影響因素，這兩種方法都可以提高整體推薦清單的多樣性程度。

一、採用「形成鄰居」時，加入多樣性影響因素的研究如下：

Fuguo[1]提出一種考量相似度與多樣性的推薦方法，經由在信任模型中挑選多樣性鄰居的方法來提高使用者對推薦項目的滿意度，實驗結果反應其能提高推薦多樣性。

Ahu et al.[5]提出語意鄰居群(semantic neighborhoods)的概念，在形成鄰居集合的過程中除了參考使用者對項目的評分資料外也考慮使用者對項目的興趣數值來選擇正確

的鄰居，透過此種方式可增加整體推薦清單的多樣性。

二、採用「產生推薦清單」時，加入多樣性影響因素的研究如下：

Ziegler et al.[4]研究推薦系統的多樣性，提出一種藉由從候選項目中挑選出具有「最少清單內相似度」的推薦清單，其最能滿足使用者的多樣化興趣的推薦清單方法。結果顯示客戶較喜愛多樣性高的推薦清單，項目間採用分類方式，但是沒有考慮到推薦鄰居的多樣性。

David[6]提出Retrieval Network 概念，稱作R-Net，其演算法可增加推薦清單的多樣性，其方法為採用Smyth和McClave的Boundyed Greedy(BG)演算法，藉由選擇可同時增加相似度和多樣性的候選項目做為下一個推薦項目。

Gediminas Adomavicius et al. 提出參數化Item-Based 評分方法，透過設定評分門檻值(Threshold)，使得維持一定的推薦準確性中仍能保有推薦項目的多樣性，藉此在準確性與多樣性中保持一定的平衡效果。

上述多樣性相關研究都只採用一種增加推薦系統多樣性的方法，本研究則同時採納二種增加推薦系統多樣性的方法，即在形成鄰居和產生推薦清單時加入多樣性影響因素。

## 3. 研究方法

### 3.1. 系統架構

本研究架構供包含三個步驟：將電影評分資料載入系統，用具備相似度與多樣性的混合權重的方法找出以使用者為基礎的K位鄰居後，再根據這K位鄰居所評的電影的評分分數，用具備相似度與多樣性的混合權重方法篩選出以項目為基礎的Top_N的電影「推薦清單」。

