國　立　成　功　大　學

資 訊 管 理 研 究 所

碩 士 論 文 計 劃 書

產品集的個人化推薦：以美妝保養社群網站為例

審查委員：

林明毅博士　劉任修博士

研 究 生：黃琬玲

指導教授：王惠嘉 博士

地 點：61208教室

日 期：2016.04.28 13:10~14:00

**摘要**

隨著網路技術發展，除了社群概念的興起外，許多觀念也透過網際網路等傳播媒體逐漸普及：美妝保養便是一個很好的例子。美妝保養產品原本被視為是附加於消費者身上的奢侈品，但隨著資訊的傳播，使得大量與美妝保養相關的資訊不停被置入至大眾的日常生活中，人們因此越來越重視美妝保養。但隨著美妝保養資訊的大量增加，造成使用者資訊過載，使用者無法消化龐大資訊，且過去在此領域中並無相關之資訊過濾系統協助使用者過濾出符合其需求之資訊，因此使用者需要一資訊過濾系統協助取得符合其需求的美妝保養資訊。

又因美妝保養產品的使用特性，使得大部分消費者購買美妝保養產品時，不會單獨購買其中一類產品，而是會按其保養程序，購買不同類別的產品搭配使用。這樣的消費特性使得美妝保養產品具有集合的性質，舉例來說：大部分的消費者在購買化妝水後，會再購買精華液、面膜，或者是乳霜等產品，因此在進行產品推薦時，以產品集合的形式進行推薦會較符合使用者需求。

為了能夠有效且正確的推薦美妝保養社群網站的使用者可能適合其膚質的美妝保養產品集合，本研究蒐集美妝保養社群網站上之資料，設計一適用於美妝保養領域的混合式產品集推薦系統：以使用者特性、其使用過之產品，利用餘弦相似度計算使用者間相似度，並利用K-Medoids分群法進行分群。分群後之再蒐集群內使用者使用過的產品，建立產品類別的連結，找出該群使用者所使用的產品類別關係，以建立產品集的雛形。有了產品類別關係後便可以如以往的研究一樣針對每一類別之產品進行推薦。除此之外，本研究利用使用者撰寫過之心得評論進行分析，取得群內使用者對於產品之評分後，對產品進行排序，找出前n個產品作為推薦清單，進而進行跨類別美妝保養產品集合推薦，以協助使用者省去過濾資訊的時間並推薦出符合其特質之產品。

關鍵字：個人化推薦系統、產品集合推薦

**目錄**

[第1章 緒論 1](#_Toc448679283)

[1.1 研究背景與動機 3](#_Toc448679284)

[1.2 研究目的 7](#_Toc448679285)

[1.3 研究範圍與限制 8](#_Toc448679286)

[1.4 研究流程 9](#_Toc448679287)

[1.5 論文大綱 10](#_Toc448679288)

[第2章 文獻探討 12](#_Toc448679289)

[2.1 社群網路服務 12](#_Toc448679290)

[2.2 自然語言處理 13](#_Toc448679291)

[2.2.1 中文斷詞處理 13](#_Toc448679292)

[2.2.2 詞性標記 15](#_Toc448679293)

[2.3 文件分析 17](#_Toc448679294)

[2.3.1 向量空間模型 17](#_Toc448679295)

[2.3.2 使用者相似度計算 18](#_Toc448679296)

[2.4 情感分析 19](#_Toc448679297)

[2.5 分群 21](#_Toc448679298)

[2.6 推薦系統 23](#_Toc448679299)

[2.7 小結 26](#_Toc448679300)

[第3章 研究方法 27](#_Toc448679301)

[3.1 研究架構 27](#_Toc448679302)

[3.2 資料前處理 32](#_Toc448679303)

[3.3 使用者分群 33](#_Toc448679304)

[3.4 產品集合 37](#_Toc448679305)

[3.5 產品集合推薦 39](#_Toc448679306)

[3.6 小結 42](#_Toc448679307)

[第4章 系統建置與驗證 44](#_Toc448679308)

[4.1 系統環境建置 44](#_Toc448679309)

[4.2 實驗方法 46](#_Toc448679310)

[4.2.1 資料來源 47](#_Toc448679311)

[4.2.2 評估指標 47](#_Toc448679312)

[4.3 實驗結果 47](#_Toc448679313)

[4.3.1 實驗一 47](#_Toc448679314)

[4.3.2 實驗二 47](#_Toc448679315)

[4.3.3 實驗三 47](#_Toc448679316)

[4.3.4 實驗四 47](#_Toc448679317)

[第5章 目前及預期進度 47](#_Toc448679318)

[參考文獻 48](#_Toc448679319)

**表目錄**

表 2‑1 CKIP部分詞性標記之意義 17

表 2‑2 分群方法比較表(林宜瑩, 2010) 23

表 3‑1 使用者與產品使用關係矩陣 37

表 3‑2 使用者間相似度矩陣 37

表 3‑3 分群後使用者與產品關係範例 39

表 3‑4 使用者使用之產品名稱對照表 39

表 3‑5 使用者使用之產品類別歸納範例 41

表 3‑6使用者所使用之產品類別歸納範例 41

表 4‑1系統建置環境 46

表 4‑2 資料集分佈表 49

**圖目錄**

圖 1‑1 Amazon.com推薦畫面示意圖 2

圖 1‑2 獲取美妝保養資訊管道(InsightXplorer創市際市場研究顧問, 2015a) 4

圖 1‑3 產品集示意圖 7

圖 1‑4 研究流程圖 11

圖 2‑1 Stanford Parser中文斷詞範例 15

圖 2‑2 CKIP詞性標記範例 17

圖 2‑3 廣義知網範例圖示 22

圖 2‑4 推薦系統的種類(Isinkaye et al., 2016) 25

圖 3‑1 產品類別示意圖 29

圖 3‑2 資料集結構圖示 31

圖 3‑3 研究架構圖 33

圖 3‑4 使用者膚質分類示意圖 34

圖 3‑5 使用者資料範例 35

圖 3‑6 心得評論斷詞範例 35

圖 3‑7 K-medoids分群演算法 38

圖 3‑8 產品集合演算法 40

圖 3‑9 單篇心得評論情緒分數計算圖示 42

圖 4‑1 UrCosme網頁局部圖示 49

# 緒論

隨著科技蓬勃發展，透過網路，不僅使得資訊傳遞更加迅速，也讓使用者更易於取得即時訊息。在網際網路架構而成的世界裡，人們利用網際網路技術在虛擬世界中聚集並分享知識或是對於特定人事物的評論，這樣來往分享的互動關係，使得單向傳遞的資訊逐漸變成雙向互動的資訊流通在網路世界中，進而形成了社群網路(social network)。

社群網路觀念的出現，不僅影響了人們的通訊方式及日常生活，也影響了許多產業的發展模式。在早期，美妝保養產品被視為是附加的奢侈品，但隨著科技進步、時代更迭以及人們思維的改變，美妝保養觀念逐漸在人們日常生活中佔有重要的一席之地。過去，人們必須親自進到百貨公司詢問專櫃人員，或是透過親朋好友的口耳相傳，才能取得美妝保養的相關資訊；如今，隨著網路技術的發展，許多資訊透過網路便能即時地傳遞到網路上的各個角落，美妝保養資訊也亦然。

這樣的轉變使得網路上與美妝保養相關的資訊以驚人的速度增加，造成許多使用者難以消化這些大量的資訊，而這樣的難題使得資訊過濾系統因此被提出，以協助使用者濾除不必要的資訊。其中，推薦系統便是一種資訊過濾系統，能在龐大資訊流中過濾出尚未被發現的重要資訊，並能對特定使用者提供決策協助。以往的推薦系統進行產品推薦時，所產出的結果通常是與特定目標具有相關性的產品，且大部分的產品為獨立推薦，例如在Amazon.com上購買美妝保養產品時，其推薦系統是針對其他同樣購買該美妝保養產品的使用者之購物或瀏覽紀錄進行推薦，但系統並不會推薦出一個具有產品類別關係的產品集合給使用者，如圖 1‑1所示。

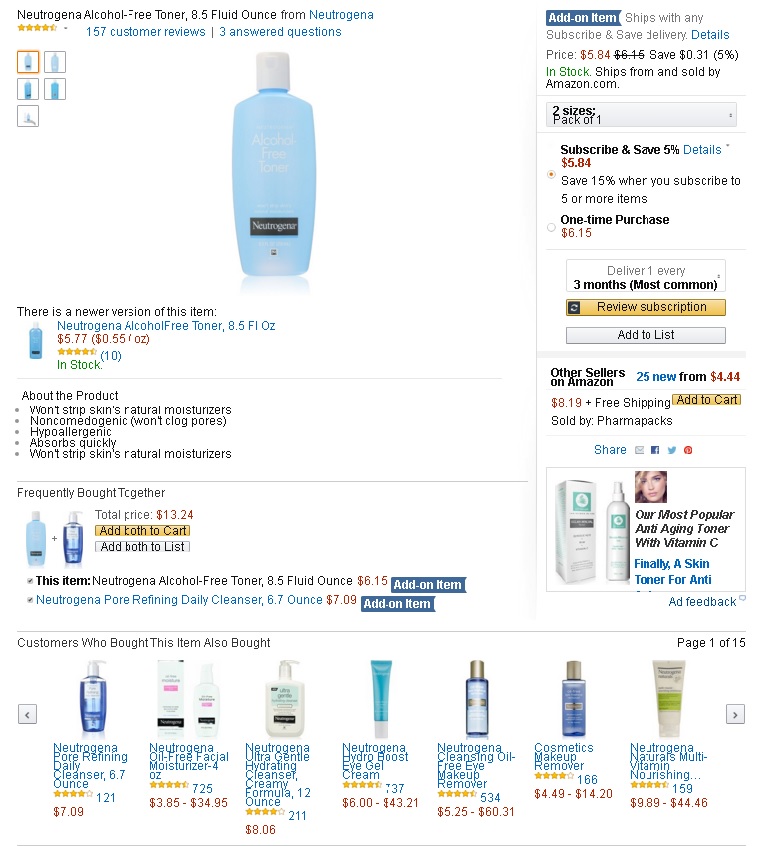


圖 1‑1 Amazon.com推薦畫面示意圖

美妝保養產品的購買與其他類型的產品較為不同，其不同之處為：大部分消費者購買美妝保養產品時，較不會單獨購買其中一類產品，而是會按照其保養程序，購買不同類別的產品搭配使用([瑞麗美人國際媒體, 2014](#_ENREF_43))，使得美妝保養產品具有集合的性質，舉例來說：大部分的消費者在購買化妝水後，會再購買精華液、面膜，或者是乳霜等產品，因此傳統的推薦系統推薦單一相關產品的特性並無法完全滿足美妝產品的推薦需求。且受到網路社群的影響，越來越多消費者在進行購物前會先參考網路上其他使用者的心得或是對於產品所發表的評論作為購買決策的參考依據，這種現象尤其在尚未購買過類似產品的使用者身上最為常見。

　　因此本研究旨在利用美妝保養社群平台上的使用者資訊及其發表之產品評論，設計一個推薦系統，將推薦之結果集結形成一產品集的完整推薦清單，希望可以藉此推薦出符合使用者需求的美妝保養產品集合。

## 研究背景與動機

隨著科技的發展以及網際網路普及，網路已成為使用者互相連結並分享大量資訊的地方([Morente-Molinera, Pérez, Ureña, & Herrera-Viedma, 2016](#_ENREF_25))。透過網路，不僅能使資訊傳遞更加迅速，也能讓使用者更易於取得更多即時資訊。人們透過網際網路技術在網路上聚集並分享知識、趣聞，抑或是對於特定人事物的評論，使得資訊的流動從單向轉為雙向，進而形成了社群網路(social network)。在這些來往的雙向互動中，網際網路扮演著傳遞資訊的角色，也讓使用者可以更便利地取得他人的評論或意見。近幾年，社群的觀念逐漸成形，成為人們日常生活中不可或缺的存在，並深深影響現實生活中人們的思想、意識、文化等([維基百科, 2015c](#_ENREF_48))。

現今網路上充滿許多的網路平台，有些平台單純透過社群概念的建立形成巨大的網絡，例如：Facebook、Twitter，或是新浪微博；有些平台則具有主題性，即是指在這些平台上的社群皆是透過討論特定主題的使用者群聚而成，例如：以音樂為主題的Last.fm或是美妝保養相關知識為主題的UrCosme及FashionGuide等。透過社群網路，人們不僅可以分享資訊，更可以向他人尋求建議，進而作為決策時的參考。根據資策會MIC於2014年針對台灣地區使用網路的消費族群進行「網路社群使用」之調查結果顯示，有81%的消費者在購物前會搜尋口碑訊息([資策會產業情報研究所, 2014](#_ENREF_44))。

隨著網路技術發展，除了社群概念的興起外，許多觀念也透過網際網路等傳播媒體逐漸普及：美妝保養便是一個很好的例子。美妝保養產品原本被視為是附加於消費者身上的奢侈品，但隨著資訊的傳播以及人們觀念的轉變，加上許多開架式品牌不斷創立，使得大量與美妝保養相關的資訊不停被置入至大眾的日常生活中，人們因而越來越重視美妝保養，且會對於相關產品的使用效果保有一定的期望。

早期的大部分消費者在購買美妝保養產品時，會選擇進到百貨公司參考專櫃人員的說明後再考慮是否購買；近幾年，藥妝店在台灣快速成長普及，消費者便從專櫃消費轉往鄰近住家的藥妝店購買開架式產品，藥妝店也會有專業人員可以讓消費者進行諮詢，以作為消費決策的參考。網際網路出現後，因其具有立即性與全球性的特性，使得越來越多的消費者在購美妝保養商品時，除了參考電視廣告、平面標語或是百貨公司、藥妝店專員的意見外，也逐漸轉向透過網際網路尋求其他消費者使用產品過後的評論，並加以比較後才會決定是否購買([InsightXplorer創市際市場研究顧問, 2015a](#_ENREF_13))。圖 1‑2為創市際市場研究顧問於2015年3月針對如何獲取美妝保養資訊所進行的調查，由此調查結果可以發現有51.6%的美妝保養產品使用者透過網路取得美妝保養的相關資訊。

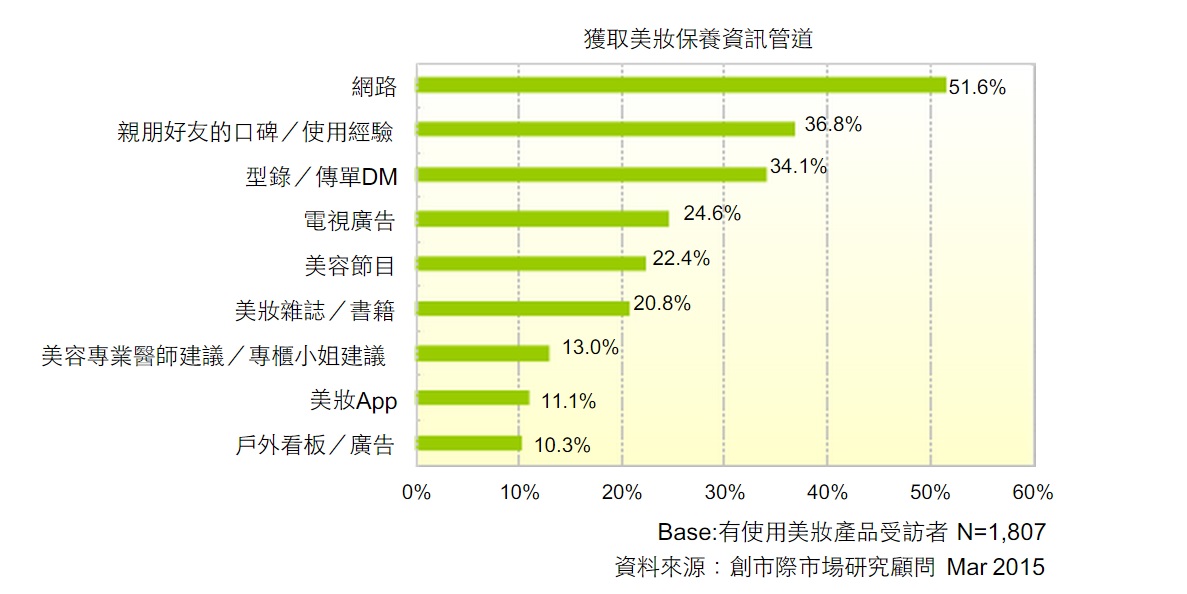


圖 1‑2 獲取美妝保養資訊管道([InsightXplorer創市際市場研究顧問, 2015a](#_ENREF_13))

根據前文所述，網際網路的出現以及社群觀念的形成改變了消費者的消費習慣，越來越多的網路平台也傾向藉由使用者的主動參與建構出屬於其平台的社群。在這些平台上，使用者在註冊成為會員後不僅可以瀏覽其他使用者的心得評論，也可以在平台上發表自己心得評論。以美妝保養社群網站為例，使用者在註冊成為會員後不僅可以瀏覽其他使用者對於特定產品的使用心得或評論，也可以在平台上發表自己對於特定產品的使用心得或評論。另外，在這些平台上普遍都具有搜尋功能，使用者可以依據自己的需求進行產品資訊與使用心得的搜尋；但儘管平台上之搜尋的功能已相當完善，有些平台卻缺乏推薦產品的功能或是其推薦功能尚有可以改進的空間，例如美妝保養產品其實具有跨產品類別的集合特性，但是在平台上卻少見推薦結果為ㄧ跨類別的集合。

　　推薦系統雖已是被廣泛討論研究的議題，但其存在也被證實是人們發現或找尋與其需求相關的資訊、產品，或是服務的重要管道([Garcia Esparza, O’Mahony, & Smyth, 2012](#_ENREF_11))。隨著網際網路的蓬勃發展，資訊量增加快速，造成許多使用者難以消化這些大量的資訊，這樣的難題使得資訊過濾系統因此被提出，以協助使用者濾除不必要的資訊。推薦系統即是一種資訊過濾系統，能將一些尚未被發現的資訊過濾出來，以對特定使用者提供決策建議([Isinkaye, Folajimi, & Ojokoh, 2016](#_ENREF_15); [Zahra et al., 2015](#_ENREF_37))。

推薦系統的目的即是要幫助使用者在許多選擇中決定出最適方案；而根據其資訊過濾的方式又可分為多種推薦系統([Garcia Esparza et al., 2012](#_ENREF_11))。根據[Zhang, Hu, Chen, Chen, and Shi (2012)](#_ENREF_38" \o "Zhang, 2012 #16)及[Isinkaye et al. (2016)](#_ENREF_15)指出，發展最為成熟且最常被使用的推薦系統類型為協同過濾式推薦系統(collaborative filtering recommendation system)，其最大的優點在於不需要對推薦對象的內容特徵屬性進行分析，且在一定資料範圍的情況下，具有高品質的推薦結果。但協同過濾式推薦系統仍然會面臨一些問題，導致推薦結果不佳，因此[Kumar and Fan (2015)](#_ENREF_17)提出了ㄧ種混合式方法，能在透過其他資源的輔助下達到更佳的個人化推薦結果。故本研究將以混合式推薦系統作為推薦工具：以使用者的個人資訊（如其膚質狀況及其曾經使用過之產品）進行過濾後，可以排除群內使用者未曾使用的產品，將「使用者—產品矩陣」中的產品維度進行降維，再以使用者對產品之心得評論進行產品推薦。

