

## 2022年6月20日

書籍

推薦系統

數據分析資料工程研讀報告

資工2B S09350246張子中

目錄 (Table of Contents)

章節一：摘要…………………………………………………………………………………3

研究背景與動機………………………………………………………………………3

研究目的………………………………………………………………………………3

推薦系統簡介…………………………………………………………………………3

章節二：相關文獻……………………………………………………………………………5

關聯規則………………………………………………………………………………5

內容過濾………………………………………………………………………………6

協同過濾………………………………………………………………………………6

混合推薦………………………………………………………………………………7

章節三：系統設計與實作……………………………………………………………………8

資料集…………………………………………………………………………………8

資料處理………………………………………………………………………………8

演算法…………………………………………………………………………………8

資料分析………………………………………………………………………………9

章節四：就業相關市場……………………………………………………………………10

章節五：心得報告…………………………………………………………………………13

心得分享……………………………………………………………………………13

未來展望……………………………………………………………………………13

章節六：參考文獻…………………………………………………………………………14

## 章節ㄧ 摘要

研究背景與動機：

在網路成為人與人之間的主要媒介後，伴隨而來的是巨大的商機，而為了讓使用者更投入在產品之中，各大電子商務公司與社群媒體公司等等，為了增加銷量或者提供更好的服務，都紛紛投入研究推薦系統，對使用者來說與其面對琳瑯滿目的選擇而手足無措，如果可以立馬看到喜歡的東西那有多好。

根據 Alex Iskold 於 Read/WriteWeb 發表的一 篇文章指出，人們上網時有兩項基本的線上活動 -搜尋(Search)與瀏覽(Browse)。拜 Google 之賜，現在大家關注的焦點都擺在搜尋上，然而網際網路活動不僅只是搜尋而已，使用者需要的更多。例如當使用者明確知道自己要找尋的項目時，他可以使用搜尋引擎，也就是我們常說的“用 Google search 一 下"。但倘若使用者並無特定的搜尋對象時，他就只是在瀏覽網頁。在這瀏覽的狀態下，使用者需要 一些建議，因此這正是推薦系統(Recommendation System)發揮作用的絕佳機會。

研究目的：

建立圖書推薦系統以爬蟲網路商城博克來為資料集。

推薦系統簡介：

Alex Iskold 在文 章中總結下列幾種推薦類型:

* 個人化推薦(Personalized recommendation) – 基於個人的行為模式進行推薦
* 社 會 化 推 薦 (Social Recommendation) –
* 基於相似用戶過去的行為模式進行推薦
* 項目化推薦(Item recommendation) – 基於項目本身進行推薦
* 上述三種方式的混合推薦

現在網際網路上幾個成功的推薦系統案例(如 Amazon.com、eBay 等)，都是上述多種推薦類型的混合。 推薦系統是一種資訊過濾(Information Filtering) 的技術，主要使用在電影、音樂、書籍、新聞及網頁網站上，引導使用者挑選出最感興趣或有用的關聯項目，使顧客願意回流並再次交易，建立更強健的以網站為基礎之顧客關係管理(Web-Based CRM)。因此於網站上建立一個好的推薦系統， 不但可留住舊客戶亦可吸引新的客戶，將可給網站帶來極大效益;同時對使用者而言，推薦系統是有 別於網頁搜尋活動的另一項選擇，可以幫助使用者 找到自己尚未發掘的關聯項目。

## 章節二 相關文獻

### 關聯規則：

Agrawal, Imielinski and Swami於SIGMOD會議(1993)上提出關聯法則(Association Rules) 是資料探勘領域中一個重要的研究課題，其中心思 想是在一組包含用戶交易資料的資料集中，發現不 同商品或項目間同時發生的頻率並找出項目間的 關聯性規則，用這些規則可以預測用戶的購買行為 模式，並可以幫助企業制定商務決策。