本研究所提出的推薦系統運作架構與流程(見圖3-1)。

本研究所提出的系統中，採用一般資訊系統運作模式，分成輸入、處理及輸出三個主要程序，分述如下：

一、輸入

本研究先將所有資料載入系統。

二、處理

處理階段由形成鄰居模組合及推薦模組所組成。

三、輸出

系統的輸出即為由推薦模組提供給目標使用者的推薦清單，依據K位鄰居所評的電影評分分數，找出Top_N部電影推薦清單，提供一串的電影清單供使用者選擇的依據。

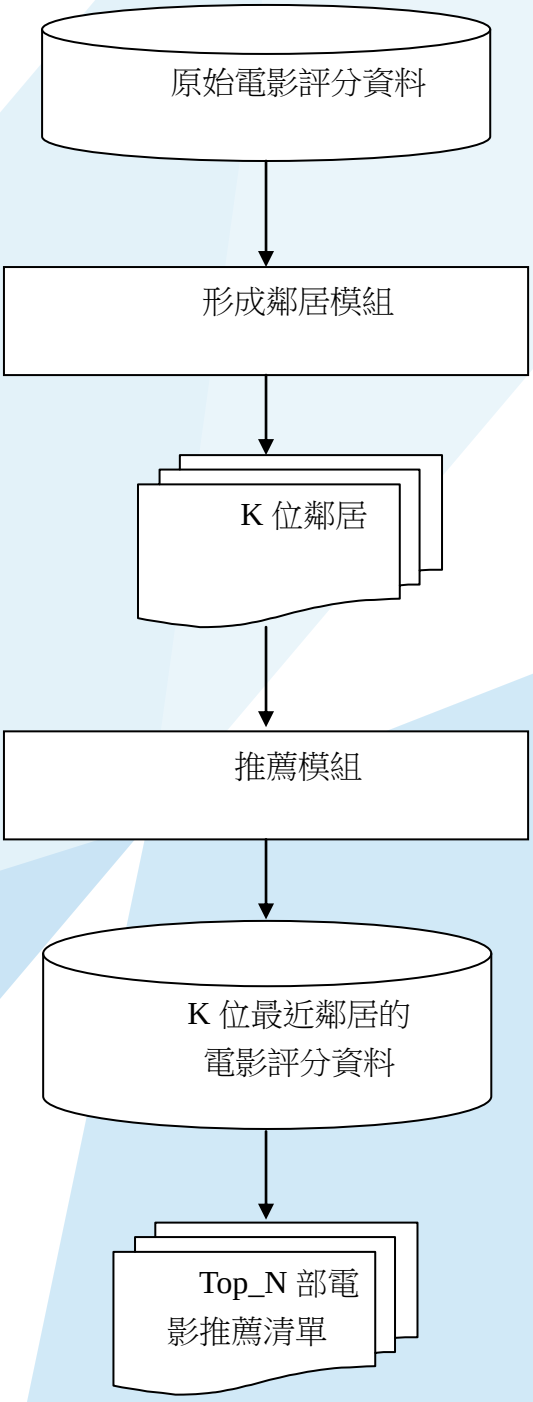


圖3-1 系統架構圖

### 3.2. 形成鄰居模組

本研究之形成鄰居模組採用結合相似度與多樣性之混合權重法計算並取得K位鄰居。

相似度依據使用者的電影評分資料計算目標使用者與其他使用者間的相似度，即sim(x,y)，相似度計算演算法採用Person 方法，計算公式3-1如下所示

$$sim(x,y)=\frac{\sum_{i\in P_{co\_rated}}(R_{xi}-\overline{R_y})}{\sqrt{\sum_{i\in P_{co\_rated}}(R_{xi}-\overline{R_x})^2}\sqrt{\sum_{i\in P_{co\_rated}}(R_{yi}-\overline{R_y})^2}} \quad (3\text{-}1)$$

$P_{co\_rated}$ 為使用者x與使用者y已經同樣給予評分的項目集合，即 $P_{co\_rated}\in \mathrm{R}_x\cap R_y$ 。$R_x=\{R_{x1},R_{x2},\ldots,R_{xi}\}$ 與 $R_y=\{R_{y1},R_{y2},\ldots,R_{yi}\}$ 表示使用者x與使用者y的評分記錄，$i$ 代表項目編號，$R_{xi}$ 表示使用者x對項目 $P_i$ 的評分值；$\overline{R_x}$ 表示使用者x對已經評分過的所有項目的平均評分值；$R_{yi}$ 表示使用者y對項目 $P_i$ 的評分值；$\overline{R_y}$ 表示使用者y對已經評分過的所有項目的平均評分值。

樣性計算演算法的計算公式3-2如下所示

$$div(x,y)=1-sim(x,y) \quad (3\text{-}2)$$

最後藉由混合權重方法取得最高權重值的K位鄰居。

步驟1：將電影評分資料載入系統後，先找出最相似的2K位鄰居。

步驟2：先設置相似度最高的使用者 $u_{i1}$ ，$T=\{u_{i1}\}$ 。

步驟3：再根據混合權重來決定下一位使用者 $u_i$ ，計算公式3-3、3-4、3-5如下所示

$$sim(T,u_i)=\frac{\sum_{u\in T}sim(u,u_i)}{\frac{N(N-1)}{2}} \quad (3\text{-}3)$$

$$div(T,u_i)=\frac{\sum_{u\in T}div(u,u_i)}{\frac{N(N-1)}{2}} \quad (3\text{-}4)$$

$$W(T,u_i)=\lambda sim(T,u_i)+(1-\lambda)div(T,u_i) \quad (3\text{-}5)$$

### 3.3. 推薦模組

本研究之推薦模組採用結合相似度與多樣性之混合權重法計算並取得Top_N推薦清單。

相似度依據K位鄰居的電影評分資料計算電影項目間的相似度，即 $sim(I_i,I_j)$ ，相似度計算演算法採用Person 方法，計算公式3-6如下所示

$$sim(I_i,I_j)=\frac{\sum_{u\in Uij}(R_{ui}-\overline{R_u})(R_{uj}-\overline{R_u})}{\sqrt{\sum_{u\in Uij}(R_{ui}-\overline{R_u})^2}\sqrt{\sum_{u\in Uij}(R_{uj}-\overline{R_u})^2}} \quad (3\text{-}6)$$