此外，以往推薦系統進行產品推薦時，最後的推薦結果都是與特定目標具有相關性的單一產品，且大部分的產品為獨立推薦；然而因為美妝保養產品的使用特性，使得大部分消費者購買美妝保養產品時，不會單獨購買其中一類產品，而是會按其保養程序，購買不同類別的產品搭配使用。這樣的消費特性使得美妝保養產品具有集合的性質，舉例來說：大部分的消費者在購買化妝水後，會再購買精華液、面膜，或者是乳霜等產品，因此在進行產品推薦時，以產品集合的形式進行推薦會較符合使用者需求。根據前文所述，消費者在購買美妝保養產品時，會網路尋求購買決策的參考，使得社群網路所產生之資訊有其可利用性。因此，若是能有效利用美妝保養社群網站上所取得之資料，蒐集使用者資訊，先以此對使用者進行分類，再以其所用過之產品作為特徵，將使用者進行分群後，分析群中使用者大部分所使用的美妝保養產品，且又因為群中每位使用者所使用的產品並不會被侷限於一特定類別中，而是會分別落在不同的類別中，舉例來說：曾經使用過洗面乳並發表心得的使用者，也可能會在使用過某牌乳霜後發表心得；我們便可以利用此特性，將群中使用者所使用過之產品類別形成產品集合，再分析每個產品的心得，得到群內使用者對於該產品的情緒分數後，以產品被該群使用者之使用次數作為某產品類別下之產品排序的權重，最後將情緒分屬及使用次數加以計算後得到使用者對產品的最後分數後進行排序，並以達到門檻值的前n個產品作為推薦結果，如圖 1‑3所示。

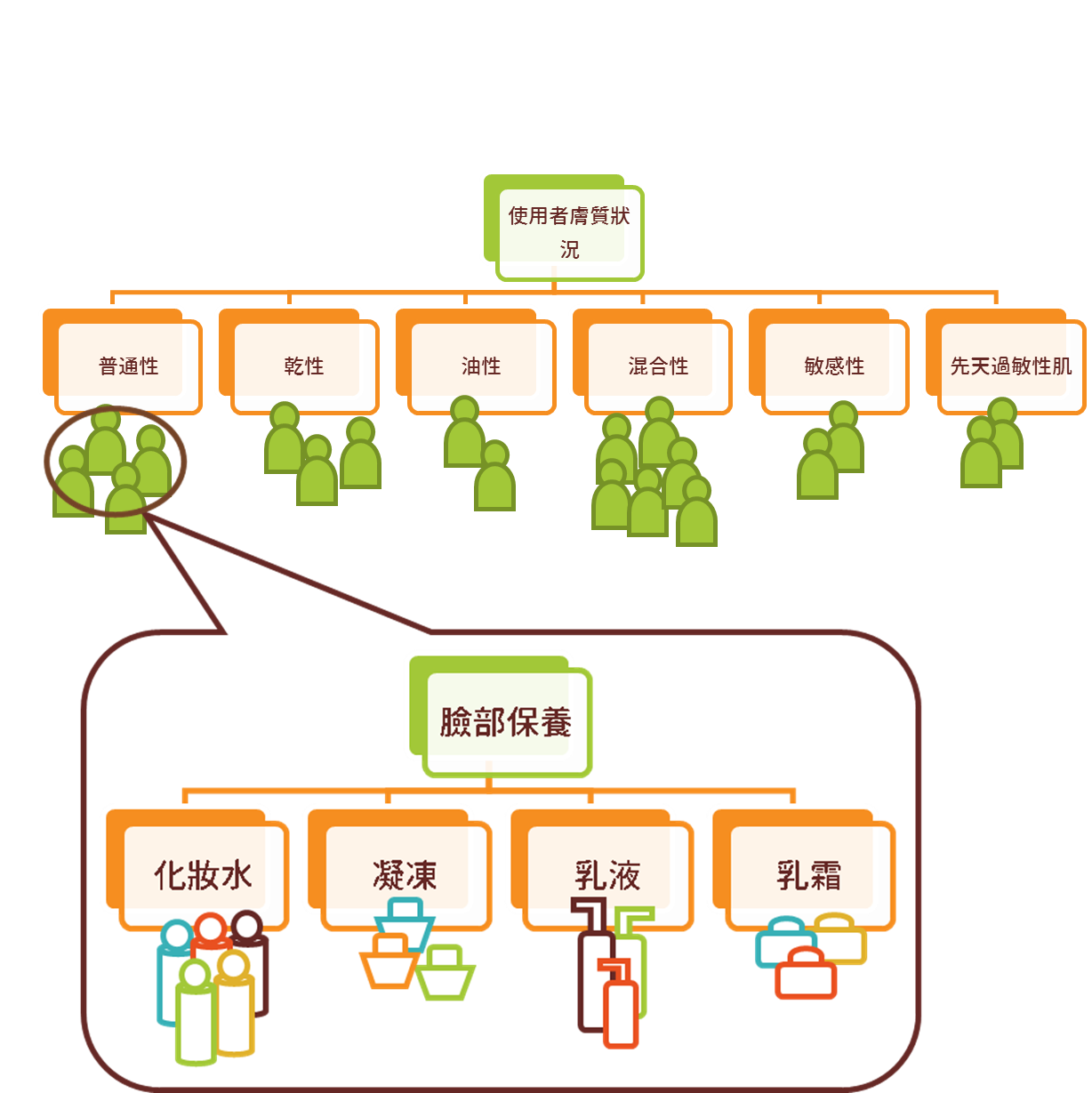


圖 1‑3 產品集示意圖

## 研究目的

基於上述之研究背景與動機，因美妝保養資訊的大量增加，造成使用者資訊過載，使得使用者無法消化龐大資訊，且在此領域中並無相關之資訊過濾系統協助使用者過濾出符合其需求之資訊，因此使用者需要一個資訊過濾系統協助取得符合其需求的美妝保養資訊。為了能夠有效且正確的推薦美妝保養社群網站的使用者可能適合其膚質的美妝保養產品，本研究蒐集美妝保養社群網站上之資料，設計一個適用於美妝保養領域的推薦系統，以使用者特性、其使用過之產品及對產品撰寫過之心得評論為依據，將使用者依其採購習慣方式進行相似度計算後分群，並利用分群的結果，進行產品評論分析，取得群內相似使用者對於產品之評分後，進行跨類別美妝保養產品集合推薦，以協助使用者省去過濾資訊的時間並推薦出符合其膚質之產品。

另外，在計算使用者相似度的部分，若單純只使用使用者特性，只能將之視為是分類後的結果，可能會使得推薦結果不夠準確，所以將加入使用者使用過之產品作為相似度計算依據。此處假設使用者對產品發表評論或是標記曾購買過該產品都視為是曾經使用該產品。系統主要目標如下：

1. 利用使用者相似度計算以及使用者分群結果，找出美妝保養產品中類別與類別間的關係，以形成未來推薦的產品集，例如：在中性肌膚質下的相似使用者經過分析後，得到相似使用者使用之產品類別關係為化妝水類－乳霜類－乳液類。
2. 利用相似使用者對於產品的心得評論，找出相似使用者對於該產品之意見字詞。意見字詞會有不同程度的差別，本研究將意見字詞之程度稱為情緒分數，透過此分數可以了解使用者對該產品的喜好程度。故本研究欲分析出相似使用者對於產品的平均情感分數，以此作為類別下產品的排序依據。
3. 有鑒於在過去研究中，沒有以集合形式的產品推薦系統，因此為了達到更個人化的推薦結果，本研究利用美妝保養社群資料中的使用者資料對使用者進行分類及分群。因為美妝保養產品的使用特性，使用者會傾向於接受與其特性相似之使用者的產品使用心得評論，因此若先將使用者進行分類再進行分群，不僅可以降低「使用者－產品矩陣」的維度，也能得到與目標使用者相近的群體所產出的推薦結果。

## 研究範圍與限制

　　本研究欲針對美妝保養社群相關網站之資料進行分析，故本研究有以下研究範圍、假設與限制：

1. 本研究之研究資料為美妝保養相關之社群網站，故研究範圍將會限縮在美妝保養產品相關領域，且美妝保養相關社群網站之資料具有使用者基礎分類，本研究也將利用此基礎分類進行後續研究，因此可能不適用在其他不具有類似基礎分類之資料集。
2. 情感分析時須將心得評論切割至以字詞為單位，因此本研究提出之方法僅能使用於具有斷詞工具的語言或原本就具有斷開字詞之語言。
3. 本研究假設為假設買過或發表產品心得的使用者都視為使用過該產品。
4. 因本研究欲推薦之結果為一產品類別集合下的產品組合，且是利用群內相似之使用者使用產品反推得到產品類別，因此使用者若只有使用單一產品時，在意義上代表該使用者所能得到之產品類別只有一個，和本研究想要找出之多個產品類別關係形成之產品類別集合相牴觸，且其數量會影響後續推薦結果，所以在實驗過程中，不會將只使用單一產品之使用者納入考量。另外，由於使用者可能使用多項產品，但該產品皆屬同一類別，此時在推薦產品集合階段時，也會將群內相似使用者所產出之產品類別集合排序清單中的只有單一類別的項目移除，不納入推薦之清單中。

## 研究流程

本研究的研究流程如圖 1‑4所示，詳細流程之說明如下：

1. 確立研究動機及目的：從日常生活中思考並找尋待解決的問題。隨著網際網路技術的發展及社群觀念與網路的結合，越來越多使用者在消費前會參考社群網路上其他使用者的心得評論；而推薦系統雖已被廣泛討論，但在推薦項目上仍有可以發揮的空間，特別是當推薦結果通常為依據使用者的瀏覽行為或是與當前鎖定之項目相似程度高者而生時，因此若能在推薦過程中結合社群評論觀念，產生出一個與特定主題相關且跨類別的項目列表，或許能改善推薦的效果。
2. 確立研究主題及研究範圍：美妝保養產業在網際網路技術廣泛深入至人們的日常生活中後，有了巨大的改變。越來越多的消費者透過網路或社群網路其他使用者的評論取得美妝保養的相關資訊，因此有越來越多與美妝保養相關的社群出現。本研究期望透過取得美妝保養社群網站資料，設計一推薦系統，針對平台上所能取得之使用者資料進行推薦，此推薦結果為跨類別的產品推薦，並期望能將產品形成一產品集。
3. 相關文獻蒐集與探討：根據上述研究主題蒐集相關文獻，包含：社群網站服務、自然語言處理、文件分析、情感分析、分群及推薦系統等相關技術之文獻。
4. 研究方法設計：在此階段會進行本研究之研究方法設計，並說明如何利用各項技術於本研究之系統架構中。
5. 進行實驗與驗證：根據本研究所提出之系統架構，進行系統實作並驗證其成效。
6. 結論及未來研究方向討論：提出本研究的結論及貢獻，並找出未來尚可以繼續發揮之研究方向。

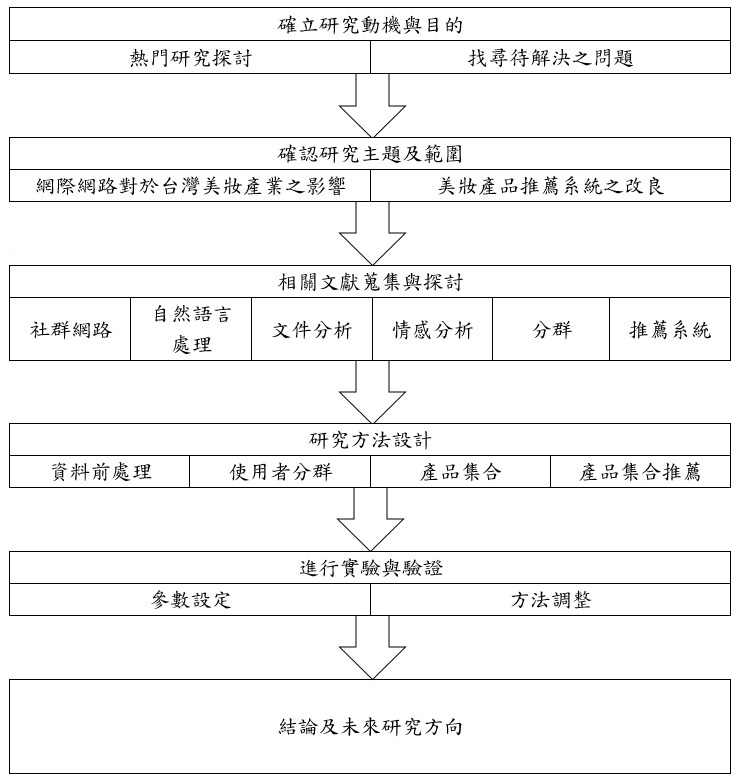


圖 1‑4 研究流程圖

## 論文大綱

為了讓讀者能夠更迅速的了解本研究，茲將本研究的章節架構條列如下，共分為五章，以下將針對各個章節進行簡要的說明：

1. 緒論：此章節主要為探討網際網路技術對於美妝保養資訊及消費者消費行為的影響，並在此章節中說明此篇論文的對於美妝保養社群以及產品的推薦之研究背景與動機，期望可以對使用者進行更完整的產品推薦。
2. 文獻探討：此章節主要是探討與本研究內容相關的技術，包括：社群網路的相關背景、自然語言處理相關技術之應用、文件分析之相關介紹、情感分析應用、分群手法簡介以及推薦系統介紹等。
3. 研究方法：在此章節中將詳細說明本研究之系統架構以及核心方法。
4. 系統建置與驗證：此章節依照本研究提出之系統架構實作系統，評估此系統的結果並進行分析以及探討。
5. 結論及未來方向：此章節會對本研究的結果進行總結，並建議未來有哪些研究方向可以深入探討。

# 文獻探討

## 社群網路服務

英文中的social network一詞最早被翻譯為社會網路。社會網路是一種由許多人或組織之間的交錯關係所建構而成的結構，在網路中的人及組織通常被稱為節點。隨著科技的發展及時代的更迭，現實生活中的社會網路也因為網際網路技術的發明與普及，形成另一種新的概念：社群網路服務(social network service, SNS)([維基百科, 2015b](#_ENREF_47))。

社群網路服務又被廣泛稱為社群網站，主要是由一群擁有相同興趣或活動的人所組成的線上社群，通常是透過網際網路作為聯繫及交流的管道。社群網路的出現提供了新的資訊交流與分享之途徑上，例如：聊天、信件往來、影音及檔案分享，或是部落格撰寫等。有些社群網站單純以建立朋友關係而成立，因而擁有超過百萬的註冊使用者，並成功的成為使用者日常生活中不可或缺的存在，知名的大型社群網路有Facebook、Twitter，及Google+等([維基百科, 2015b](#_ENREF_47))；而有些社群網站則是因為討論特定主題而吸引許多對此主題有興趣的使用者註冊，例如：以音樂為主的Last.fm及以電影為主的MovieLens。台灣同樣也有許多大大小小的社群網路，例如與美妝保養領域相關的社群網站UrCosme及FashionGuide。

而社群網路中的社群媒體(social media)指的是能讓人們進行創作、分享、交流意見、觀點以及個人經驗的虛擬社群平台。和一般傳統的大眾媒體最不一樣的地方在於，社群媒體讓使用者享有更多選擇以及編輯的空間，並且可以自行集結成社群。社群媒體可以多種不同的形式來呈現，包括：文字、音樂、圖像或是影像。知名的社群媒體包括：Facebook、Twitter及Google+([維基百科, 2015a](#_ENREF_46))。以美妝保養社群網站UrCosme為例，使用者註冊後便可以在平台上發表心得，且心得內容也能包含圖片，其它使用者可以瀏覽來自不同使用者的使用心得，也可以選擇追蹤對象。綜合上述條件，UrCosme除了是社群網站以外，同時也屬於社群媒體，能讓使用者保持一定的空間發表並選擇想要閱聽的內容。

## 自然語言處理

自然語言處理(Natural Language Processing, NLP)概念的出現是為了讓使用者以平常所使用的語言與電腦進行溝通，也就是要讓電腦能夠瞭解人類所使用的語言，因而針對文件、語句及字詞等進行處理。現今對於自然語言處理之研究相當廣泛，在此僅探討與本研究相關之技術，包含：中文斷詞處理以及詞性標記(Part-of-Speech Tagging, POS)。

### 中文斷詞處理

文件的最小組成單位為詞(word)，因此在進行文件的相關分析時，不論任何語言都必須先辨識出文件中所包含的詞才能進行後續的處理。但是不同的語系對於語言處理的方式不盡相同，例如在英文文件中會以空格來區隔每一個單詞，但在中文文件中，詞與詞之間通常不會以空格做區隔；又例如在英文的文法上，單詞有時會因時態的不同而有不一樣的寫法，因此在處理英文文件時必須對單詞進行字根還原，但中文並不需要。且在中文中，每一個字都有其獨立意義，字與字組成詞後又會產生不一樣的意思，造成中文在斷詞處理的困難。

在過去與中文斷詞的相關研究中，可以發現中文斷詞處理大致上可以分成以統計為基礎的機器學習(statistical and machine learning)方法以及以字典為基礎(dictionary-based)的方法([Wu & Tseng, 1993](#_ENREF_34))，以下將針對此兩種技術進行介紹：

* 以統計為基礎的機器學習方法：

此方法是利用大量的訓練文件，即為訓練語料(corpus)，以統計的方法找出字詞之間的關聯性及正確的斷詞位置。有許多研究應用統計方法以偵測字詞邊界，例如使用隱藏馬可夫模型(Hidden Markov Model, HMM)找出斷詞時的字詞邊界([Teahan, McNab, Wen, & Witten, 2000](#_ENREF_29))，或是利用該模型找出文章中的實體名稱([Fu & Luke, 2005](#_ENREF_10))。也有研究是利用n-gram方法將文件中的句子進行切割後，找出字詞之間的共同資訊量(mutual information, MI) ([Yang, Luk, Yung, & Yen, 2000](#_ENREF_36))，共同資訊量的數值越低，此現象之發生點越有可能為斷詞之位置。上述方法皆是以字元為文件中之最小單位，利用其出現之位置、頻率或機率等資訊進行分析。這些方法最大的問題在於需要大量的訓練語料作為參考才能得到較好之結果，因此可知訓練語料數量會影響精確度，且訓練過程得付出龐大的時間成本以建構出複雜的模型。

此方法中較知名的斷詞系統為美國史丹佛大學的自然語言處理團隊所開發之Stanford Parser([Chang, Galley, & Manning, 2008](#_ENREF_4); [Levy & Manning, 2003](#_ENREF_18))。此斷詞系統主要是利用統計結果及文法結構進行斷詞處理以及詞性標記，經由系統處理後可以得到字詞之間相依關係資訊的樹狀結構。Stanford Parser在加入經由[Xue, Chiou, and Palmer (2002)](#_ENREF_35" \o "Xue, 2002 #42)所提出之Penn Chinese Tree Bank後也可以進行中文的斷詞處理，其斷詞結果如圖 2‑1 Stanford Parser中文斷詞範例所示。



圖 2‑1 Stanford Parser中文斷詞範例

* 以字典為基礎的方法：

以字典為基礎的斷詞法是透過使用事先定義好的字典與欲處理之文件進行比對，以找出文件中出現在字典中的字詞。[Wong and Chan (1996)](#_ENREF_33)的研究指出比對的方法分為三種，分別為由前往後掃描的Forward Matching(FM)、由後往前掃描的Reverse Matching(RM)，以及雙向掃描的Bidirectional Matching(BM)。這些方法雖然較為簡單快速，但因為此方法幾乎完全仰賴字典的完整度以及規則的數量，所以當現實生活中不斷有字詞被創造出來時，若尚未被字典收錄，則在處理這些含有尚未被字典收錄的字詞之文件時，會發生字詞無法辨識的情況，此種現象稱為Out-of-vocabulary(OOV)問題。

現有的中文斷詞系統中，最廣泛使用的是由中央研究院資訊科學所詞庫小組所研發的CKIP斷詞系統([Chen & Bai, 1998](#_ENREF_5); [Chen & Liu, 1992](#_ENREF_6); [Chen & Ma, 2002](#_ENREF_7); [Ma & Chen, 2003](#_ENREF_21))。CKIP是以字典為基礎的中文斷詞系統，除了能夠將文章中的句子經由處理後變成字詞外，還能將每個字詞標註其詞性(Part-of-Speech, POS)；另外，此系統也加入能夠擷取未知詞彙的演算法以解決OOV問題。