分析主要透過計算支持度(Support)與信心水準(Confidence)來挖掘數據之間的關聯性，支持度(Support) :表示物品集(Ex. 如果只有一個物品({A),如果有想個物品{A,B})在擁有N個資料(ALL\_DATA)的數據庫中出現的次數比例，信心水準(Confidence): 表示兩物品同時出現的條件機率，簡單來說就是在已經出現商品A的情況下，出現商品B的機率，還有更多其他指標可以建立關聯分析，衍伸出各種演算法例如：Apriori演算法、FP-growth算法。

一張含有 文字 的圖片

自動產生的描述

支持度(Support) 表示物品集，Ex. 如果只有一個物品({A),如果有想個物品{A,B})在擁有N個資料(ALL\_DATA)的數據庫中出現的次數比例。

一張含有 文字 的圖片

自動產生的描述

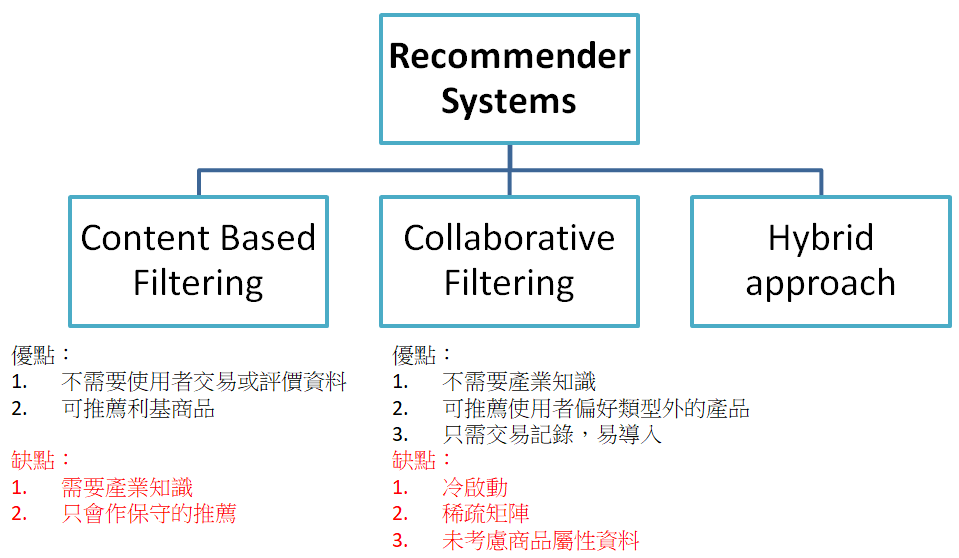
信心水準(Confidence)表示兩物品同時出現的條件機率，簡單來說就是在已經出現商品A的情況下，出現商品B的機率。

一張含有 文字, 室外, 標誌, 夜空 的圖片

自動產生的描述

提升度(Life) 表示當經出現商品A的情況下，出現商品B的機率，但會看出只出現商品B的機率的問題，提升度(Lift)代表著數據間的關聯性。

進階的推薦系統：



### <https://ithelp.ithome.com.tw/articles/10220962>

### 內容過濾：

### 優點：不需要使用者交易或評價資料、可推薦利基商品。

### 缺點：需要產業知識、只會做保守的推薦。

### 內容過濾就是以物件內容或屬性為基礎來代替每個物件間建立屬性向量，在判斷物件是否相關時，常以計算物件向量間的餘弦值，也就是以兩向量間的夾角來代表相似度，這兩向量間的夾角越小，代表這兩個物件的相關性越高，反之則相關性越低。找出商品跟商品之間的關聯，把商品做分類，若以顧客和商品間的關係來看，我們可以把顧客和商品分別用一物件表示，兩個物件的向量夾角大小，就是顧客對商品可能的喜好度。這種方式不需要有用戶資料也可以進行推薦，適合推薦陌生用戶使用，也可以針對使用者以前喜歡的商品，推薦類似的商品給使用者。