項目i與項目j的相似度是由計算相關係數 $sim(I_i,I_j)$ 而得，先區隔出已經同樣給予項

目i與項目j評分的使用者集合$U_{ij}$，$i$代表項目編號；$R_{ui}$表示使用者對項目$I_i$的評分值；$\overline{R_u}$表示u位使用者對項目i與項目j的平均評分值。

多樣性依據電影項目所屬類別計算目標電影項目與其他電影項目間的多樣度，即$div(I_i,I_j)$。

最後藉由混合權重方法取得具有最高混合權重數值的Top_N電影項目，即Top_N電影推薦清單。

步驟1：將k位鄰居所評的電影項目中，找出相似度最高的2N個電影項目。

步驟2：先設置推薦分數最高的項目$I_1$，$T=\{I_1\}$。

步驟3：再根據混合權重來決定下一個項目$I_i$，計算公式3-7、3-8、3-9如下所示

$$sim(T,I_i)=\frac{\sum_{I\in T} sim(I,I_i)}{\frac{N(N-1)}{2}} \tag{3-7}$$

$$div(T,I_i)=\frac{\sum_{I\in T} div(I,I_i)}{\frac{N(N-1)}{2}} \tag{3-8}$$

$$W(T,I_i)=\lambda sim(T,I_i)+(1-\lambda)div(T,I_i) \tag{3-9}$$

本研究在類別多樣性演算法中，將採用電影分類目錄喜好分數(interest score)的計算方式，藉此計算電影項目的多樣性數值。例如電影分類架構如圖3-4：

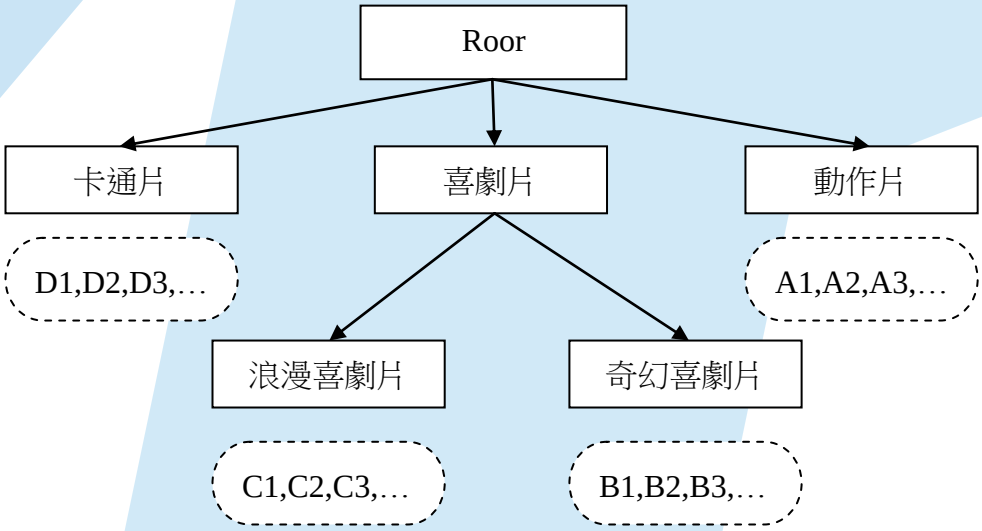


圖3-4 電影分類架構圖

如果使用者a喜愛項目C1，不只代表使用者a喜愛浪漫戲劇片也同時表示使用者喜愛喜劇片，但是喜劇片的分類目錄喜好分數會隨著與其距離整體電影分類架構的葉端(leaf node)而遞減。

# 4. 研究結果

## 4.1. 實驗設計

本研究使用 Movielens 的資料驗證多樣性因素對推薦清單具多樣性特性的影響，以 MovieLens 電影評分為例，資料來源為 www.grouplens.org，共計 100,000 筆有效資料，包含 943 位顧客對 1682 部電影的評分資料。

本研究將資料分為訓練組與測試組，943 位顧客評分資料中的 200 位顧客評分資料做為測試組，其餘 743 位顧客評分資料為訓練組。研究中所使用的參數，分述如下：