根據[高照明 (2012)](#_ENREF_42)的研究報告指出，因為Stanford Parser在正體中文處理上是採用簡體中文的斷詞程式，儘管可以透過兩語言間的互相轉換以完成分析，但是在由繁體轉簡體、簡體轉繁體的兩次語言轉換中，容易產生轉換錯誤等問題，因此在正體中文斷詞的部分以CKIP作為基礎可以得到較好的結果。

### 詞性標記

詞性標記(Part-of-Speech)的主要目的在於對句子中的每個字詞進行詞性分析並加上詞性標籤(POS Tag)。在自然語言處理領域中，詞性標記是相當重要的環節，經由詞性標記這個步驟後，便能分析句子的文法架構，並推論出每個句子所表達的意涵。

|  |
| --- |
| 暗沉(VH)　膚色(Na)　立即(D)　美白(VH)　亮透(Na)　，(COMMACATEGORY)  肌膚(Na)　很(Dfa)　保濕(VH)　鎖水(Na)　，(COMMACATEGORY)  皮膚(Na)　摸起來(D)　觸感(Na)　很(Dfa)　水嫩(VH)　，(COMMACATEGORY)  上妝(VA)　都(D)　不易(VH)　脫妝(Na)　。(PERIODCATEGORY) |

圖 2‑2 CKIP詞性標記範例

以圖 2‑2為例，可以看到經由CKIP斷詞系統分析後，每一個字詞皆被標上詞性標籤，在此範例中可以看出四個句子的詞性規則分別為“(VH)(Na)(D)(VH)(Na)”、“(Na)(Dfa)(VH)(Na)”、“(Na)(D)(Na)(Dfa)(VH)”、“ (VA)(D)(VH)(Na)”，可以看出評論中潛在的詞性排列規則，這些規則可以視為是一種詞性的排列模式(POS Pattern)。利用這些排列模式便可以找出心得評論中的特徵與意見，例如範例中的“膚色美白”、“肌膚保濕”…等特徵及意見，詞性標記及其對應之意義如表 2‑1所示。

表 2‑1 CKIP部分詞性標記之意義

|  |  |  |
| --- | --- | --- |
| 簡化標記 | 對應詞類標記 | 詞性種類 |
| A | A | 非謂形容詞 |
| D | Dab,Dbaa,Dbab,Dbb,Dbc,Dc,Dd,Dg,Dh,Dj | 副詞 |
| Na | Naa,Nab,Nac,Nad,Naea,Naeb | 普通名詞 |
| Nb | Nba,Nbc | 專有名稱 |
| Nc | Nca,Ncb,Ncc,Nce | 地方詞 |
| VA | VA11,12,13,VA3,VA4 | 動作不及物動詞 |
| VB | VB11,12,VB2 | 動作類及物動詞 |
| VC | VC2, VC31,32,33 | 動作及物動詞 |
| VH | VH11,12,13,14,15,17,VH21 | 狀態不及物動詞 |

## 文件分析

　　從大量非結構化的資料中解析出對使用者而言重要的資訊稱為文件分析，而此技術可以應用在各種來自不同領域的資料集，也能分析並整理出不同類型的資料，讓使用者在應用時可以更加便利。目前在文字探勘領域的實作上以[Salton, Wong, and Yang (1975)](#_ENREF_28)所提出的空間向量模型最常被使用。

### 向量空間模型

向量空間模型(vector space model, VSM)主要由[Salton et al. (1975)](#_ENREF_28" \o "Salton, 1975 #22)等提出，為常用的資訊檢索模型，是透過數學概念來表達文件的一種方式，也就是利用幾何空間中的向量來詮釋文件，以方便計算文件之間的相似程度，或利於探討文件與詞彙間的各種關係([圖書館學與資訊科學大辭典, 2012](#_ENREF_45))。

當文件能以向量表示時，就可以文件向量求得其夾角，進而評估出兩向量的差異：當夾角為零度時，代表兩向量相同；若夾角為九十度時，則代表兩向量完全不相似。由於一篇文件*i*可以由文件向量=表示，其中指的是詞彙*j*在文件*i*中的權重，且通常大於或等於0，所以倘若詞彙*j*沒有出現在文件*i*中，則其權重就以數值0表示。[王京盛 (2012)](#_ENREF_39" \o "王京盛, 2012 #20)則指出向量空間模型就是將一份文件利用字詞集合將其轉移成多維向量空間中的一個資料點。舉例來說，若有一文件集合，而第*j*份文件就可以被表示成=，其代表的意義為字詞集合中的*k*個字詞在文件中的重要性。

最常見的空間向量模型計算方式則是使用Term Frequency-Inverse Document Frequency(TF-IDF)，其公式如下：

|  |  |
| --- | --- |
| = | (2-1) |
| = | (2-2) |
| = | (2-3) |

其中是文件*i*中字詞*j*的出現頻率(term frequency, TF)，是字詞*j*在文件*i*中出現的次數，代表文件*i*中的字詞總數；則是字詞*j*在文件集中的逆向文件頻率(inverse document frequency, IDF)，代表文件集中的文件總數，代表文件集中包含字詞的文件個數；代表字詞*j*在文件*i*中的重要性。

TF用來判斷該字詞在該文件中的出現頻率，IDF則是考慮到一字詞若同時頻繁出現在其他文件中，表示需要減低此字詞對該文件的重要性，故利用IDF來降低其在此情況下的重要性，反之若一字詞只在該文件中較頻繁出現，則表示該字詞為該文件的重要特徵，則其TF-IDF值越高。

### 使用者相似度計算

在推薦系統中，存在許多種相似度計算方式，其中最常被使用的兩種相似度計算方式為：以相關程度為基礎的計算方式(correlation-based)以及以餘弦為基礎的計算方式(cosine-based)([Isinkaye et al., 2016](#_ENREF_15))。此兩種方式中又各以Pearson相關係數(Pearson correlation coefficient)以及餘弦相似度(cosine similarity)最常被為使用([Anand & Bharadwaj, 2010](#_ENREF_1))。

Pearson相關係數為用來衡量兩變數之間的線性相依程度，在推薦系統中是利用兩使用者對於項目的評分做為向量並以此計算兩向量間的相關程度，為一統計方法；而cosine similarity則與Pearson相關係數不同，它屬於空間向量模型，在衡量兩使用者間的相關性時比較偏向線性代數技巧，而非統計技術，為利用兩使用之評分向量間的夾角來判斷兩使用者是否相似。在傳統文件探勘技術中，餘弦相似度是最被廣泛使用作為衡量文件間距離的方式([Mahdavi, Chehreghani, Abolhassani, & Forsati, 2008](#_ENREF_23))。兩份文件彼此之間的相似度可以透過餘弦相似度計算公式而得，其公式如下：

|  |  |
| --- | --- |
|  | (2-4) |

其中為ㄧ篇文章之向量，為比較文章之向量，*t*為文件所包含的字詞數量，則表示第*i*個字詞在第*j*篇文章中的權重。在本研究中將以使用者是否用過某產品的關係形成「使用者－產品矩陣」，在此矩陣中將會記錄使用者與產品之間的關係，使用者若用過該產品，則其值為1，反之則為0，若單獨只採用值為1欄位作為向量維度，則所產生之使用者向量維度兩兩並不完全相等。相較於其他相似度計算方式，餘弦相似度計算更適用於兩向量維度不相等的相似度計算情況，因此在相似度計算的公式選擇上([Mahdavi et al., 2008](#_ENREF_23))，本研究將選用最常被用來衡量兩向量間夾角的餘弦相似度進行運算。

## 情感分析

情感分析又稱作意見探勘(opinion mining)、評論探勘(review mining)、評價擷取(appraisal extraction)或態度分析(attitude analysis)，任務為偵測、擷取或分類文字所傳達之對於特定主題的態度([Ravi & Ravi, 2015](#_ENREF_27))，即為決定文字態度的極性，例如：正向、負向或中性。情感分析的技術可以依照不同層級分成許多類別，主要可分為兩類：文件層級及特徵層級。文件層級是以整篇文章為對象決定意見極性，而特徵層級則是針對每個特徵來決定意見極性。[Quan and Ren (2014)](#_ENREF_26)研究指出文件層級與特徵層級相較之下，特徵層級較能夠針對特定目標分析，且有更廣的應用空間。

情感分析的方法根據[Cambria, Schuller, Yunqing, and Havasi (2013)](#_ENREF_3)所歸納之結果可以分為四類：Keyword Spotting、Lexical Affinity、Statistical methods以及Concept-based approaches。Keyword Spotting方法是利用明確的意見字詞(如“Happy”、“sad”、“afraid”、“bored”)人工建立情緒字典，此方法的缺點在於無法辨識否定的意見字詞且只能從字面意義判斷字詞極性。Lexical Affinity為透過訓練語料得到字詞帶有情緒的機率值，因此此方法不僅能夠偵測出明顯的意見字詞，也能夠對任意字詞賦予包含情感的可能性；但此方法的缺點在於當意見字詞具有否定以及其他涵義時，此方法無法辨識句中的字詞為意見字詞或是其他涵義，例如：“I avoid an accident”和“I met my friend by accident”兩個句子中的accident，前者為否定意見字詞，而後者則代表突然、不經意的意思。且Lexical Affinity對字詞賦予情感機率時，會因為不同的資料集來源而使得結果產生偏頗。Statistical methods則是利用監督式機器學習方式對文章進行情感分類，因此需要事先針對資料集中的意見字詞進行人工標記。由於是利用統計的方式得到結果，所以只有在資料集數量夠大時才會有較佳的結果，且此類方法僅可處理情感分析層級範圍較廣的文件層級，無法在較小單位的文字(如：句子)上進行分析。Concept-based approaches為利用知識本體架構(ontology)或是語義網(semantic networks) 分析意見字詞。由於此方法仰賴大量的語義知識庫，因此可以透過自然語言方式進行語義分析，而無須盲目地使用關鍵字或字詞共同出現頻率等方式分析意見字詞。此方法較能找出字詞中較為精細的情感，效果優於文法分析方式。但此法因依賴知識庫的完整性，倘若知識庫不夠廣泛，在分析與辨別自然語言字詞字義時就會有困難。

因Statistical methods需要事先利用人工標記方式對訓練資料集進行意見字詞標計處理，但本研究希望透過一非監督式方式得到結果，故不採用此方法。因其他方法皆為透過事先建立的情緒字典及知識庫進行情感分析，且根據前面所述可知使用知識本體架構能夠偵測到字詞更精細的情感，因此本研究將使用情緒字典搭配知識本體架構進行情感分析。目前在繁體中文部份所能取得之資源僅限於台灣大學情緒字典NTUSD；在簡體中文方面則有HowNet能夠提供簡體字轉繁體字的情感字詞分析功能。知識本體架構的部份則有中央研究院所開發之廣義知網知識本體架構。

以下將針對廣義知網進行概況說明。廣義知網知識本體架構(E-HowNet)是中央研究院資訊所中文詞知識庫小組擴建繁體字知網(HowNet)而得，為一詞彙意義實體關係模型。其範例架構如圖所示，由圖 2‑3可知廣義知網知識本體架構是以階層式的架構表達語意及詞彙。且從圖中也可以看出廣義知網知識本體架構提供部份字詞的極性分數，例如圖中“細緻”的極性分數為0.2972，表示該字詞較可能的極性為正向，且其正向程度為0.2972。因此，使用廣義知網知識本體架構可協助本研究透過同意字詞概念為無極性分數之字詞取得其極性分數。

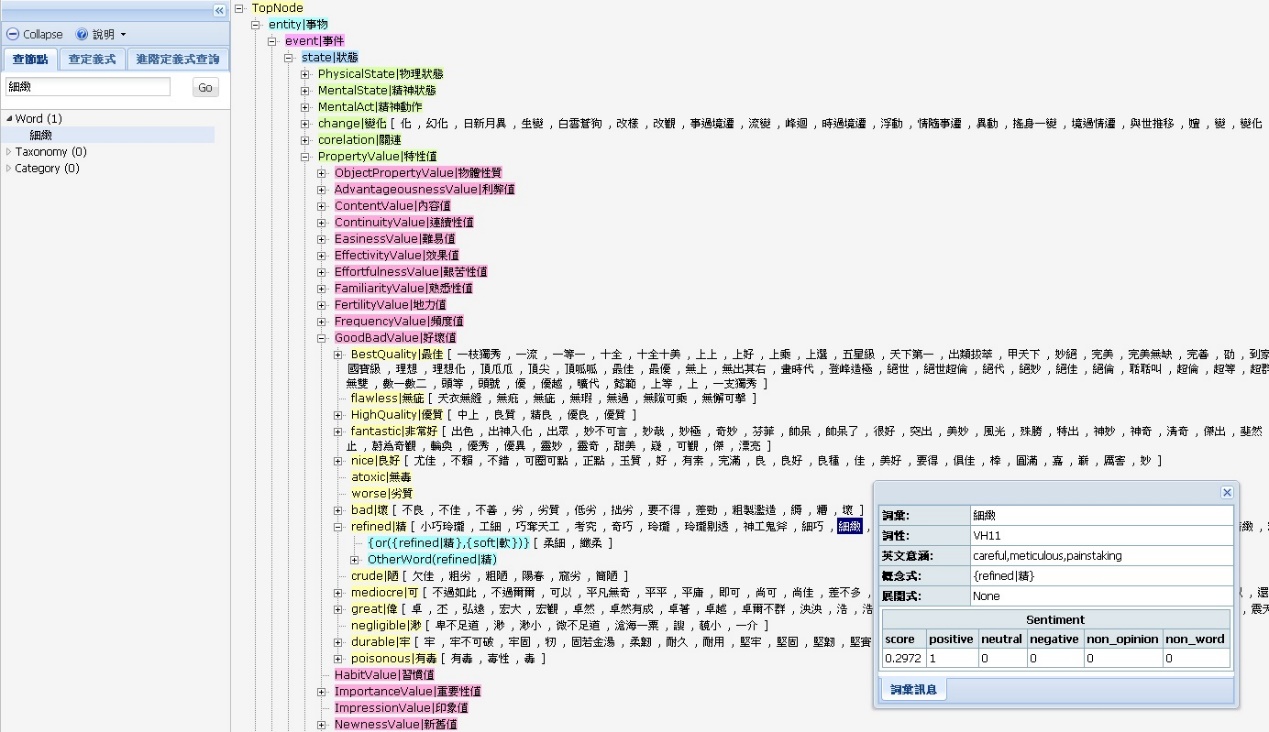


圖 2‑3 廣義知網範例圖示

## 分群

分群(clustering)技術主要是為了因應在龐大的資料集合中，找出集合內各個物件的分布情況及其隱含意義需求而生的技術。分群的目的是將大型資料集合切割成數個較小且特定的群集，使得群集與群集之間的相似程度最小，群集內資料的相似程度最大。資料集在進行分群前，並沒有已知的群集數量及每個群集的類別標籤(class label)，因此分群技術是一種非監督式分類方法([Zahra et al., 2015](#_ENREF_37))；分群技術也被廣泛的應用在不同的技術上，例如：機器學習、影像分析，或是離群值偵測等。

[Halkidi, Batistakis, and Vazirgiannis (2001)](#_ENREF_12)將分群的技術簡單劃分為四種類型：分割式分群(partitional clustering)、階層式分群(hierarchical clustering)、密度式分群(density-based clustering)及網格式分群(grid-based clustering)。分割式分群試圖將資料集直接分解為數個互相沒有交集的群集，此方法需要事先定義欲分割的群集數目以使分群結果最佳化。階層式分群則是透過合併或分裂的方式將資料集進行分割，此方法的方群結果為一樹狀圖，由此圖可以看出群與群之間的關聯性。密度式分群的主要概念為透過密度的調控以決定是否將互為鄰居的資料點劃分為同一群集。網格式分群的特點則是在於其將空間切割成若干個有限數量的子空間，再利用這些空間進行分群作業。

針對上述四種分群法的描述，[林宜瑩 (2010)](#_ENREF_41" \o "林宜瑩, 2010 #19)將各方法的優缺點整理如表 2‑2所示，可以由表得知分割式分群法適用於擁有大量文件的資料集，且其運算成本低，較能快速回應使用者，因此雖其分群結果綜觀而言雖非最佳，但分群品質仍保有一定準確度，故本研究將採用分割式分群法作為分群方式。

表 2‑2 分群方法比較表([林宜瑩, 2010](#_ENREF_41))

|  |  |  |
| --- | --- | --- |
|  | 優點 | 缺點 |
| 分割式分群 | 運算成本低且適用於大量文件分群 | 分群結果易受初始群心以及離群值影響 |
| 階層式分群 | 分群結果較其他方法為佳 | 分群結果易受到門檻值影響且運算時間長、成本高 |
| 密度式分群 | 對於雜訊值的容忍度較高 | 不適用於大型資料集合 |
| 網格式分群 | 運算效率較好 | 分群品質易受到邊界影響且複雜度高 |

分割式分群做法為先指定欲分群的群數，藉著反覆疊代運算，降低誤差目標函數的值，直到目標函數不再變化，即達到分群的最後結果；其中K-means分群法即是常見的分割式分群法，也是推薦系統中常用之分群法([Isinkaye et al., 2016](#_ENREF_15))。以下將針對K-means分群法進行說明。其作法是先選定k筆資料作為起始群心，再將每一筆資料劃分到與其距離最近的群心之群集，經過反覆運算後，將資料分割為k個群集，詳細步驟如下([Mahdavi et al., 2008](#_ENREF_23))：

1. 從資料集中隨機挑選出k筆資料作為起始群心。
2. 計算每一筆資料與剛剛所挑出來之k個起始群心的距離，找出距離最近的群心，並將此筆資料劃分到該群心所屬的群集中。
3. 全部的資料點都分群完畢後，再使用每一群中資料點重新計算各群集的新群心。
4. 找到新群心後，重複步驟2.和3.，直到符合終止條件為止。

儘管K-means演算法為最常被使用的分群手法，但其分群結果易受到起始群心以及分群數目的影響，也容易受到資料集內雜訊與離群值影響而導致分群品質不佳([Karol & Mangat, 2013](#_ENREF_16))。過去也有學者在進行分群研究時採用K-Medoids分群法([林宜瑩, 2010](#_ENREF_41))，其概念與K-means十分相似，但在群心選擇上，K-Medoids採用群集中最接近中心的資料點作為群心，與K-means採用群集中各資料的平均值作為中心點不同，但分群方式差異不大。

總結上述，分割式分群法的成本較其他方法低，且分割式分群法為目前最廣泛使用，也被視為最適用於大量文件分群的方法，又[Madhulatha (2011)](#_ENREF_22" \o "Madhulatha, 2011 #60)將K-means與K-Medoids方法進行比較後得到結果發現：K-means的優點在於其運算時間較短，但使用者必須事先定義群數，且在大資料集上的表現不如K-Medoids好，故本研究將採用K-Medoids作為分群法。

## 推薦系統

網路的蓬勃發展使得資訊量快速成長。對使用者而言，儘管資訊量大、豐富度高，但在搜尋資料時卻不容易找到符合需求的資料，且雖然搜尋引擎具有過濾資訊的功能，但往往搜尋結果仍是相當可觀的資料筆數，使用者依然需要自行過濾或篩選出需要的資訊，因此推薦系統應運而生。推薦系統為利用使用者過去的使用習慣或喜好來主動發掘使用者可能會有興趣的主題或資源並推薦給使用者，如此一來便可以減少使用者自行搜尋並過濾資料的時間成本([Chiang & Huang, 2015](#_ENREF_8); [Liang, Yang, Chen, & Ku, 2008](#_ENREF_19); [Zahra et al., 2015](#_ENREF_37))。