### 

### 協同過濾：

### 優點：不需要產業知識、可推薦使用者偏好類型外的產品、只需交易紀錄，易導入。

### 缺點：冷啟動、稀疏矩陣、未考慮商品屬性資料

### 協同過濾，顧名思義就是集合眾人的意見協同合作，進而推薦或篩選物品，一樣是以銷售記錄進行分析，不同的是，並不進行商品組合分析，而是將銷售記錄轉成『使用者/商品對應的矩陣』(User-Item Matrix)，如下圖，記錄哪些使用者買過哪些商品，計算顧客間或商品間的相似度，再推薦相似顧客曾買過的商品，或推薦與目前商品最相似的其他商品，進行 cross selling。

而一般網購的顧客、會員或潛在客戶很多，如果使用 USER-USER 協同過濾，USER-USER Similarity Matrix 就會非常大，計算非常耗時，這時就可以採用 ITEM-ITEM 協同過濾計算比較快。反之，如果是屬於內容為主的網站，例如新聞、部落格，則因商品是新聞或業配文，文章推陳出新，亦即商品很多，就適合使用 USER-USER 協同過濾法。此外，USER-USER 協同過濾法是推薦最多使用者購買的商品，所以，常會推薦熱銷品，相對的，ITEM-ITEM 協同過濾就比較容易推薦到長尾(long tail)的商品(銷量不大，但一直持續有人購買)，這對公司整體營收會有較良性的助益。

## 混合推薦（Hybrid Recommendation）：

## Cold Start (冷啟動)的問題：

## 在推薦系統中，冷啟動的問題是指「沒有充足的訊息」而造成推薦效果不佳。以電商平台來說，冷啟動的問題可以歸類成兩個面向：產品與使用者。新的電商平台可能沒有足夠的產品上架或是會員資料，導致推薦系統的訊息不足，無法有效運作，進而影響推薦效果。

## 此外冷啟動在協同過濾系統中一直是很大的問題。對於曝光度低的冷門產品(long-tail item)，無論是哪一種協同過濾系統(User-based / Item-based)，都不會把商品推薦給使用者，使得那個產品會一直處在冷門的狀態。

## Data sparsity 資料稀疏的問題：

## 當電商平台的資訊量過多的時候，也可能造成所謂「資料稀疏」的問題。舉例來說，如果平台上的某個品類有太多相似的產品，當使用者瀏覽頁面的時候無法看過大部分產品，就會造成蒐集資料上的稀疏。資料的規模越大，資料稀疏的問題就越嚴重，這也是目前許多學者們們想解決的問題。

## 「混合推薦」顧名思義就是結合上述的內容推薦與協同過濾等方法來建立推薦系統，以增進推薦的效率。由於資料的複雜程度，以及每種商品都有其適合的推薦方式，業界許多推薦系統的應用都會混合不同的方法，甚至結合機器學習演算法來做推薦系統。

## 章節三 系統設計與實作

### 資料集：

資料集由博客來網站中文書新書依照暢銷度往下爬蟲，總共收集到3872筆資料，包含“文學小說”、“商業理財”、“藝術設計”、“人文社科”、“心理勵志”、“宗教命理”、“自然科普”、“醫療保健”、“飲食”、“生活風格”、“旅遊”、“童書/青少年文學”、“國中小參考書”、“親子教養”、“影視偶像”、“輕小說”、“漫畫/圖文書”、“語言學習”、“考試用書”、“電腦資訊”、“專業/教科書/政府出版品“共21種分類中文新書，欄位包括“分類”，“書名”,“圖片網址”,“作者”,“出版社”,“出版日期”,“優惠價”，“內容”,“商品連結”，“也買了”共10種，為什麼需要“也買了”這一欄呢？因為它相當於顧客購買資料。