一、最近鄰居個數：10、20、30、40、50 位

二、權重：1、0.75、0.5、0.25

三、推薦個數：10、20、30、40、50

## 4.2. 實驗結果

於本研究之形成鄰居模組，其主要功能是先找出與目標對象最接近的 k 位鄰居，再經由推薦模組，其主要功能為由與目標對象最接近的 k 位鄰居預測目標對象可能會喜歡的 TopN 電影推薦清單。

實驗方向為探討不同最近鄰居個數(K)、不同權重(λ)和不同 TopN 推薦清單對推薦清單多樣性的影響。

4.2.1.不同最近鄰居個數和不同權重對多樣性的影響

探討不同最近鄰居個數(K)和不同權重(λ)對推薦清單多樣性的影響，如表 4-1

表4-1　不同最近鄰居個數和不同權重的多樣性

| K \ λ | 1 | 0.75 | 0.5 | 0.25 | 0 |
|---|---|---|---|---|---|
| 10 | 0.60 | 0.60 | 0.60 | 0.54 | 0.54 |
| 20 | 0.64 | 0.64 | 0.57 | 0.59 | 0.59 |
| 30 | 0.60 | 0.60 | 0.50 | 0.57 | 0.57 |
| 40 | 0.63 | 0.63 | 0.64 | 0.57 | 0.57 |
| 50 | 0.57 | 0.57 | 0.60 | 0.65 | 0.65 |

在實驗固定最近鄰居個數下，不同權重對於推薦清單多樣性的影響。於最近鄰居個數(K)=10、20、30 和 40 時，權重(λ)=1 和 0.75 的多樣性皆高於權重(λ)=0.25 和 0；相反狀況發生在最近鄰居個數(K)=50 時，權重(λ)=1 和 0.75 的多樣性皆低於權重(λ)=0.25 和 0。

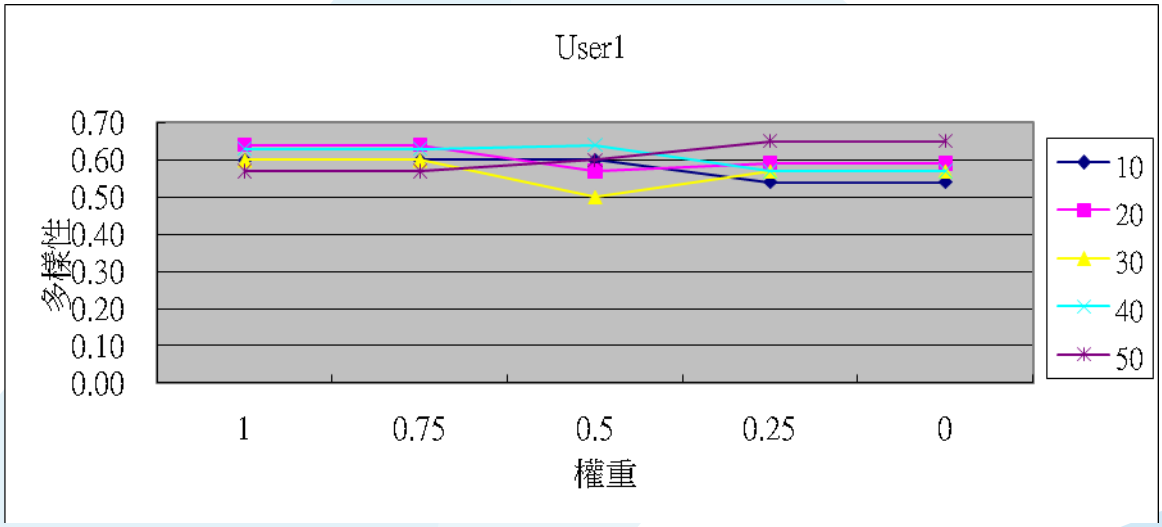


圖1　固定最近鄰居個數之不同權重的多樣性

4.2.2.不同 TopN 和不同權重對多樣性的影響

探討不同 TopN 和不同權重(λ)對推薦清單多樣性的影響，如表 4-2

表4-2 不同TopN和不同權重的多樣性

| N \ λ | 1 | 0.75 | 0.5 | 0.25 | 0 |
|---|---|---|---|---|---|
| 10 | 0.56 | 0.56 | 0.50 | 0.50 | 0.50 |
| 20 | 0.56 | 0.56 | 0.52 | 0.59 | 0.59 |
| 30 | 0.60 | 0.60 | 0.50 | 0.57 | 0.57 |
| 40 | 0.60 | 0.60 | 0.52 | 0.55 | 0.55 |
| 50 | 0.61 | 0.61 | 0.50 | 0.56 | 0.56 |