[Liang et al. (2008)](#_ENREF_19)及[Chiang and Huang (2015)](#_ENREF_8" \o "Chiang, 2015 #4)的研究皆指出推薦系統所使用的資料過濾方法主要分為四種，分別為：基於規則的過濾技術(rule-based filtering)、基於內容的過濾技術(content-based filtering)、協同過濾技術(collaborative filtering)以及混合式過濾技術(hybrid filtering)，分類如圖 2‑4所示。以下將簡單介紹這四種方法：

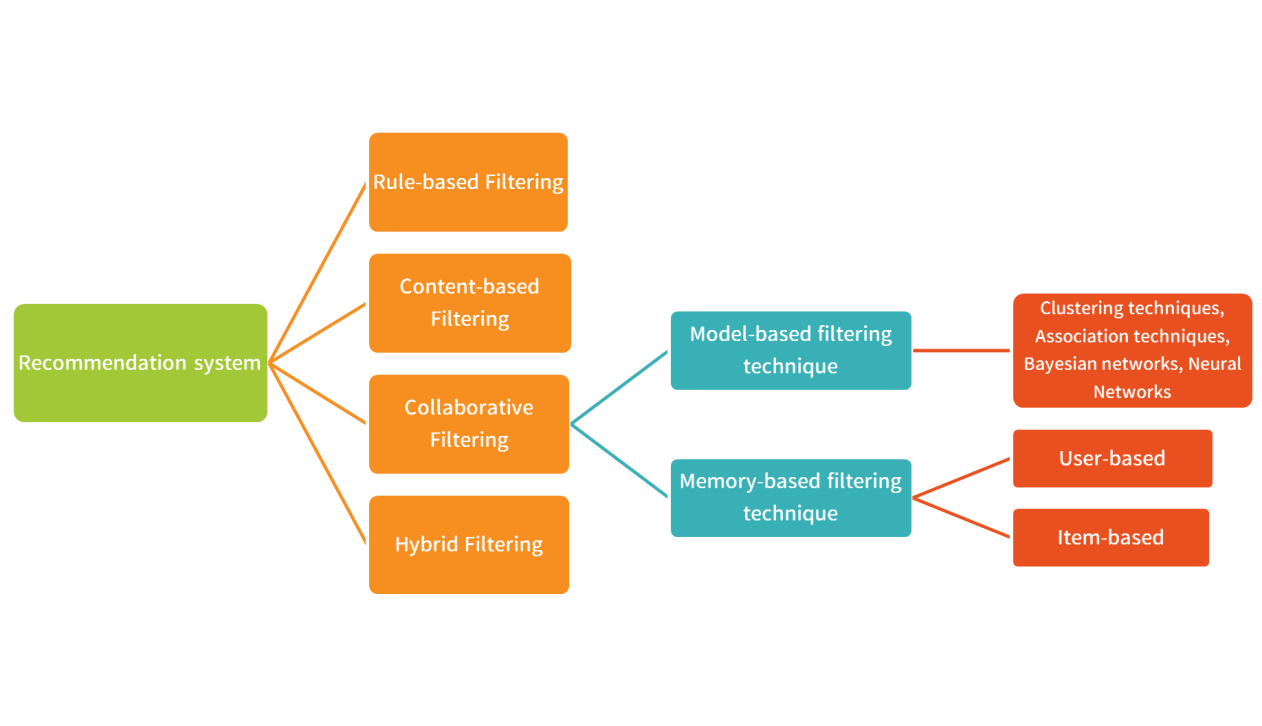


圖 2‑4 推薦系統的種類([Isinkaye et al., 2016](#_ENREF_15))

* 基於規則的過濾技術(rule-based filtering)：基於規則的過濾技術主要是利用事先定義好的規則(if-then)以選出相關的資訊進行推薦，如此一來便可以過濾掉不相關的資訊。
* 基於內容的過濾技術(content-based filtering)：基於內容的過濾技術則是根據使用者過去的習慣或偏好，建立出使用者的資料輪廓，再透過資料輪廓找出相似的項目以進行推薦([Isinkaye et al., 2016](#_ENREF_15))。
* 協同過濾技術(collaborative filtering)：此方法的概念是先分析使用者的興趣或偏好，然後再從其他使用者中找到擁有類似興趣或偏好的群體，利用這些群體對某一新增加項目的偏好對目前尚未表現出興趣或偏好的使用者進行預測([Barragáns-Martínez et al., 2010](#_ENREF_2))。
* 混合式過濾技術(hybrid filtering)：此方法是結合基於內容的過濾技術以及基於使用者的協同過濾技術的特性而成。基於內容的過濾法在考量使用使用者的偏好及興趣時，會只針對項目內容中最重要的特徵值，而造成推薦上的限制；而基於使用者的協同過濾技術則是會忽略項目的內容，只針對相似使用者的偏好來進行推薦，因此會產生推薦正確性降低的風險。由於基於內容的過濾技術以及協同過濾技術各有優缺點，所以當單一的過濾技術無法解決特定問題時，便可以考慮使用混合式的過濾技術([Barragáns-Martínez et al., 2010](#_ENREF_2); [吳虹瑩, 2012](#_ENREF_40))。

其中，協同過濾技術廣泛應用在推薦系統中，其概念為假設擁有相似品味的使用者會對產品或是物品擁有相似的偏好([Kumar & Fan, 2015](#_ENREF_17))；也就是說，當目標使用者在進行選擇時，會對與其相似之使用者所提出的建議感興趣。協同過濾技術與基於內容的過濾技術不同的地方在於協同過濾技術旨在先找到與目標使用者相似的使用者，再依據其偏好來推薦給目標使用者。由此方法所產生的推薦結果會因為使用者之間的相關性而產生不同的結果。其主要目的是尋找與目標使用者之相似使用者對於項目的評價，以推敲目標使用者對於項目可能會產生的評價([Mishra, Kumar, & Bhasker, 2015](#_ENREF_24))。

而根據[Kumar and Fan (2015)](#_ENREF_17" \o "Kumar, 2015 #24)的研究，協同過濾演算法主要又可以分成兩種：以使用者為基礎的協同過濾演算法(user-based collaborative filtering algorithm, user-based CF)以及以項目為基礎的協同過濾演算法(item-based collaborative filtering algorithm, item-based CF)。

* 以使用者為基礎的協同過濾演算法(user-based CF)：此演算法是找尋與目標使用者相似，對項目擁有相似偏好(或可解釋為對項目擁有相似評分)的使用者，並將這些使用者高度評分的項目推薦給目標使用者。
* 以項目為基礎的協同過濾演算法(item-based CF)：此演算法則是考量項目間的相似程度，並會將其他與目標項目相似且被使用者高度評分的項目推薦給目標使用者。

但因為單純使用「以使用者為基礎的協同過濾演算法」會遇到一些問題；而單純使用「以項目為基礎的協同過濾演算法」則其結果可能與使用者需求較不貼近，於是[Kumar and Fan (2015)](#_ENREF_17)提出了ㄧ種結合兩種協同過濾演算法的混合式方法，能達到更佳的個人化推薦結果，此方法為混合式基於使用者—項目之協同過濾方法(hybrid user-item baesd CF method)。

## 小結

本章節針對與本研究相關之過去與現有技術的理論發展狀況進行簡介與整理，包含：社群網路服務、自然語言處理、分群、推薦系統，及文件分析等。在自然語言處理的部分，因為Stanford Parser在正體中文處理上是採用簡體中文的斷詞程式，儘管可透過兩語言間的互相轉換以完成分析，但容易產生轉換錯誤等問題，因此在進行中文斷詞處理時，本研究將利用CKIP中文斷詞系統對產品心得評論進行斷詞處理及詞性標記。情感分析的部分，因為本研究將計算使用者對於產品的情緒分數，故將使用包含字詞極性及分數的廣義知網本體架構來處理字詞與極性分數間的比對與轉換。

在使用者相似度計算部分，因本研究將以使用者與產品間的使用關係作為相似度計算的依據，而非以使用者對於產品之評分為依據，因此較適用以計算兩向量夾角的餘弦相似度計算方法進行計算，故本研究將以此方法進行相似度計算。而K-means演算法雖為最常被使用的分群手法，但其分群結果易受到起始群心、分群數目及資料集內雜訊與離群值影響而導致分群品質不佳。又[Madhulatha (2011)](#_ENREF_22)將K-means與K-Medoids方法進行比較後得到結果發現：K-means的優點在於其運算時間較短，但使用者必須事先定義群數，且在大資料集上的表現不如K-Medoids好，故本研究將採用K-Medoids作為分群法。另外，在推薦方法上，本研究會利用使用者與產品間的使用關係進行協同過濾，並分析心得評論得到使用者對於產品的情緒分數，以此作為產品排序的權重，並在計算之後得到產品的推薦清單，綜合以上手法，本研究所採用的推薦系統為一混合式推薦系統。各方法的優缺點如表所示，而詳細的研究架構與研究方法將於第三章進行深入的說明。

表格 2‑1本研究各方法選擇之優缺點比較表

|  |  |  |  |
| --- | --- | --- | --- |
| 各方法比較 | | 優點 | 缺點 |
| 中文斷詞處理 | 中研院CKIP 斷詞系統 | 以事先建立好的字典進行比對，較簡單快速 | Out-of-vocabulary(OOV)問題 |
| 相似度 計算方式 | 餘弦相似度計算 | 利用計算兩向量夾角，求得兩文件相似程度，不易受維度影響，且較適合本研究資料 | 會忽略兩文件的本身的長度或是文件間的距離 |
| 情感分析 | NTUSD  搭配  廣義知網 | 能夠偵測到字詞更精細的情感 | 倘若知識庫不夠廣泛，在分析與辨別自然語言字詞字義時就會有困難 |
| 分群 演算法 | K-medoids 分群法 | 在大資料集上的表現較K-means好 | 須事先定義群數 |
| 推薦系統 | 混合式推薦系統 | 結合多種過濾方法或外部資源，避免使用單一種過濾技術所產生的問題，以期達到更佳的個人化推薦結果 |  |

# 研究方法

## 研究架構

本研究之研究目的為採用美妝保養社群網站所提供的使用者膚質狀況、產品資訊，以及曾經對於產品所發表過的產品使用心得評論資料等，經過分析後，找出使用者可能感興趣的產品並將其集結成產品集後，對目標使用者進行不同種類產品所形成的產品集合推薦。在一些美妝保養社群網站上，產品除了依照品牌進行分類外，也針對產品不同的功能進行階層式分類，例如：臉部保養類、化妝產品類、美體健康類等。每一個大類別又被分類成許多子類別，以臉部保養類產品為例，若使用者以基礎保養作為需求，則可以看到臉部保養被分為臉部卸妝、化妝水、乳霜等八個細項([FashionGuide華人第一女性時尚美妝傳媒, 2015](#_ENREF_9); [UrCosme, 2015](#_ENREF_30); [Wikipedia, 2015](#_ENREF_32))，如下圖 3‑1所示。

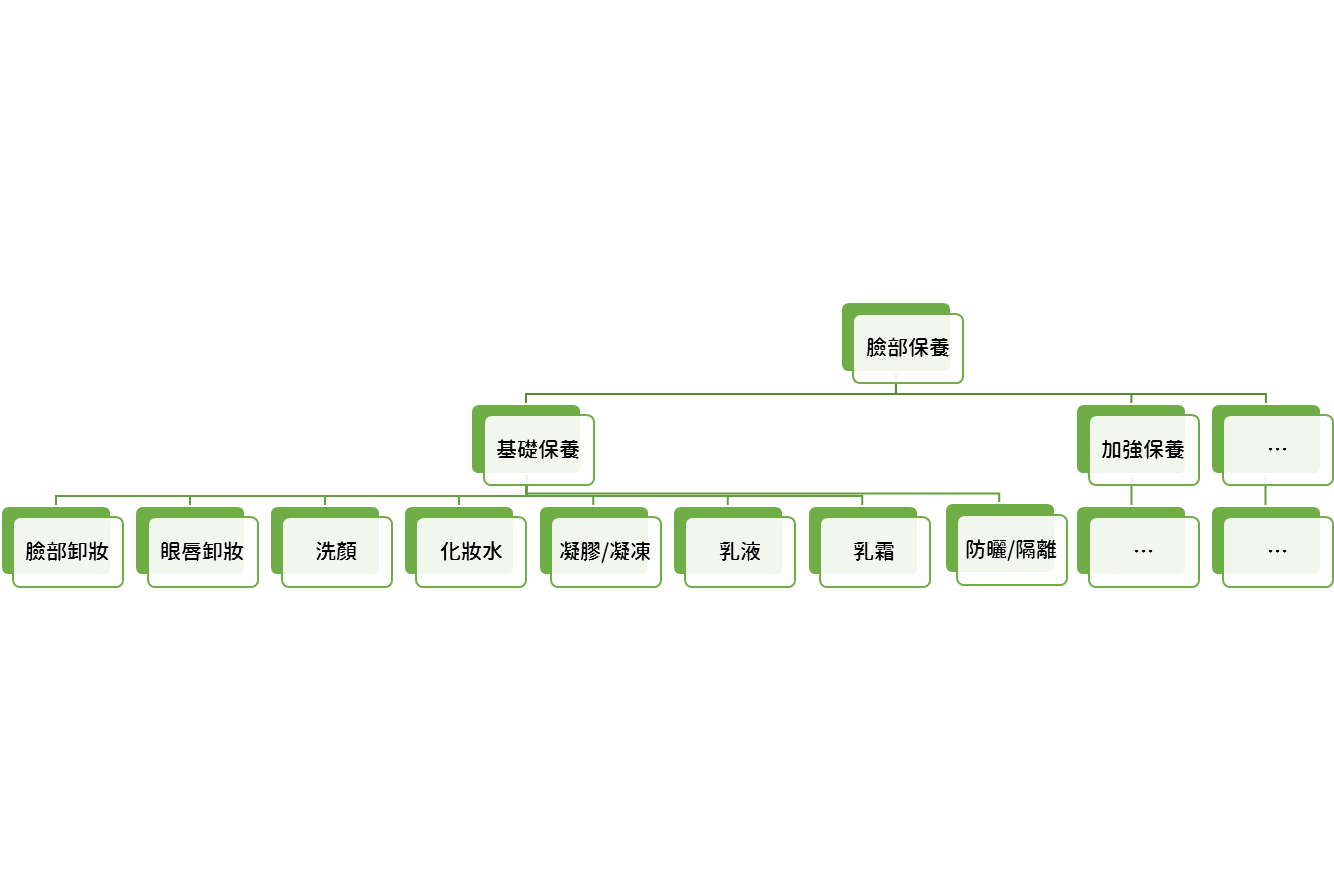


圖 3‑1 產品類別示意圖

有些美妝保養社群網站直接將產品的類別進行詳細的劃分，並直接將產品種類以清單方式呈現，使用者在搜尋產品相關評論時可以直接透過該產品所屬的類別直接搜尋，比起前面所敘述的階層式分類更為直覺。但無論是哪一種呈現方式，其對於產品種類的分類都相當明確，所以本研究在蒐集資料時會同時將產品所屬種類記錄下來，並將產品類別的表達方式進行統整，以利後續實驗實作。

當使用者在網站上註冊成為會員時，網站上都會紀錄其基本資料，且在本研究中假設不論是曾經購買或對產品發表心得者，都視為使用者曾經使用過該產品。本研究將會以上述假設為依據，並同時考量使用者的膚質狀況及使用者曾經使用過之產品，計算使用者與使用者彼此之間的相似程度並將使用者進行分群，再將群內使用者所使用過之不同類別之產品集合起來推薦給目標使用者；因為在網站上，使用者使用過的產品會分別散落在不同的類別中，所以若是將與目標使用者相似之使用者所使用過之產品進行整理歸納，便可以得到一個與美妝保養相關且跨產品種類的產品推薦集合。

使用者膚質分類可以依照肌膚含水量、含油量以及敏感程度劃分為五類：中性肌膚(normal skin type)、油性肌膚(oily skin type)、乾性肌膚(dry skin type)、混合性肌膚(Combination skin type)及敏感性肌膚(sensitive skin type)([FashionGuide華人第一女性時尚美妝傳媒, 2015](#_ENREF_9); [UrCosme, 2015](#_ENREF_30); [WebMD, 2015](#_ENREF_31))。美妝保養產品使用者在選擇其需求時會考慮其膚質狀況進行產品選擇，因此本研究將以此為依據先對資料集內的使用者進行膚質狀況分類，分類表如圖 3‑4所示，以利後續實驗。

大部分的產品推薦通常針對使用者的瀏覽紀錄或是使用者對於項目的評分來推薦相似的物品，本研究則希望可以在進行推薦時，將使用者的相似程度與其對產品所發表的評論一起納入考量，並將推薦之產品形成產品集，以對使用者進行跨種類的產品推薦，而非只是針對相似產品進行推薦。本研究將使用自行撰寫的程式進行資料收集，所需要的資料包含：使用者資訊(含其編號及膚質狀況)和產品資訊(含產品類別、產品名稱、心得及篇數)。資料收集完畢後會進行統計整理，最後會產生符合本研究需求的資料集，資料集結構圖示如下圖 3‑2所示。



圖 3‑2 資料集結構圖示

本文研究架構分為資料收集與前處理模組、使用者與產品分群模組、產品集合模組及產品集推薦模組，並欲透過這些模組找出相似的同一群使用者所建構之包含不同類型產品的產品集，以作為其他與此群使用者相似之目標使用者進行購買決策之參考依據。圖 3‑3為本研究之研究架構圖，以下將針對各模組的功能與目的進行簡單的說明：

1. 資料前處理(data preprocessing)模組：

在此模組中將會進行資料收集以及資料前處理。

* 資料蒐集：本研究利用自行撰寫之程式蒐集美妝保養社群網路平台的使用者資訊，包含：使用者編號、使用者膚質狀況、使用者產品心得，及產品資訊(包含產品名稱及產品心得篇數)等。取得資料後會將資料儲存在資料庫中並進行簡單的整理，例如將產品種類進行統合，再依據統合後的產品種類將產品所屬種類重新記錄。
* 資料前處理：資料蒐集完畢後，為了方便後續實驗計算，所以在此模組中會先針對資料集進行前置處理。首先，資料集中的使用者資訊部分，因為本研究計畫在資料集中找出與目標使用者相似之使用者群集，因此需要對使用者進行分群處理。此部分實驗需要以使用者之特徵進行運算，因此將會對資料集中之使用者資訊及其使用產品之狀況進行特徵擷取，以利於計算使用者間之相似度。

1. 使用者分群(Clustering)模組：

在此模組中將會對資料進行相似度計算以及分群處理。

* 相似度計算：資料經過前處理模組之後預計會得到使用者與使用過之產品所構成的「使用者—使用者使用產品」矩陣，得到此矩陣後將會進行相似度計算，找出資料集中使用者與使用者間的相似程度。在相似度計算部分目前計畫採用餘弦相似度作為計算方法。
* 分群：相似度計算完成後便能對使用者進行分群。對於使用者分群之目的是為了將資料集中的使用者劃分成不同群集，如此才能對目標使用者找出與其最為相似之使用者群集；而完成使用者分群後，便可以蒐集每一群使用者所使用之產品，再針對這些相似使用者所使用過之產品進行產品類別判定，並計算該群使用者所使用之產品所在之類別及個數。

1. 產品集合(Set Grouping)模組：

當分群程序完成後，此模組將會再針對每一群使用者所使用之產品進行產品集合的處理，其概念是針對分群結果中與目標使用者相似之群集的每一相似使用者進行檢視，將使用者所使用過之產品所屬的產品類別進行對應，若經過對應步驟發現使用者使用之產品落在不同產品類別中，則使此兩產品類別有一連結，依此類推直到相似使用者依序檢視完畢。此時再針對產品類別與類別之間之連結進行數量計算，重複出現最多次的連結組合則視為此群使用者之最適產品種類集合。