### 資料集完成時間：2022年6月14日

### 資料處理：

### 本次爬蟲取得之資料須符合python編碼規範，需將所有資料檔案進行編碼，本研究將原本格式轉成utf-8 csv檔，以利後續操作。取得的欄位當中“圖片網址”與“商品連結”是絕對用不到的所以就直接把他們去掉，在爬蟲的時候透過商品連結取得“也買了”個欄位，但當初爬取時是把目標站都放入矩陣然後轉成文字，所以那個項目裡的資料會有不餘的符號如“,‘[]，需要一併去除，以免影像到後續處理商品名稱發現明明是同樣的商品，卻因為多了符號而被誤判的情形發生，最後透過迴圈將“書名”與“也買了”中各個商品做對應，製作成二維矩陣，以利後續演算法執行。

### 演算法：

Apriori重要假設: Apriori是計算頻繁項集的一種演算法，它假設當項集是頻繁的，也就是假設B這個物品在數據中是頻繁出現的，那它的子集也會是頻繁的，也就是說{B、C}、{B、C、E}等也是頻繁的，反之就是不頻繁的。

一張含有 桌 的圖片

自動產生的描述

### <https://chwang12341.medium.com/machine-learning-關聯分析-apriori演算法-詳細解說啤酒與尿布的背後原理-python實作-scikit-learn一步一步教學-76b7778f8f34>

Step1: 上圖由左至右進行疊代，有了最左邊的數據庫後，接下來進行第一次掃描，可以看出每樣商品出現的次數，由於我們自行設定的最小支持度為50%，所以如果次數小於3(總數據量6 x 50%)，就會被當成是不頻繁項集淘汰，也就是圖中的D商品，為什麼會被淘汰?因為前面有提到Apriori的重要假設，當項目集是不頻繁的，那它的子集也不會頻繁，接下來就形成了新的數據表。

Step2: 接下來進行我們的第二次掃描，也就是項目集中會有兩樣商品來做分析，一樣列出出現次數後，淘汰掉次數低於3的項目集，形成新的數據表。

Step3: 最後進行我們的第三次掃描，項目集中會有三樣商品來做分析，列出出現次數，由於只有一組項目集，分析過程也就完成囉，最後一樣會形成一個關聯數據表。

### 資料分析：

### 由於資料量大且重複性較低所以設定最小支持度=0.0001,最小信心水準=0.0001,最小提升度=3,經過演算法計算，最後打印出商品關係對以及其支持度、信心水準與提升度。

### 章節四 相關就業市場

### 相關企業：麻布數據科技股份有限公司

### 公司網址：<https://ezprice.com.tw/>

### 資本額：1億

### 員工人數：50人

### 職務名稱：Machine Learning Engineer (NLP / Recommendation System)

### 工作待遇：待遇面議（經常性薪資達 4 萬元或以上）

### 技術需求：

1. 精熟自然語言處理、機器學習模型、Data Mining 理論和方法，並有一定的相關研究或實務經驗。

2. 實作過但不限於以下 NLP-related task： Word Segmentation、Text Classification、Named Entity Recognition、Relation Extraction、Recommendation System、Knowledge Graph。

3. 二年以上的深度學習框架實作經驗 (如：Pytorch、Keras、Tensorflow)，並實際應用於商業案例。

4. 二年以上的 BigData Ecosystem Operation / Management 經驗。

5. 熟悉分析產品需求，將 modeling 過程中的痛點轉成工程計畫。

6. 熱愛思考商業需求，將演算法動手打造成產品。

7. 喜歡追求高品質的程式，面對需求能夠提出最小可行性系統架構及了解其中的 tradeoff。

職務內容：

1. 這個職位主要負責巨量發票數據分析， ML 模型及系統建置，提升資料商轉應用的效率。 2. 視業務和專案需求，引入適合的系統及技術，並規劃開發流程，包含：

＊ 規劃機器學習架構流程 (包含演算法訓練，實際上線所需的系統建置)。

＊ 建置機器學習專案系統 (API / 計算流程)。

＊ 規劃 Data Pipeline 開發流程與架構，數據資料前處理規劃 (如：斷詞、標記等)。

＊ 快速交付可行性架構跟改進數據產品系統相關計畫。

3. 內部環境自由，長期可能碰觸到的內容：

(a) 參與數據分析產品、資料分類系統、廣告平台、推薦系統等相關專案，有機會接觸到的技術：

＊ Google Cloud Platform

＊ Containers & Serverless (Kubernetes, GAE ...)