在實驗固定 TopN 下，不同權重對於推薦清單多樣性的影響。於 TopN (N)=10、20、30、40 和 50 時，權重(λ)=1 和 0.75 的多樣性皆高於權重(λ)=0.5、0.25 和 0。

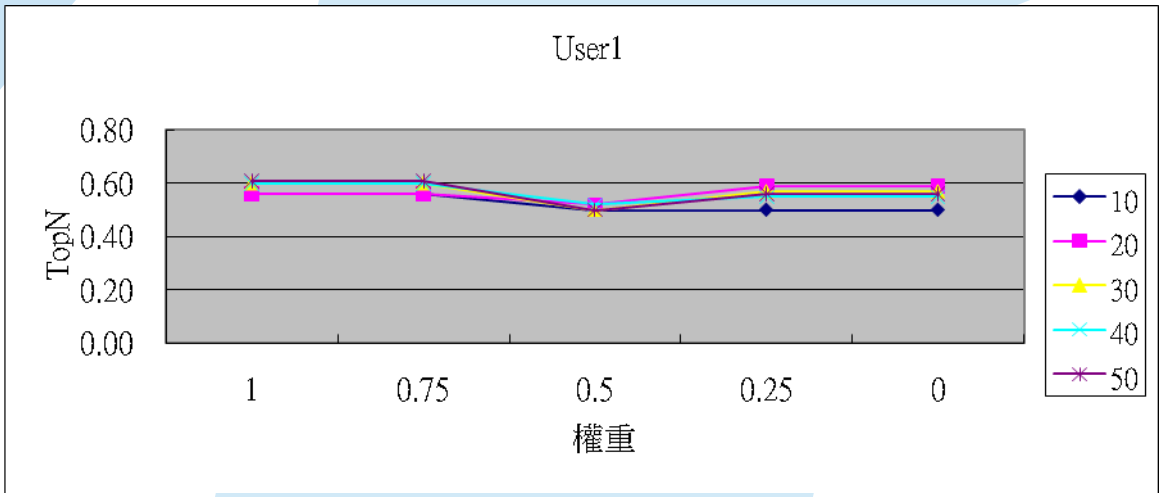


圖1　固定TopN之不同權重的多樣性

## 5. 結論

根據本研究從不同觀點之實驗結果比較，可以得到下列結論：

一、只有當最近鄰居個數(K)=50，才能夠顯現在形成鄰居模組時置入多樣性影響因素會對整體推薦清單的多樣性具有較佳的影響。

二、本研究的 TopN 介於 10 至 50，在所有 TopN 的情況下於形成鄰居模組置入多樣性影響因素皆不會對推薦清單的多樣性具有影響。

## 6. 參考文獻

[1] Fuguo, Zhang., “Research on Recommendation List Diversity of Recommender Systems.” *Proceedings of the 2008 International Conference on Management of e-Commerce and e-Government*, 2008,pp. 72-76.

[2] Fuguo, Zhang., “Research on Recommendation List Diversity of Recommender Systems.” *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference*, 2009,pp. 222-225.

[3] Cong Yu, Lakes V.S. Lakshmanan,Sihem Amer-Yahia., ”Recommendation Diversification Using Explanations Data Engineering,” *ICDE '09. IEEE 25th International Conference on*,2009,pp. 1299-1302.

[4] N. Ziegler,S. M. McNee,J. A. Konstan,G. Lausen., “Improving recommendation lists throuth topic diversification.” *Proceedings of the 14th international conference on World Wide Web*,2005.

[5] Ahu Sieg,Bamshad Mobasher,Robin Burke., “Improving the effectiveness of collaborative recommendation with ontology-based user profiles.” *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems.*,2010.

[6 ]David McSherry(2002)., “Recommendation Engineering.” *Proceedings of 15th European Conference on Artificial Intelligence.*,2002,pp. 86-90.

**中文博、碩士論文**

[7] 蔡怡萍，民 98，運用序列分群於產品推薦之研究，中國文化大學資訊管理研究所碩士論。

**中文網站**

[8] Wiki，民 100，http://zh.wikipedia.org/zh-hk。