1. 產品集合推薦(Recommendation)模組：

得到產品集合之定義後，再對產品群中之產品進行排序；基於本研究之假設，對於某產品曾經發表心得評論或是購買此產品之使用者，皆視為曾經使用該產品。在進行產品推薦時，需要對該類別之產品進行排序，但若以該產品被使用次數作為排序標準，有可能會發生使用者使用過該產品但對該產品之喜好程度不如預期，所以排序之依據無法單單使用產品被使用次數，而是需加入經由情感分析進行心得評論計算後所得之使用者對於產品的情感分數，以產生較為較貼近可能使用結果之排序。

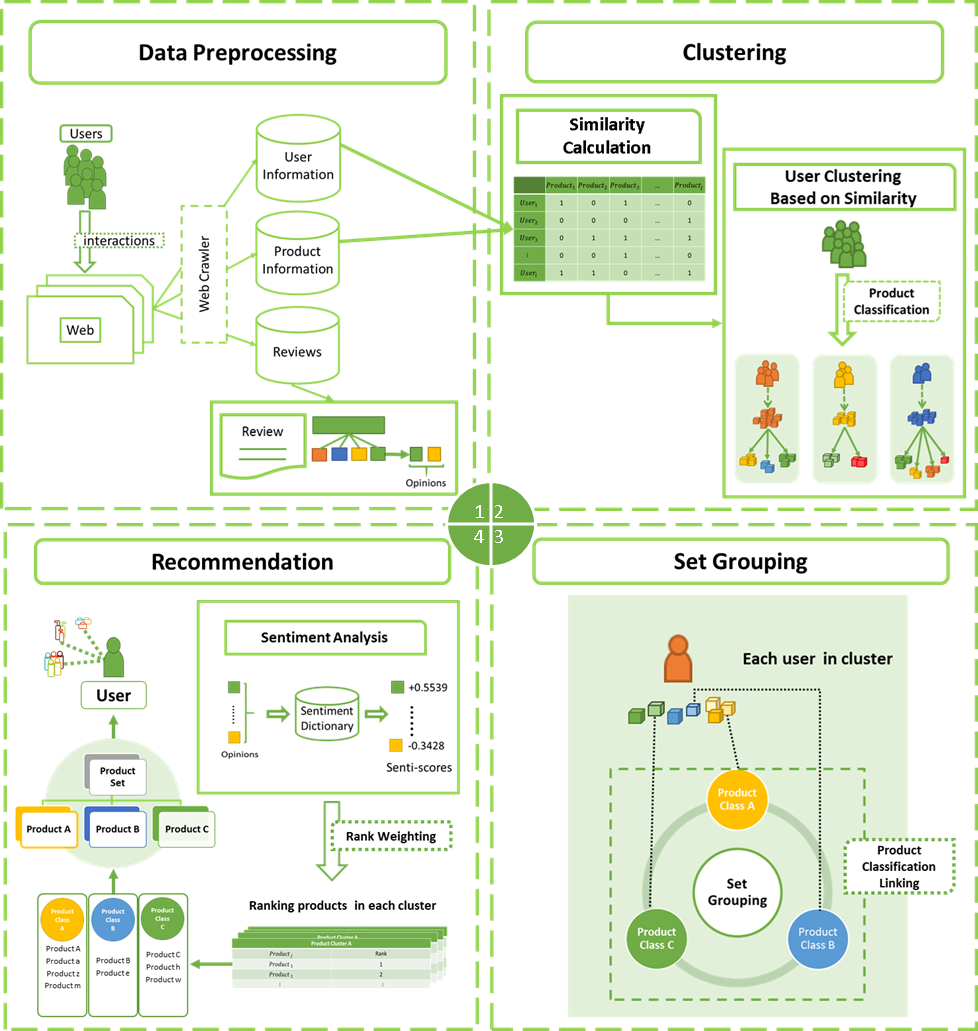


圖 3‑3 研究架構圖

## 資料前處理

本研究利用自行撰寫之程式蒐集美妝保養社群網路平台的使用者資訊，包含：使用者編號、使用者膚質狀況、使用者產品心得，及產品資訊(包含產品名稱及產品心得篇數)等。

根據第一章所述，使用者在購買或使用產品時會依賴社群網路上的評論或心得以進行購買的決策；又因社群網路已與日常生活密不可分，使得越來越多的使用者在使用某產品後，會選擇將自己的評論或心得在網路上進行發表。因此在產品評論網站中，使用者與產品之間的關係可以透過使用者是否曾經購買、是否曾經使用或是否撰寫過對於該產品的評論或心得來判定。在本研究中假設凡是使用者曾經對某產品發表評論或心得或曾經對產品進行評分，則將該使用者視為曾經使用過該產品。

在此模組中，資料經過整理後會得到膚質集合，*i*為第幾種膚質，膚質分類示意圖如圖 3‑4所示。

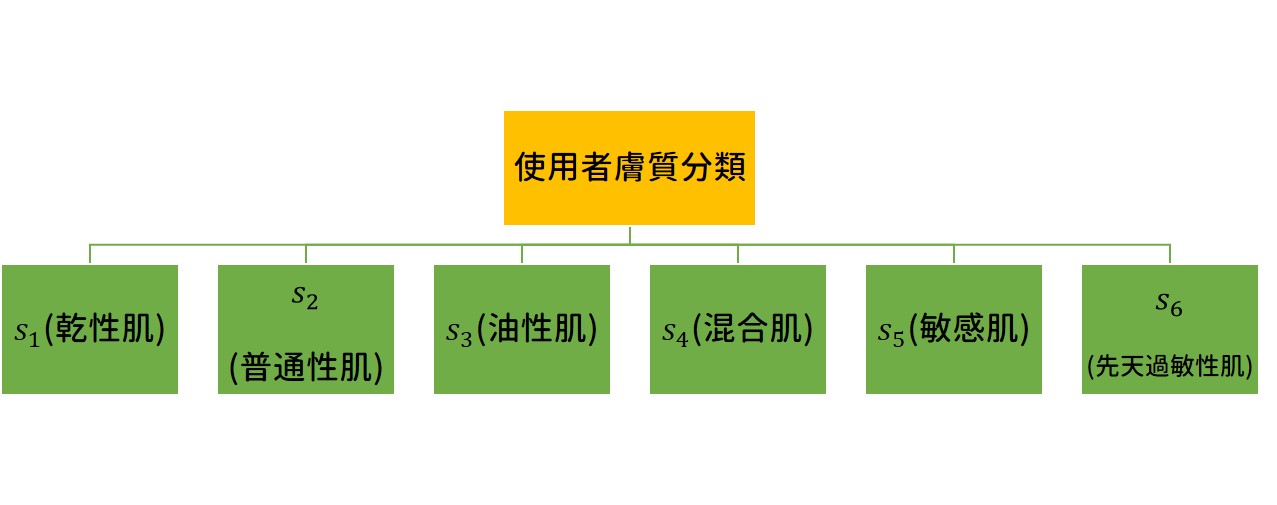


圖 3‑4 使用者膚質分類示意圖

產品集合；產品，其中代表第*j*個產品名稱，則為其產品類別；使用者的評論集合，其中代表使用者對產品*j*的心得評論集合；且定義經過中文斷詞處理後的評論集合為；每一個使用者的資料為，其中為使用者的膚質；每一使用者使用過的產品集合；使用者範例說明如圖 3‑5所示，評論經過中文斷詞處理後的範例如圖 3‑6所示。

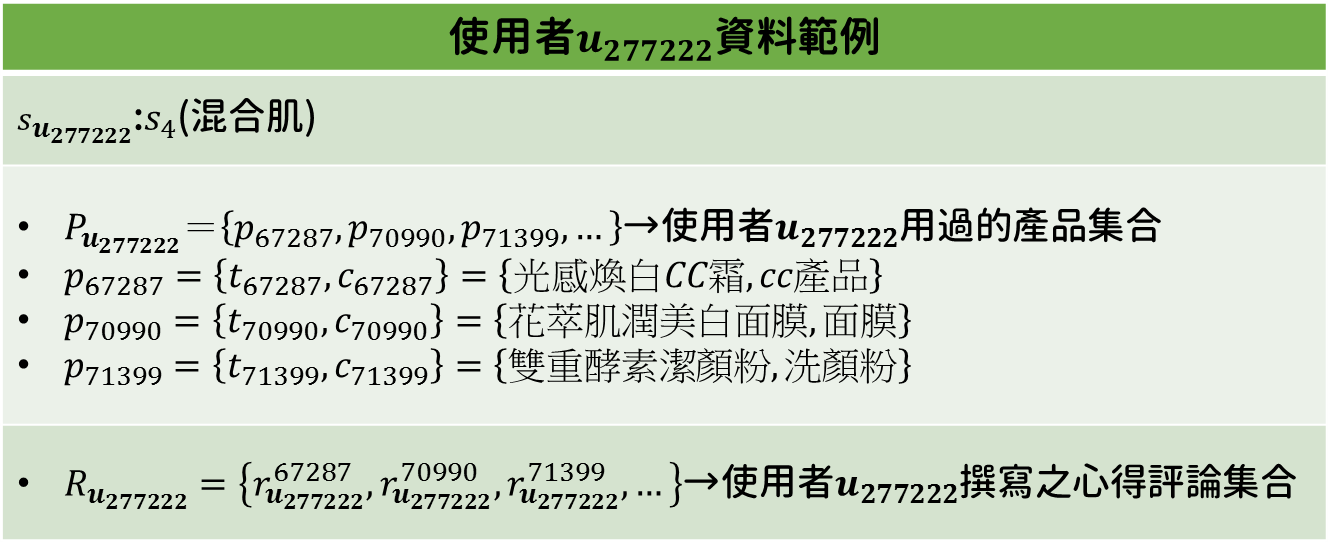


圖 3‑5 使用者資料範例

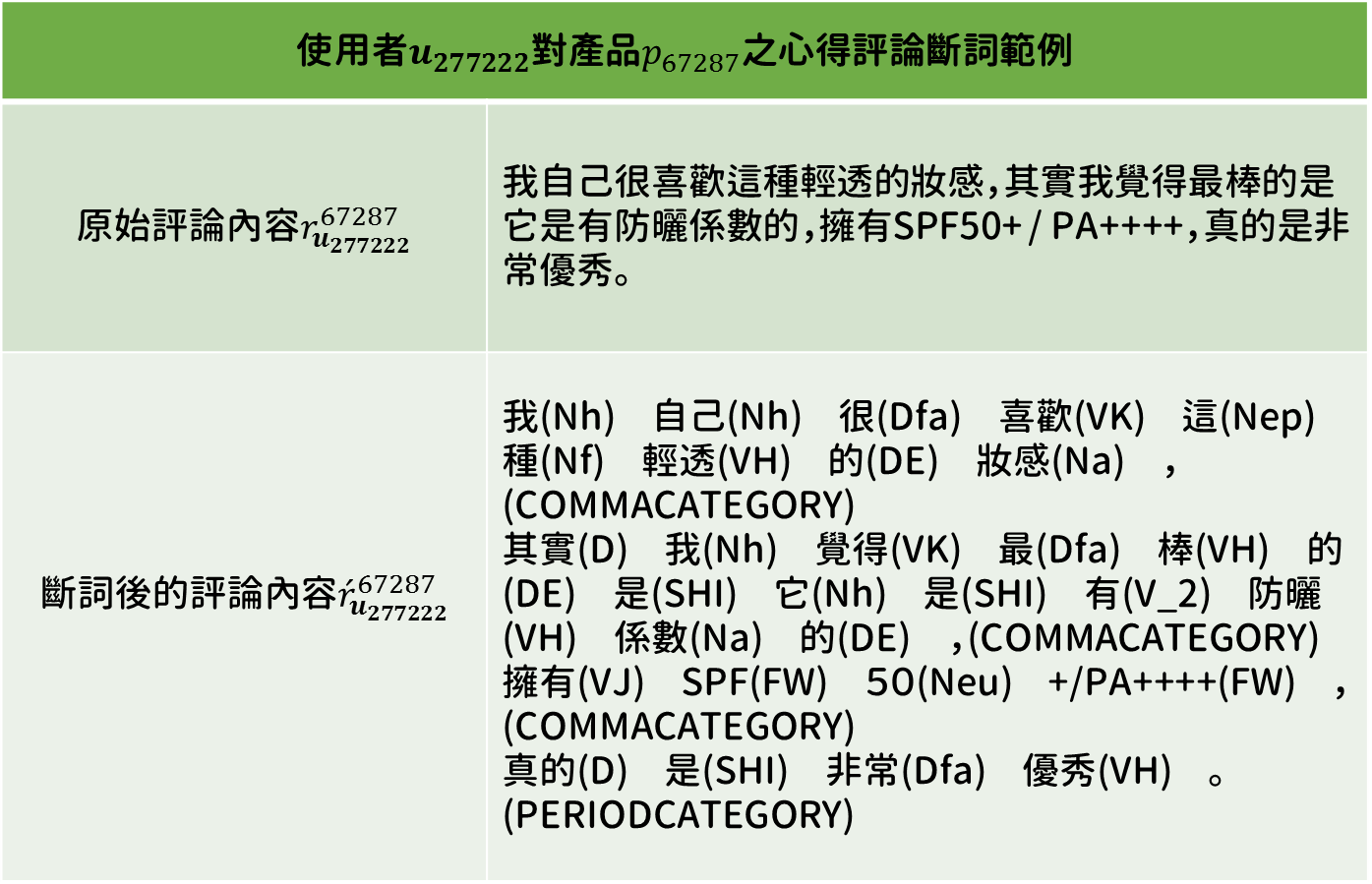


圖 3‑6 心得評論斷詞範例

## 使用者分群

此模組中將會對資料進行相似度計算以及分群處理。首先，取得使用者資料後，會先以使用者膚質對其進行分類，再針對某一類別下的使用者進行相似度計算。由於資料經過前處理後預計會得到由使用者與其使用過之產品所構成的「使用者－產品矩陣」矩陣。取得此矩陣後，將會進行相似度計算，希望能計算出資料集中使用者與使用者之間的相似程度，以利於後續分群處理。「使用者－產品矩陣」中將會記錄使用者與產品之間的關係，使用者若用過該產品，則其值為1，反之則為0。相較於其他相似度計算方式，餘弦相似度計算更適用於兩向量維度不相等的相似度計算情況，因此在相似度計算的公式選擇上([Mahdavi et al., 2008](#_ENREF_23))，本研究將選用最常被用來衡量兩向量間夾角的餘弦相似度進行運算公式如下所示，其中代表的是使用者m所用過之產品向量，則為使用者*Uq*所使用之產品向量。

|  |  |
| --- | --- |
|  | (3-1) |

以下將針對使用者分群進行範例說明。取得使用者資訊並依照膚質進行分類處理後，可將使用者分為五類，如圖 3‑4所示。以膚質中之混和肌類()舉例說明使用者分群過程：先針對混合肌類別下之使用者進行相似度計算(因使用者及其使用產品眾多，在此僅隨機挑選4位使用者及5項產品進行說明)，如表 3‑1所示。在相似度計算上，會以兩兩使用者進行計算，以表 3‑1中之使用者與使用者為例，將其使用過之產品關係向量代入餘弦相似度計算公式後，可得此兩者相似度約為0.3536。依此類推，經過計算過後即可得使用者相似度矩陣如表 3‑2所示。

表 3‑1 使用者與產品使用關係矩陣

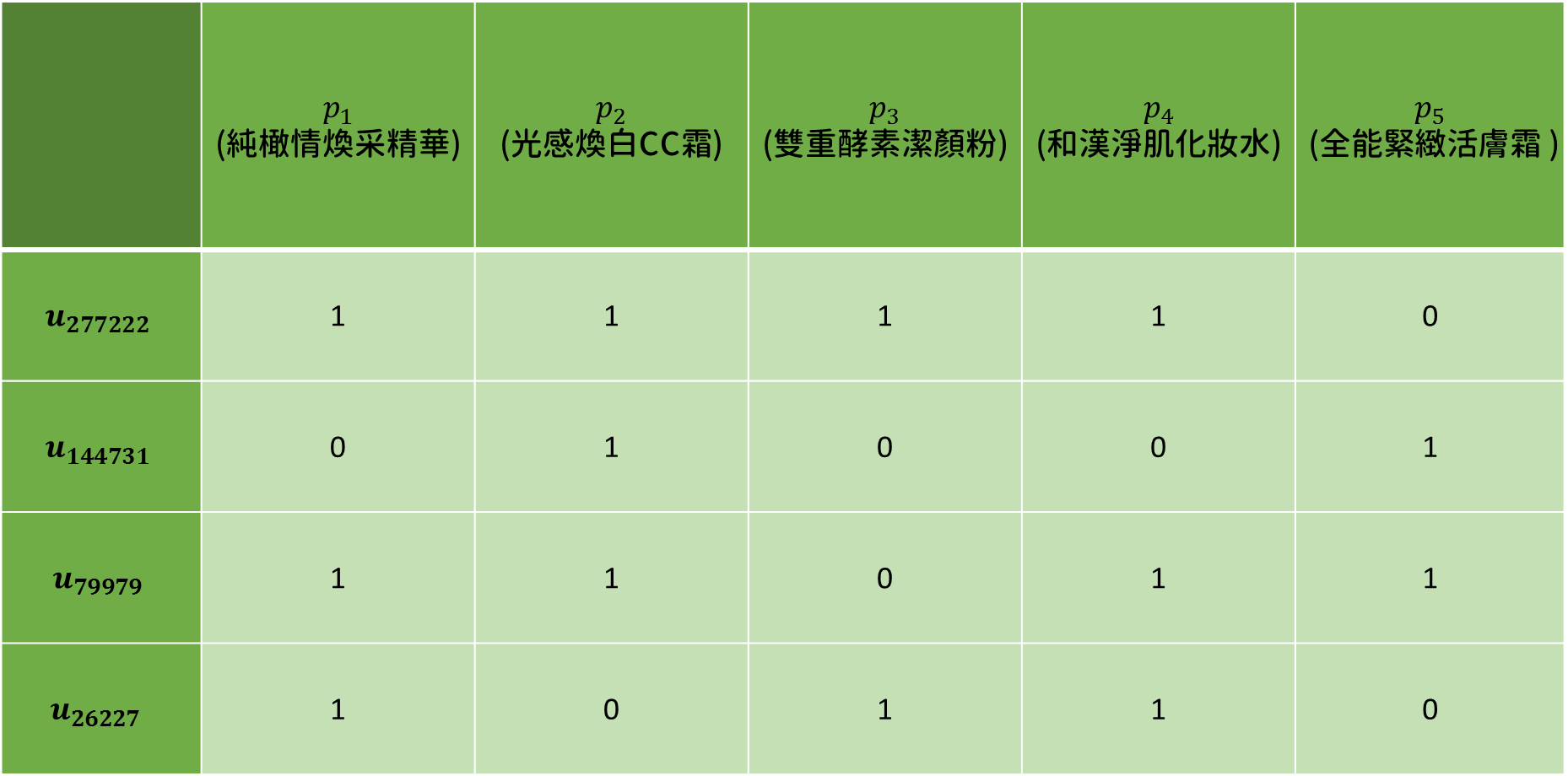
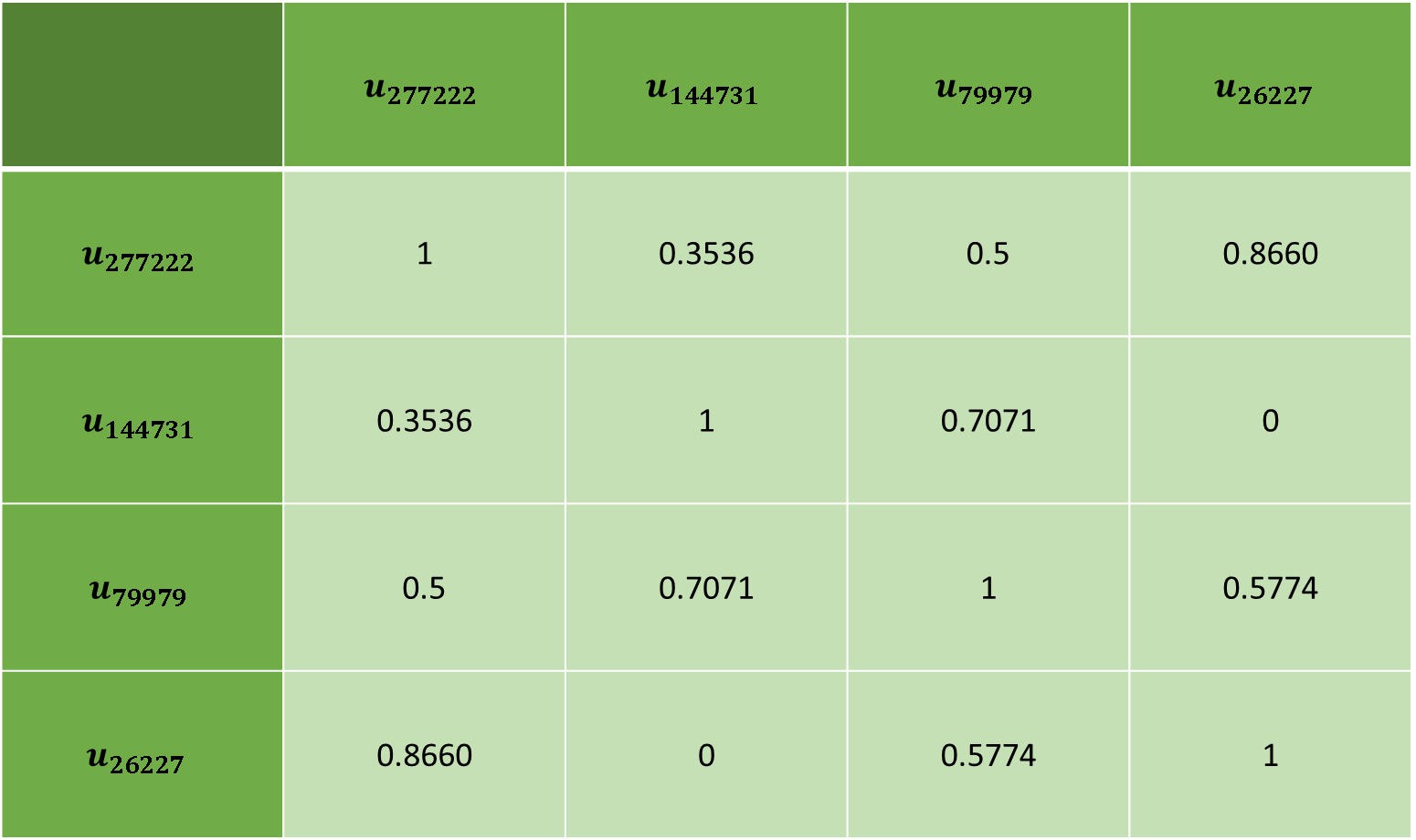