＊ Redis ＊ Firestore ＊ Data Warehouse (BigQuery, ClickHouse)

＊ Pub/Sub

＊ Cloud build

(b) 和其他 Data Component 團隊合作，開發機器學習相關系統建置。

### 相關企業： 東森新媒體控股股份有限公司

### 公司網址： http://www.ettoday.net/

### 資本額：10億

### 員工人數：650人

職務名稱：資料工程師 Data Engineer

### 工作待遇：待遇面議（經常性薪資達 4 萬元或以上）

基本條件：

1. 具有三年以上的軟體（包含演算法與資料結構）的開發經驗

2. 利用 SQL, Python, R 來建立資料模型

3. 擁有 Linux 作業系統的基本操作經驗

4. 善於解決問題，獨立思考且結果導向，有優秀的動手作能力。喜歡自動化、Code Review、Design Review

5. 基於雲端解決方案，規劃技術架構並建立大規模 ETL 流程的經驗

加分條件：

1. 根據資料講故事及表達的能力

2. 有作過推薦系統的經驗

3. 有作過搜尋引擎的經驗

4. 善於在跨職能小團隊中團隊合作與溝通

職務內容：

1. 與PM/UI/UX、前後端、資料、演算法工程師一同合作，開發與維護品。

2. 負責處理及挑戰 Data 相關的工程問題，包含建立 data pipeline、資料建模和資料分析等

3. 與數據團隊協同合作，打造基於機器學習、資料探勘與人工智慧等相關技術的數據服務（例如 Recommendation system, Search engine, Data warehouse 等），並應用到 APP 上。

相關企業：雄獅資訊科技股份有限公司

資本額：1800萬

員工人數：206人

公司網址：<http://www.liontech.tw>

職務名稱：python工程師

工作待遇：待遇面議 （經常性薪資達 4 萬元或以上）

職務需求：

1.熟悉 Flask 框架

2.熟悉 Git 開發管理專案

3.熟悉 SQL 及 MongoDB 資料庫

4.熟悉 ML 相關知識

5.熟悉Python程式語言，包含Pandas, Numpy, Sklearn,Tensorflow, Pytorch等模組

工作內容：

1. 負責公司之 Data Science API維護和新功能開發, 協助產品成效優化及發展方向

2. 負責資料收集、資料清理、模型建置等工作事項

3. 商品推薦系統建置及優化

4. 系統測試代碼實現，自動化測試

5. 參與.討論.協助專案架構設計

### 章節五 報告心得

### 心得分享：

### 在資料爬蟲上吃了不少苦頭，尤其遇到要爬取的內容當中有項目是以AJAX的方式呈現，還有博客來伺服器端的反爬蟲機制，明明status code已經回傳200卻顯示空的內容,（即使header已經設定）,最後還是得開webdriver等待網頁載入完成後，才能拿到資料，不過有些神奇的事明明已經指定標籤縮小範圍，卻還是能捕捉到預計目標以外東西，且有些資料似乎在載入的時候也不順利，這時候就填入nan。博客萊在書籍分類網址設計上有一定的規律，但是也有一些例外，有些例外好處理，有些例外卻無法處理，例如18+的商品、書籍他會要求先登入會員驗證，此類商品的推薦商品也直接做nan處理。收集資料階段耗時非常久，而且商品推薦欄位的資料大概也接近30%資料是沒有收集到的，原因可能是被網站擋了下來或者是出現了例外的錯誤與不同的設計等等。