表 3‑2 使用者間相似度矩陣



其次則是需要進行分群處理。結束相似度計算過程後，便能對使用者及產品進行處理。首先會對使用者進行分群，得到使用者分群結果後，才會對相似使用者所形成之使用者群集內的使用者所使用之產品進行分類。對使用者分群的目的是為了將資料集中的使用者成功劃分成不同的群集，如此才能找出與目標使用者其最為相似之使用者群集，使此研究之推薦結果能更貼近使用者需求；完成使用者分群後，便可以收集每一群使用者所使用之產品，再針對這些相似使用者所使用過之產品進行類別對應。

分群部分所要使用的演算法為K-medoids演算法，其概念與K-means十分相似，但在群心選擇上，K-Medoids採用群集中最接近中心的資料點作為群心，與K-means採用群集中各資料的平均值作為中心點不同，但分群方式差異不大，演算法如圖 3‑7所示。

|  |
| --- |
| **Input:** , training users; *,* the number of clusters  **Output:** , *k* centroids |
| 1: Arbitrary selection of the k users as medoid points out of training users.  2: After selection of the k medoid points, associate each user in the given data set to most similar medoid.  3: Randomly select non-medoid user.  4: Compute total cost of swapping initial medoid object to non-medoid user.  5: If cost <0, then swap initial medoid with the new one (if cost <0 then there will be new set of medoids).  5: Repeat steps 2 to 5 until there is no change in the medoid.  6: **return** |

圖 3‑7 K-medoids分群演算法

接續前面之範例說明，在計算完相似度後便可以對使用者進行分群，假設在混合性膚質類別下的四個使用者被分為A、B兩群(使用者和使用者為A群，使用者和使用者為B群)。此膚質類別下之使用者使用產品分別如表 3‑1所示，包含：純橄情煥采精華、光感煥白CC霜、雙重酵素潔顏粉及和漢淨肌化妝水。則在A群使用者(使用者和使用者)中其使用過之產品(為取兩使用者使用過之產品交集)如表 3‑3所示，包含純橄情煥采精華、光感煥白cc霜、雙重酵素潔顏粉及和漢淨肌化妝水。又因每一種產品分別隸屬不同類別，如表 3‑4所示，純橄情煥采精華在美妝保養社群網站上被歸類為精華類別、光感煥白cc霜為cc產品、雙重酵素潔顏粉為洗顏粉類，和漢淨肌化妝水則屬於化妝水類，可以看出這些產品分別屬於4個類別。此時可將A群使用者所使用之產品被分為四類。

表 3‑3 分群後使用者與產品關係範例

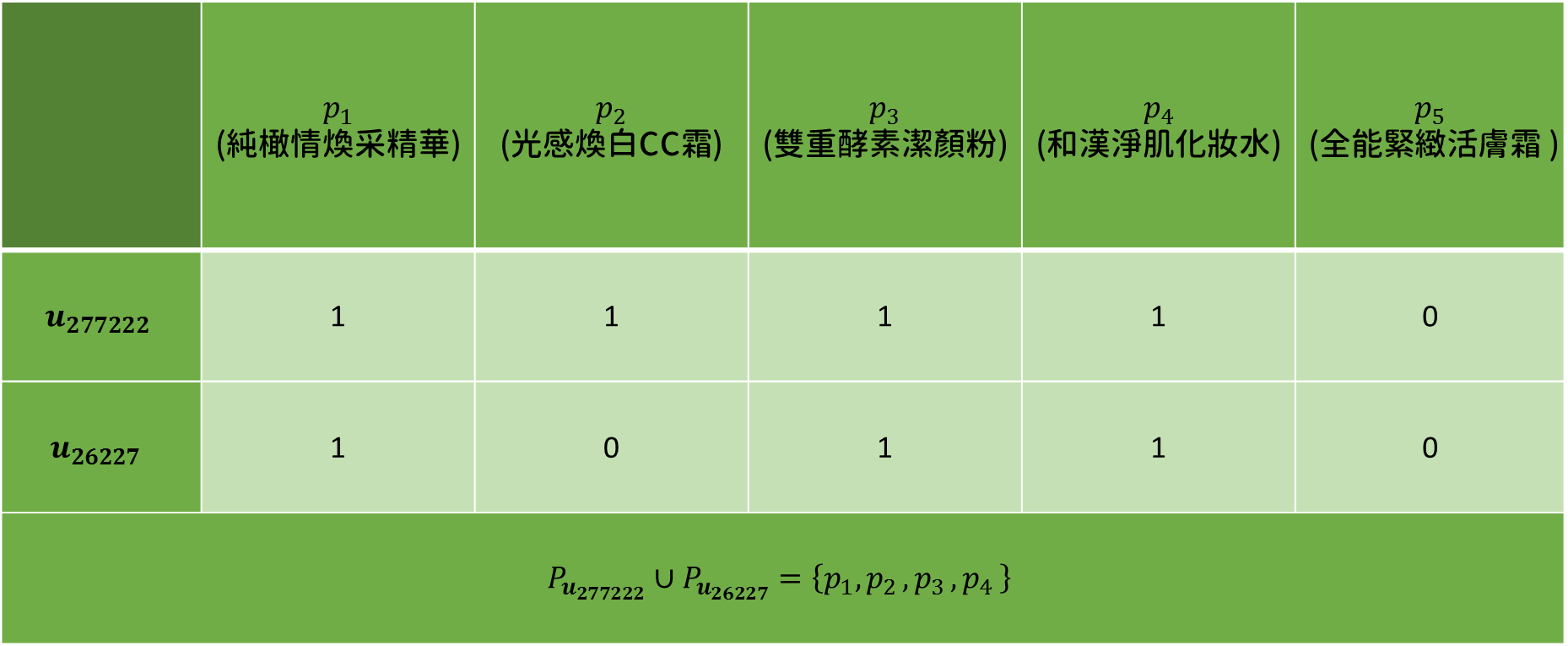
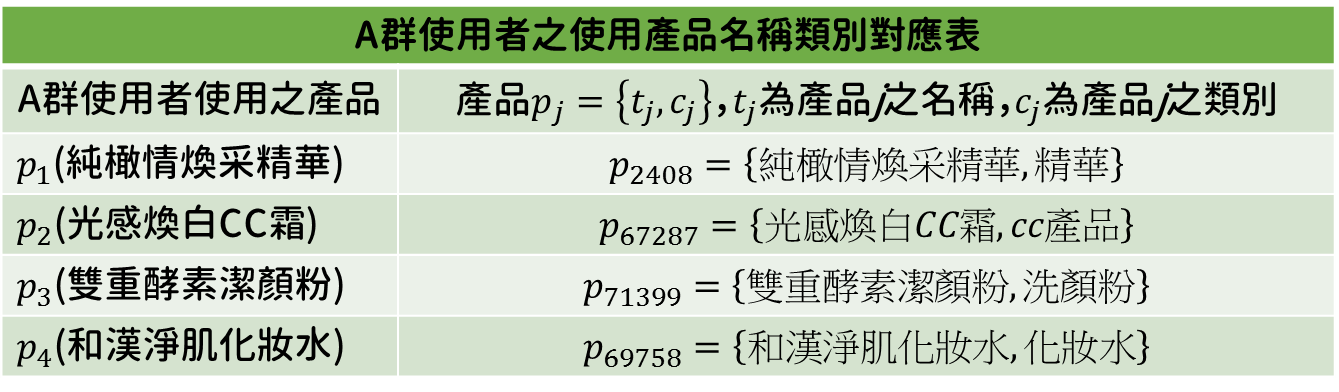


表 3‑4 使用者使用之產品名稱對照表



## 產品集合

當上一模組之分群程序結束後，便可得到某類別下分成k群的使用者群集，接下來便能在此模組中進行產品集合之定義。在這個模組中，將會按照上一模組之結果，對每一群使用者所使用之產品進行產品集合的處理，其概念是對使用者分群之結果中的每一相似使用者進行檢視，將使用者所使用過之產品所屬之產品群進行對應，若經過對應步驟後發現使用者其使用之產品落在不同產品類別中時，則令此兩產品類別間產生一連結，依此類推，直到相似使用者依序被檢視完畢。此時再針對產品群與產品群之間的連結進行數量計算，重複出現次數最多的產品連結組合則視為是此群使用者之最適產品集合，演算法如圖 3‑8所示。

|  |
| --- |
| **Input:** , users belong to cluster ; *,* the number of clusters; , products used by , , categories of products  **Output:** , links between m+1 products |
| 1: Check every user , find the product the user had use.  2: If the products used by user belong to different categories, let every two categories have a link.  3: Repeat step 2 until every user have been checked.  4: Calculate the links.  5: Find the links which have been used the most.  6: **return** |

圖 3‑8 產品集合演算法

接續前面的例子，當A群使用者所使用之產品被分為五個類別後，便可以回頭檢視此群中使用者所使用之產品分別落在哪些類別。以使用者為例，其所使用之產品分別為純橄情煥采精華、光感煥白cc霜、雙重酵素潔顏粉及和漢淨肌化妝水，則由圖 3‑8的演算法可以得到使用者所使用之產品所產生之產品類別關係為：精華類－cc產品類－洗顏粉類－化妝水類，如表 3‑5所示；以此類推，A群中的使用者所建立之產品類別連結則為精華類－洗顏粉類－化妝水類。因和使用者所建立之產品連結分別為精華類－cc產品類－洗顏粉類－化妝水類及精華類－洗顏粉類－化妝水類，恰巧精華類－洗顏粉類－化妝水類重複出現次數為2，且此群中使用者已分析完畢，因此可歸納出適合此群使用者之產品類別集合為精華類－洗顏粉類－化妝水類。

表 3‑5 使用者使用之產品類別歸納範例

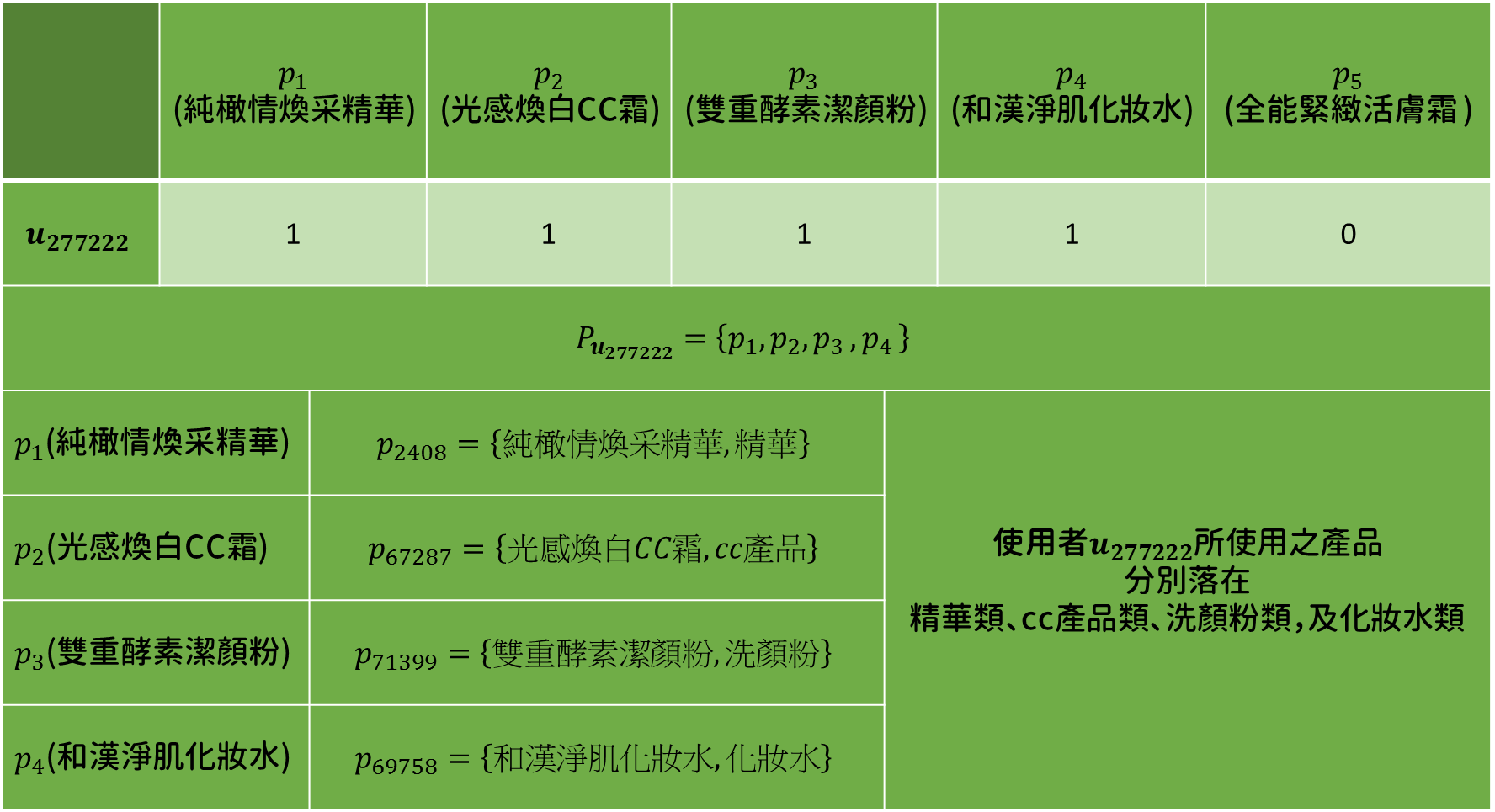
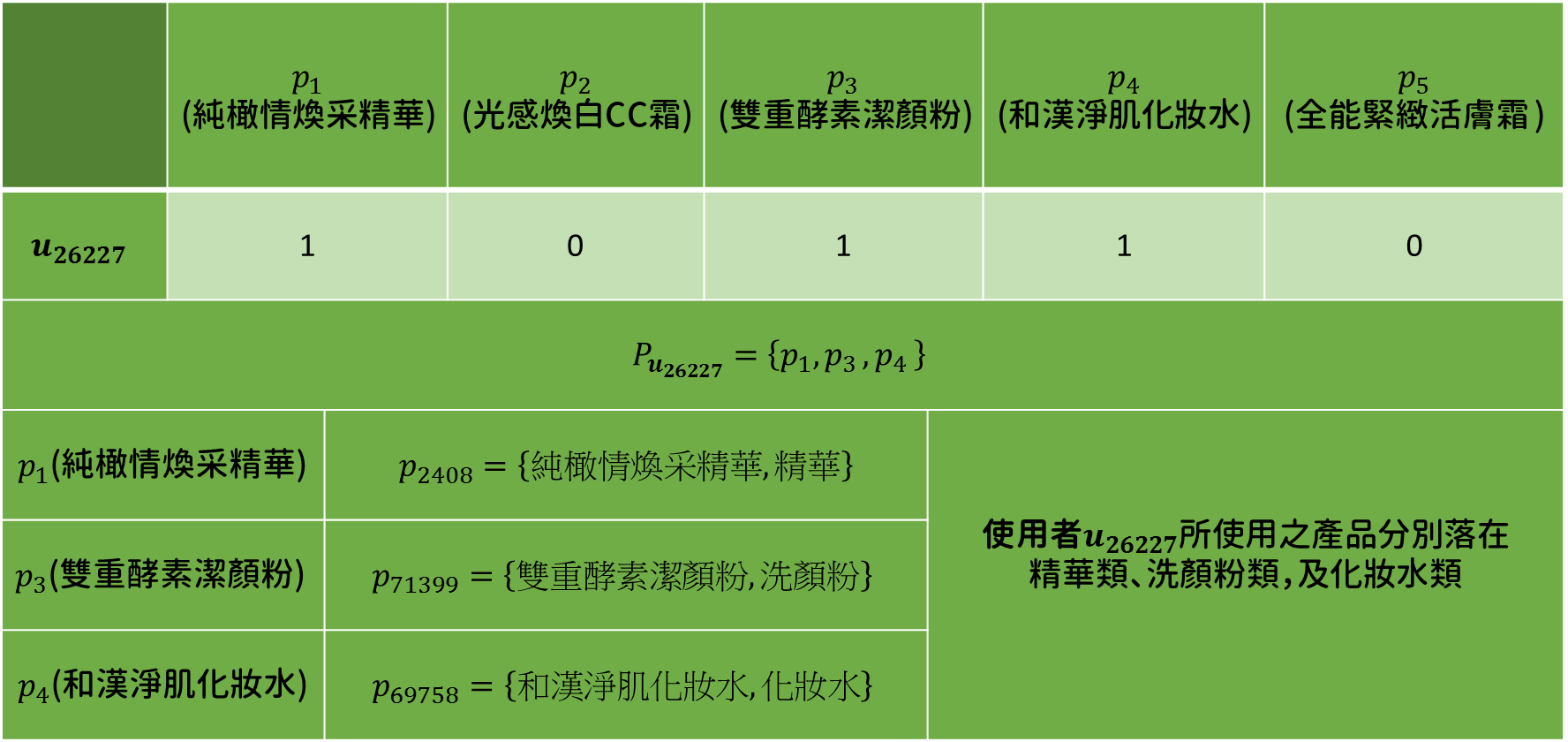


表 3‑6使用者所使用之產品類別歸納範例



## 產品集合推薦

從上一個模組得到產品集合之定義後，此模組將會再對產品類別中之產品進行排序；基於本研究之假設：「對於某產品曾經發表心得評論或是購買此產品之使用者，皆視為曾經使用該產品」，所以會將相似使用者發表過之心得評論數量及曾購買過數量的加總作為產品被使用次數。若相似使用者使用某產品次數越多，則此時該產品的排序名次會越前面，但若單純以此作為排序依據可能會發生多數使用者使用該產品，使其排名越前面，可是反觀這些使用者對該產品之評價卻呈現產品不佳的評論時，將會使得排序結果不夠公允。因此本研究又將利用使用者對產品的心得評論進行分析，先找出相似使用者對某產品之全部的評論，然後逐一對每篇評論進行情緒分析，找出使用者們對於該產品的情緒分數，單篇評論之情緒分數如下圖 3‑9所示。

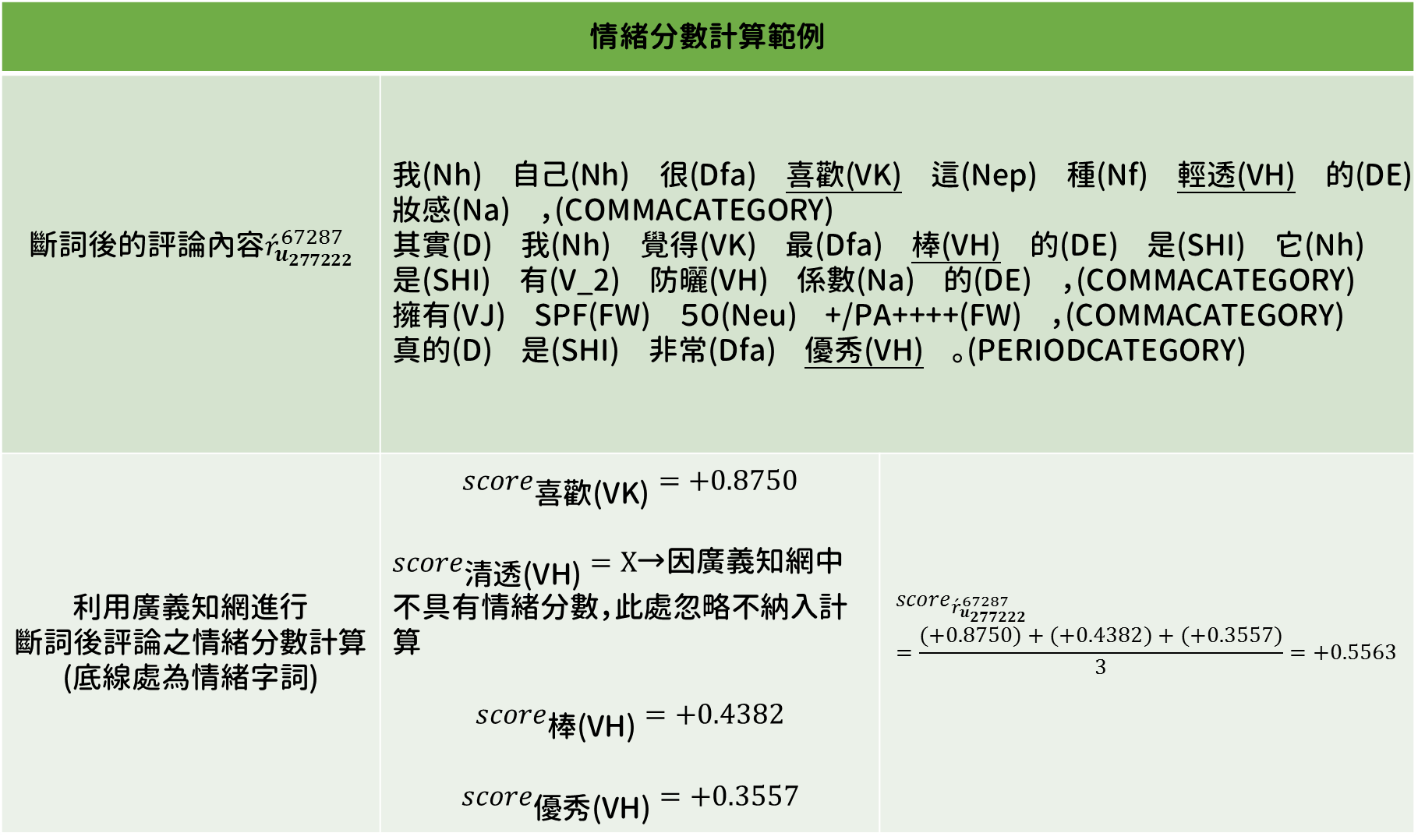


圖 3‑9 單篇心得評論情緒分數計算圖示

　　計算情緒分數之公式如下所示，計算群集*G*內使用者對產品*j*的全部評論之情緒分數，再除以相似使用者對產品*j*的評論總篇數*k*，以得到最後對於產品*j*的*SentiScore*。其中使用者對產品*j*的單篇評論情緒分數計算方式為計算每一篇評論中出現之情緒字詞的分數，在除以每一篇評論中的情緒字詞個數，以避免有些評論中情緒字詞大量出現，而有些評論的情緒字詞卻寥寥可數的情況。

|  |  |
| --- | --- |
|  | (3-2) |
|  |  |
|  | (3-3) |

在計算產品被使用次數的部分則是會將產品被使用之次數進行正規化，公式如下：

|  |  |
| --- | --- |
|  | (3-4) |

其中，*Max*代表的意思為此產品在資料集的最大被使用次數，*used*則代表某一群使用者使用此產品次數。此處是將使用次數進行正規化，目的是為了避免產品被使用次數分佈範圍太大，例如有些產品被使用多達100次，但有些產品卻只被使用不到10次，此時若缺乏正規化的手續，有可能會影響後續分數的計算，且正規化可使其數值介於0到1之間。

最後的排序則是會將使用次數以及情緒分數進行加權考量後所得到的最後分數作為排序依據，公式如下所示：

|  |  |
| --- | --- |
|  | (3-5) |

為一衡量參數，以對和給予不同的權重，實驗中將對此參數進行實驗以求最佳推薦結果。排序完成後將依據排序前*n*高的產品作為該類別的產品推薦清單，蒐集完全部產品類別集合中的產品後便能得到對目標使用者的產品集推薦。

承接前述之範例說明，當得到A群使用者所產生的產品集合：精華類－洗顏粉類－化妝水類後，再分別蒐集A群使用者對這些類別內的產品的評論進行情緒分析。以精華類為例，蒐集A群使用者對該類產品所撰寫過之所有產品評論，在本例中產品為(純橄情煥采精華)，心得篇數為2(因群內有2位使用者撰寫過心得評論)。情緒分析完畢並計算使用者對於產品單篇評論心得分數，分別得到+0.5563及+0.6227，再根據群內使用者對於該產品的所有評論之情緒分數計算，此即為A群使用者對於產品的情緒分數。

假設產品(純橄情煥采精華)在全資料集中最大被使用次數*Max*為40次，而此產品在A群使用者中的被使用次數*used*為2，則該產品。最後再根據的計算方式計算得到產品(純橄情煥采精華)的最後排序分數為0.43125(假設參數)，則)。

## 小結

本研究之研究目的為採用美妝保養社群網站所提供的使用者膚質狀況、產品資訊，以及曾經對於產品所發表過的產品使用心得評論資料等，經過分析後，找出使用者可能感興趣的產品並將其集結成產品集後，對目標使用者進行不同種類產品所形成的產品集合推薦。研究架構分為資料收集與前處理模組、使用者與產品分群模組、產品集合模組及產品集推薦模組，並欲透過這些模組找出相似的同一群使用者所建構之包含不同類型產品的產品集，以作為其他與此群使用者相似之目標使用者進行購買決策之參考依據。

在資料前處理模組中將會利用自行撰寫之程式蒐集美妝保養社群網路平台的使用者資訊，包含：使用者編號、使用者膚質狀況、使用者產品心得，及產品資訊(包含產品名稱及產品心得篇數)等，並針對資料集進行前置處理：對資料集中之使用者資訊及其使用產品之狀況進行特徵擷取，以利於計算使用者間之相似度。在使用者分群模組中，當資料經過前處理後會得到「使用者—使用者產品」矩陣，此時將利用此矩陣後進行相似度計算，找出資料集中使用者與使用者間的相似程度，相似度計算部分採用餘弦相似度。

計算完相似度後便能對使用者進行分群以對目標使用者找出與其最為相似之使用者群集；而完成使用者分群後，便可以蒐集每一群相似使用者所使用之產品，再針對這些產品進行產品集合的處理，其概念是針對分群結果中與目標使用者相似之群集的每一群內使用者進行檢視，將使用者所使用過之產品所屬的產品類別進行對應，若經過對應步驟發現使用者使用之產品落在不同產品類別中，則使此兩產品類別有一連結，依此類推直到群內使用者依序檢視完畢。此時再針對產品類別與類別之間之連結進行數量計算，重複出現最多次的連結組合則視為此群使用者之最適產品種類集合。得到產品集合之定義後，再對產品類別中之產品進行排序。而因在進行產品推薦時，需要對該類別之產品進行排序，但若以該產品被使用次數作為排序標準，有可能會發生使用者使用過該產品但對該產品之喜好程度不如預期，所以排序之依據無法單單使用產品被使用次數，而是需加入經由情感分析進行心得評論計算後所得之使用者對於產品的情感分數，以產生較為較貼近可能使用結果之排序。

# 系統建置與驗證

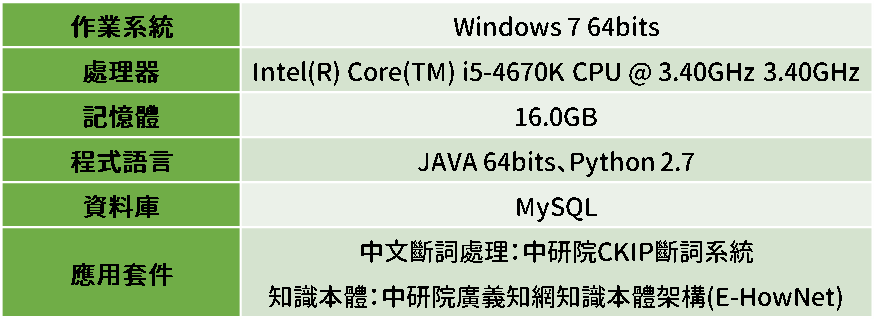
為了驗證本研究所提出之個人化產品集合推薦的有效性，在本章節中將進行實驗、驗證分析與建議。依序將介紹本研究之實驗環境建置及實驗方法設計，

並將利用特定評估指標衡量實驗結果。最後再針對實驗之結果進行分析與討論。

## 系統環境建置

本研究利用JAVA程式語言撰寫爬蟲程式(crawler)蒐集美妝產品評論網站UrCosme上的產品資料以及使用者資料；並使用MySQL資料庫管理系統儲存實驗資料集。計算相似度並將使用者分群的使用者分群模組則是利用程式語言Python進行程式開發。詳細的系統環境如下表所示。

表 4‑1系統建置環境



本研究實驗系統處理流程主要分為四個階段，分別為：資料前處理、使用者分群、產品集合，以及產品集合推薦，以下將針對前述流程進行說明。

1. 資料前處理階段：

實驗所蒐集之資料集包含使用者對產品之評論、使用者的膚質類別以及產品資訊等。其中，因資料來源網站特性，一旦網站使用者超過六個月未登錄該網站，則該網站會刪除該使用者資料但保留其對於產品所撰寫之心得評論，造成使用者編號為空值而評論存在的現象。又當使用者為空值時，並沒有辦法在資料集中辨識該使用者所使用過的產品，因而無法進行使用者相似度計算，此時必須將使用者編號為空值之資料在此階段中先移除，以利後續實驗進行。另外，因為後續實驗必須依照使用者膚質先將使用者進行分類後，再依其使用之產品計算相似度後進行分群，而所蒐集之資料集中卻存在膚質欄位為空值之使用者，因此此部分資料也需在此階段進行移除的動作，以避免影響後續實驗進行。而資料集中的使用者對產品之評論資料也會一併在此階段利用CKIP斷詞系統進行斷詞處理並標記詞性，以便於產品推薦階段進行推薦分數計算。

1. 使用者分群：

此階段會依所蒐集之資料，按照使用者膚質將使用者進行分類後，再針對每一個膚質類別下的使用者與該類別使用者所使用的產品，撰寫程式計算類別下使用者間之相似度，其中相似度計算公式採用能夠計算兩向量夾角的餘弦相似度進行運算，並利用此相似度計算公式作為K-Medoids分群演算法中計算距離的公式，再利用此分群演算法找出每一個膚質類別下適合作為群心的使用者，並將相似使用者分為同一群，做為下一階段找出產品集合的依據。在此階段中，也會進行分群演算法門檻值以及群集數量對於推薦效果影響的實驗。

1. 產品集合：

在此階段中，會按照上一階段之每一個類別下的分群結果找出每一個使用者群集所使用之產品落在哪些產品類別中，並使類別與類別間產生一連結，直到所有使用者之產品皆判定完畢後，計算最後產品與產品間的連結數量，並依此結果找出該群使用者的產品類別關係做為產品類別集合，下一階段將以此為依據進行產品集合推薦。

1. 產品集合推薦：

此為本研究最後一個實驗階段，在本階段中將會以上一階段所得之產品類別集合為依據，利用每一群使用者所使用之產品次數以及產品之評論情緒分數加入權重調合計算後，進行產品排序，並衡量其推薦效果。

## 實驗方法

本研究主要目的為找出在具有類別特性的資料集下產品類別之間的關係，並利用產品類別之間的關係進行推薦，期望能得到較貼近使用者的推薦結果。方法是先將使用者分類後再分群以降低資料稀疏性，並期望得知加入使用者評論後對於推薦效果之影響。且為了讓本實驗之資料集能夠利用K-Medoids演算法進行分群，在實驗過程中以不影響演算法之前提下增設門檻值，以期能將實驗效率優化。因此本研究主要有以下實驗：

1. 實驗一：找出演算法中能夠同時兼顧分群效率及品質的門檻值

因為K-Medoids演算法的特性，若是按照其原始算法進行，在演算法第二階段必須將每一個使用者與做為起始群心的使用者進行相似度計算，並找出與群心距離成本最小的使用者做為新的群心，而此過程在使用者數量較多的時候會十分耗時，因此為了因應資料集中所形成之「使用者－產品」矩陣可能會因為使用者數量以及產品數量龐大，而需花費較長的時間進行使用者分群，因此本研究在演算法中加入一門檻值，目的在於減少演算法中找出最適群心的時間成本。又不同門檻值對於時間具有不同影響，本實驗目的在於比較未加入門檻值以及不同門檻值對於分群時間之影響。

1. 實驗二：分群數量對於產品集合推薦之影響

本實驗在於找出分群數量不同的情況對於產品類別集合是否有所影響，並找出分群數量選擇是否對最後階段之產品集合推薦有所影響。另外，推薦不同數量的產品類別集合對於推薦效果也會有影響，因此也會在此實驗中進行推薦數量的比較。

1. 實驗三：推薦的準確性

本實驗用意在於比較以次數作為最後階段之產品排序與加入評論情緒分數一起計算後之排序結果。又在加入評論情緒分數計算時，會需一權重值來調和次數以及評論情緒分數，而此權重值之決定亦會在此實驗中進行。

### 資料來源

根據創市際市場研究顧問於2015年的調查，UrCosme為台灣Top10的美妝保養評論網站([InsightXplorer創市際市場研究顧問, 2015b](#_ENREF_14))。本研究蒐集該網站的產品評論以及使用者資訊，網站局部頁面如圖 4‑1所示。

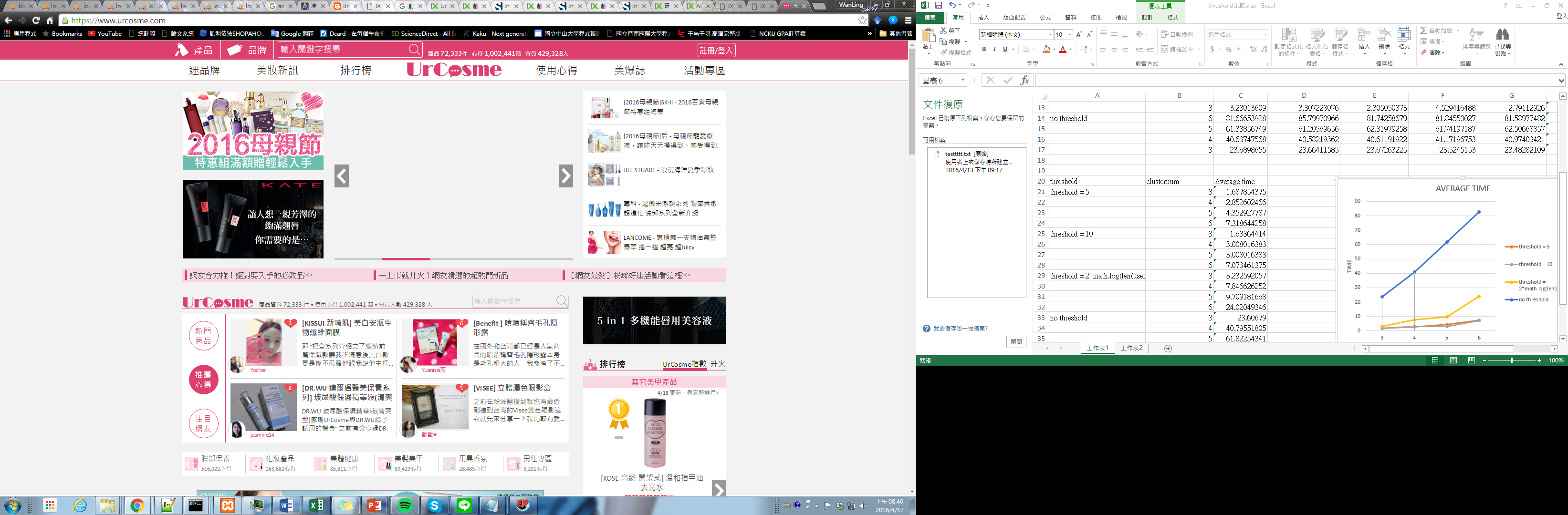
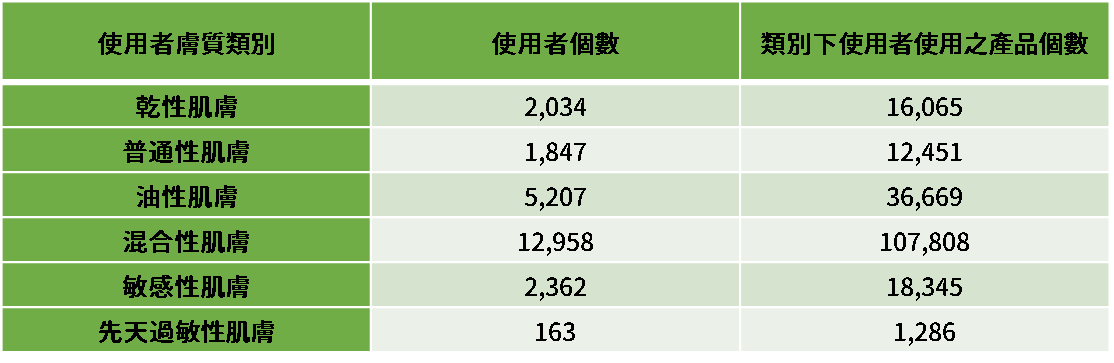