### 製作完資料之後就是要選擇如何分析？用什麼方法做推薦系統？但這邊的難點是我所掌握到的資料僅有網上看得見的，幾乎所有公司的銷售資料、顧客資料是無法獲取的，當然這是保障顧客的資訊安全，這些資料也是公司寶貴資產的一部分，本來想說身為人工智慧組的我，應該要做一些model-based的分析推薦，但考量到手中的資料量以及監督式學習需要標準答案，最後決定使用關聯分析的apriori演算法。

### apriori的優點是使用上非常簡單，只需要把資料換成二維矩陣就可導入了，而些數據中只需要有關聯的數據即可其他東西用不到，但也有他的缺點就是每個商品她都必須計算他與其他商品的相關性，所以若是資料龐大，那麼運算效率就會明顯降低，透過Apriori我們可以看出哪些商品有更高的相關性，倘若是要結合銷售，那麼綁在一起賣將有助於提高銷售額！

### 未來展望：

### 後續倘若要做相關的研究，可於提高推薦精準度與擴充推薦物品種類下手，就以市面上常看到的推薦系統來看大多都是推薦熱門的、接觸率高的，很多基於物品的推薦系統是很難分類的，因為內容特徵並不明確例如：聲音、影片，所以有效找出目標特徵也是一項重要的課題，而目前最好的推薦系統是將使用者資料與商品資料混合使用，將使用者分群，將物品分群，搭配機器學習、深度學習能獲得更好地效果。

## 章節六 參考文獻

Agrawal & Srikant （1994）《Fast Algorithms for Mining Association Rules》

### <http://www.vldb.org/conf/1994/P487.PDF>

### 余明哲 2022 《圖書館個人化館藏推薦系統》

<https://ir.nctu.edu.tw/handle/11536/70200>

唐瑤瑤 許凱平 張傑生 《基於關聯法則建置之臺灣大學電子報推薦系統》

### <http://itech.ntcu.edu.tw/tanet%202007/2/61.pdf>

### <https://ithelp.ithome.com.tw/articles/10220962>

### <https://chwang12341.medium.com/machine-learning-關聯分析-apriori演算法-詳細解說啤酒與尿布的背後原理-python實作-scikit-learn一步一步教學-76b7778f8f34>

### <https://aitmr1234567890.medium.com/%E8%B3%BC%E7%89%A9%E7%B1%83%E5%88%86%E6%9E%90-python%E5%AF%A6%E6%88%B0-%E5%A6%82%E4%BD%95%E6%89%BE%E5%87%BA%E5%95%86%E5%93%81%E6%90%AD%E9%85%8D%E7%9A%84%E7%B8%BD%E9%AB%94%E7%AD%96%E7%95%A5-%E9%99%84python%E7%A8%8B%E5%BC%8F%E7%A2%BC-d4585e951395>

### <https://www.maxlist.xyz/2018/11/03/python_apriori/>

### <https://medium.com/@winnie54liu0504/%E6%8E%A8%E8%96%A6%E7%B3%BB%E7%B5%B1%E6%98%AF%E4%BB%80%E9%BA%BC-what-is-recommender-system-445abf82795>

### <https://medium.com/edureka/apriori-algorithm-d7cc648d4f1e>

### <https://www.104.com.tw/job/7e2a0?jobsource=jolist_b_relevance>

### <https://www.104.com.tw/company/1a2x6bic2h?jobsource=jolist_b_relevance>

### <https://www.104.com.tw/job/7mumm?jobsource=jolist_b_relevance>

### <https://www.104.com.tw/company/b6ctt6w?jobsource=jolist_b_relevance>

### <https://www.104.com.tw/job/71rxc?jobsource=jolist_a_relevance>

### <https://www.104.com.tw/company/13kq7dpk?jobsource=jolist_a_relevance>