圖 4‑1 UrCosme網頁局部圖示

資料蒐集範圍為臉部保養類別下的其中七個產品類別，包含：臉部卸妝、眼唇卸妝、洗顏、化妝水、凝膠/凝凍、乳液，以及乳霜。資料筆數為208,678筆評論，移除使用者為空值後的資料筆數為208,592筆，根據使用者膚質類別分類後的資料分布情況如表 4‑2所示，其中使用者個數以及類別下使用者使用之產品個數為已經移除只對一項產品發表評論之使用者數量的結果；移除只對一項產品發表評論之使用者是因為若該使用者只使用過單一產品，除了影響分群結果外，其在實驗意義上代表該產品類別只有一個，與本研究所欲找出之多個產品類別關係形成之產品類別集合相牴觸，且其數量會影響後續推薦結果，因此在實驗過程中，並不將只使用單一產品之使用者納入考量。

表 4‑2 資料集分佈表



### 評估指標

本研究所採用之評估指標為在推薦系統中常用來衡量推薦品質的*precision*([Liu & Shih, 2005](#_ENREF_20))，其意涵為推薦的項目是否為使用者真正感興趣的，公式如下所示。

|  |  |
| --- | --- |
|  | (4-1) |

在本研究中，分別會在產品類別集合推薦以及產品集合推薦兩部分利用*precision*衡量推薦效果。在衡量產品類別集合推薦時，*precision*的計算方式為推薦的產品類別集合是否能包含使用者曾經使用之類別集合；而在衡量產品推薦效果時除了使用*precision*外，還會另外採用*hit-rate (HR)*來衡量推薦效果。*hit-rate*指的是全部的使用者中，系統所推薦的

## 實驗結果

### 實驗一

此實驗目的在於比較加入可以減緩分群時間成本的門檻值與未加入此門檻值的分群時間成本。本實驗以表 4‑2的先天過敏性肌膚類別以及油性肌膚類別做為比較對象。從表 4‑2可以看出先天過敏性肌膚類別中的使用者人數最少，而油性肌膚則擁有次多之使用者。未採用使用者人數最多之混合性肌膚類別進行本實驗之原因為在本實驗中需比較未設立門檻值與已設立門檻值的影響，當進行未設立門檻值的實驗時，在使用者人數最多之混合性膚質類別需要花費較長的時間且十分消耗電腦資源，因此轉以人數次多之油性肌膚類別進行門檻值效果的比較。

下圖為在先天過敏性肌膚類別資料集上測試門檻值設立成效的結果，可以看出在未設立門檻值的情況下，若分群的數量增加，則所需的時間成本越大。而一開始挑選門檻值為長數5時可以發現，所需的時間成本平均而言最小，其次則是長數設立為10時。但因此資料集使用者人數較少，所以在設有門檻值的情況下，平均演算法執行時間約在30秒內。

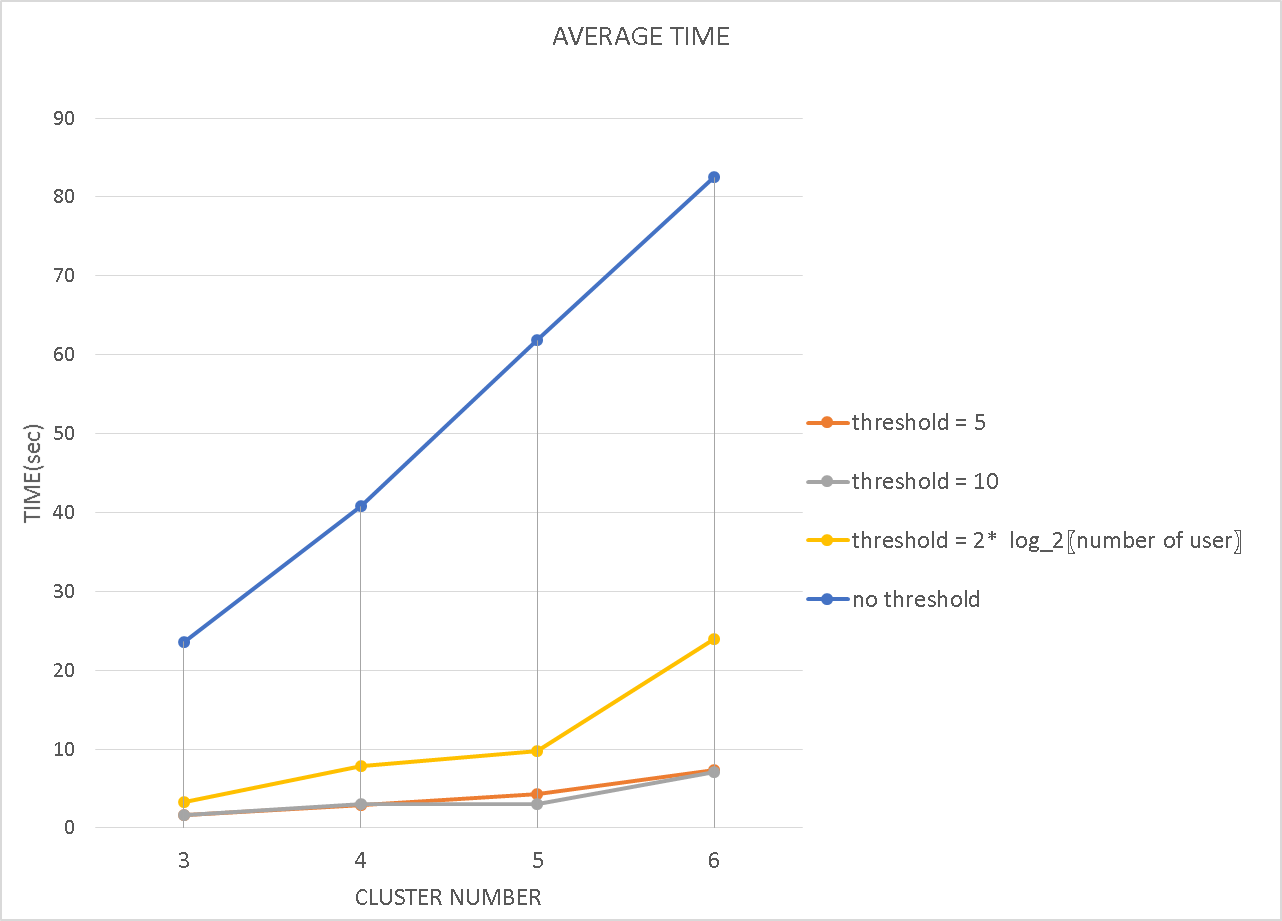


圖 4‑2 門檻值對於時間成本之影響

### 實驗二

實驗二在於找出分群數量不同的情況對於產品類別集合是否有所影響，並找出分群數量選擇是否對最後階段之產品集合推薦有所影響。如圖 4‑3所示，可以發現，平均而言分類後的precision會較未分類前的表現來的好，圖中的每一個直條代表的是不同膚質類別下各推薦一個產品類別集合的效果。

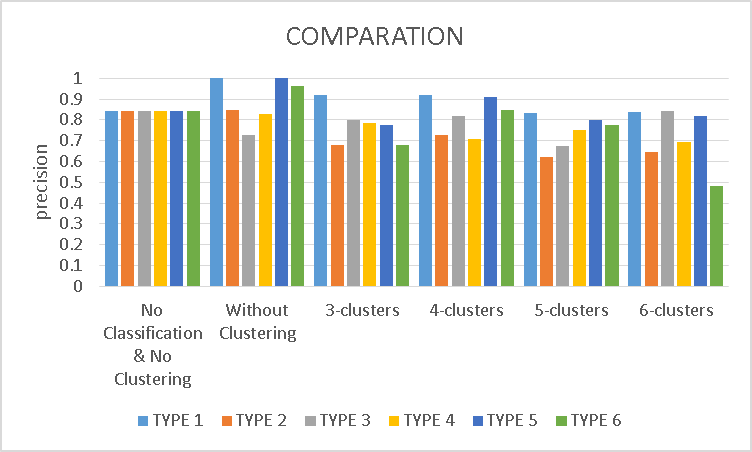
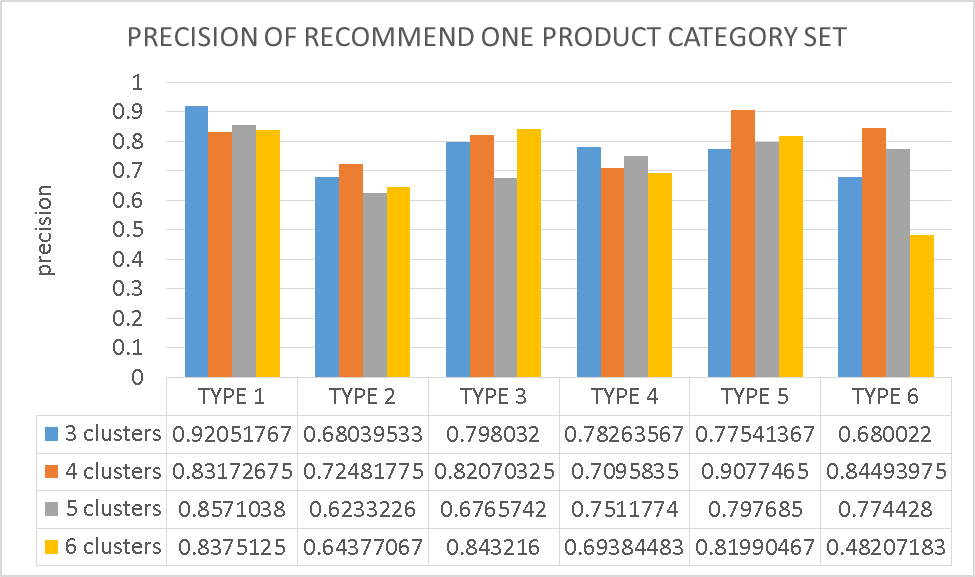


圖 4‑3未分類未分群、僅分類及分類又分群在推薦效果的比較圖

下圖則為分類後在進行分群的推薦效果比較圖，如圖所示可以發現，在不同類別集合下，分群數目不同會有不同的推薦效果。其中，在膚質類別為乾性肌膚的情況下(如圖中的TYPE\_1)，以分三群有較好的推薦效果。



### 實驗三

本實驗用意在於比較以次數作為最後階段之產品排序與加入評論情緒分數一起計算後之排序結果。又在加入評論情緒分數計算時，會需一權重值來調和次數以及評論情緒分數，而此權重值之決定亦會在此實驗中進行。評估指標將會採用*precision*以及*hit-rate*。

# 目前及預期進度

目前已利用所蒐集之資料完成分群演算法門檻值比較以及利用產品次數進行推薦的實驗，且正在進行評論情緒分數計算，以及加入情緒分數之推薦效果比較之實驗。在後續進度規劃部份，預計在口試前完成所有實驗並進行實驗分析以及實驗評估，並依據老師所給予之建議進行修改，且論文內容會持續撰寫並反覆修改。

# 參考文獻

Anand, D., & Bharadwaj, K. K. (2010). Enhancing Accuracy of Recommender System through Adaptive Similarity Measures Based on Hybrid Features. In N. T. Nguyen, M. T. Le & J. Świątek (Eds.), *Intelligent Information and Database Systems: Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part II* (pp. 1-10). Berlin, Heidelberg: Springer Berlin Heidelberg.

Barragáns-Martínez, A. B., Costa-Montenegro, E., Burguillo, J. C., Rey-López, M., Mikic-Fonte, F. A., & Peleteiro, A. (2010). A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. *Information Sciences, 180*(22), 4290-4311. doi: <http://dx.doi.org/10.1016/j.ins.2010.07.024>

Cambria, E., Schuller, B., Yunqing, X., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *Intelligent Systems, IEEE, 28*(2), 15-21. doi: 10.1109/MIS.2013.30

Chang, P.-C., Galley, M., & Manning, C. D. (2008). *Optimizing Chinese word segmentation for machine translation performance*. Paper presented at the Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio.

Chen, K.-J., & Bai, M.-H. (1998). Unknown Word Detection for Chinese by a Corpus-based Learning Method. *Computational Linguistics and Chinese Language Processing, 3*(1), 27-44. doi: citeulike-article-id:4746407

Chen, K.-J., & Liu, S.-H. (1992). *Word identification for Mandarin Chinese sentences*. Paper presented at the Proceedings of the 14th conference on Computational linguistics - Volume 1, Nantes, France.

Chen, K.-J., & Ma, W.-Y. (2002). *Unknown word extraction for Chinese documents*. Paper presented at the Proceedings of the 19th international conference on Computational linguistics - Volume 1, Taipei, Taiwan.

Chiang, H.-S., & Huang, T.-C. (2015). User-adapted travel planning system for personalized schedule recommendation. *Information Fusion, 21*, 3-17. doi: <http://dx.doi.org/10.1016/j.inffus.2013.05.011>

FashionGuide華人第一女性時尚美妝傳媒. (2015). FashionGuide華人第一女性時尚美妝傳媒. from <http://www.fashionguide.com.tw/>

Fu, G., & Luke, K.-K. (2005). Chinese named entity recognition using lexicalized HMMs. *SIGKDD Explor. Newsl., 7*(1), 19-25. doi: 10.1145/1089815.1089819

Garcia Esparza, S., O’Mahony, M. P., & Smyth, B. (2012). Mining the real-time web: A novel approach to product recommendation. *Knowledge-Based Systems, 29*, 3-11. doi: <http://dx.doi.org/10.1016/j.knosys.2011.07.007>

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*(2-3), 107-145. doi: 10.1023/A:1012801612483

InsightXplorer創市際市場研究顧問. (2015a). 美妝保養調查及台灣美妝相關網站使用概況. *創市際雙週刊, 38*.

InsightXplorer創市際市場研究顧問. (2015b). 美妝網站調查與 美容時尚類別網站使用概況.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2016). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*. doi: <http://dx.doi.org/10.1016/j.eij.2015.06.005>

Karol, S., & Mangat, V. (2013). Evaluation of text document clustering approach based on particle swarm optimization. *Central European Journal of Computer Science, 3*(2), 69-90. doi: 10.2478/s13537-013-0104-2

Kumar, N. P., & Fan, Z. (2015). Hybrid User-Item Based Collaborative Filtering. *Procedia Computer Science, 60*, 1453-1461. doi: <http://dx.doi.org/10.1016/j.procs.2015.08.222>

Levy, R., & Manning, C. (2003). *Is it harder to parse Chinese, or the Chinese Treebank?* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan.

Liang, T.-P., Yang, Y.-F., Chen, D.-N., & Ku, Y.-C. (2008). A semantic-expansion approach to personalized knowledge recommendation. *Decision Support Systems, 45*(3), 401-412. doi: <http://dx.doi.org/10.1016/j.dss.2007.05.004>

Liu, D.-R., & Shih, Y.-Y. (2005). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software, 77*(2), 181-191. doi: <http://dx.doi.org/10.1016/j.jss.2004.08.031>

Ma, W.-Y., & Chen, K.-J. (2003). *A bottom-up merging algorithm for Chinese unknown word extraction*. Paper presented at the Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17, Sapporo, Japan.

Madhulatha, T. (2011). Comparison between K-Means and K-Medoids Clustering Algorithms. In D. Wyld, M. Wozniak, N. Chaki, N. Meghanathan & D. Nagamalai (Eds.), *Advances in Computing and Information Technology* (Vol. 198, pp. 472-481): Springer Berlin Heidelberg.

Mahdavi, M., Chehreghani, M. H., Abolhassani, H., & Forsati, R. (2008). Novel meta-heuristic algorithms for clustering web documents. *Applied Mathematics and Computation, 201*(1–2), 441-451. doi: <http://dx.doi.org/10.1016/j.amc.2007.12.058>

Mishra, R., Kumar, P., & Bhasker, B. (2015). A web recommendation system considering sequential information. *Decision Support Systems, 75*, 1-10. doi: <http://dx.doi.org/10.1016/j.dss.2015.04.004>

Morente-Molinera, J. A., Pérez, I. J., Ureña, M. R., & Herrera-Viedma, E. (2016). Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies. *Information Sciences, 328*, 418-434. doi: <http://dx.doi.org/10.1016/j.ins.2015.08.051>

Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences, 272*, 16-28. doi: <http://dx.doi.org/10.1016/j.ins.2014.02.063>

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46. doi: <http://dx.doi.org/10.1016/j.knosys.2015.06.015>

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM, 18*(11), 613-620. doi: 10.1145/361219.361220

Teahan, W. J., McNab, R., Wen, Y., & Witten, I. H. (2000). A compression-based algorithm for Chinese word segmentation. *Comput. Linguist., 26*(3), 375-393. doi: 10.1162/089120100561746

UrCosme. (2015). UrCosme. from https://[www.urcosme.com/](http://www.urcosme.com/)

WebMD. (2015). Skin Types and Care: Normal, Dry, Oily, Combi... from <http://www.wikihow.com/Determine-Your-Skin-Type>

Wikipedia. (2015). Cosmetics. from https://en.wikipedia.org/wiki/Cosmetics#Types

Wong, P.-k., & Chan, C. (1996). *Chinese word segmentation based on maximum matching and word binding force*. Paper presented at the Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark.

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science, 44*(9), 532-542. doi: 10.1002/(SICI)1097-4571(199310)44:9<532::AID-ASI3>3.0.CO;2-M

Xue, N., Chiou, F.-D., & Palmer, M. (2002). *Building a large-scale annotated Chinese corpus*. Paper presented at the Proceedings of the 19th international conference on Computational linguistics - Volume 1, Taipei, Taiwan.

Yang, C. C., Luk, J. W. K., Yung, S. K., & Yen, J. (2000). Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science, 51*(4), 340-351. doi: 10.1002/(SICI)1097-4571(2000)51:4<340::AID-ASI4>3.0.CO;2-I

Zahra, S., Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for KMeans-clustering based recommender systems. *Information Sciences, 320*, 156-189. doi: <http://dx.doi.org/10.1016/j.ins.2015.03.062>

Zhang, L., Hu, C., Chen, Q., Chen, Y., & Shi, Y. (2012). Domain Knowledge Based Personalized Recommendation Model and Its Application in Cross-selling. *Procedia Computer Science, 9*, 1314-1323. doi: <http://dx.doi.org/10.1016/j.procs.2012.04.144>

王京盛. (2012). 考量語意及引用分析之研究主題趨勢分析方法.

吳虹瑩. (2012). *朋友圈餐廳推薦機制之研究.* (碩士), 輔仁大學, 新北市.

林宜瑩. (2010). 利用時間因子與名詞片語之文獻主題追蹤法. *國立成功大學*.

高照明. (2012). 語料庫建構技術-研 究 報 告.

瑞麗美人國際媒體. (2014). 四個使用保養品的正確觀念，打造陶瓷肌！.

資策會產業情報研究所. (2014). 「網路社群使用」調查.

圖書館學與資訊科學大辭典. (2012).

維基百科. (2015a). 社群媒體. from https://zh.wikipedia.org/wiki/%E7%A4%BE%E4%BC%9A%E5%8C%96%E5%AA%92%E4%BD%93

維基百科. (2015b). 社群網路服務. from https://zh.wikipedia.org/wiki/%E7%A4%BE%E4%BA%A4%E7%B6%B2%E8%B7%AF%E6%9C%8D%E5%8B%99

維基百科. (2015c). 虛擬社群. from https://zh.wikipedia.org/wiki/%E8%99%9B%E6%93%AC%E7%A4%BE%E7%BE%A4