IR Term Project Report
Dcard 使用者配對模型

資管四 B99705011 林裕欽　　資管四 B99705018 孫至緣　　資管四 B99705037 劉盈秀

1. INTRODUCTION

近年來社群相關網站蓬勃發展，facebook、twitter 等的崛起更是其中的代表，我們逐漸習慣透過網路、電腦，與其他人間有更進一步的連結，與人相識、聊天、交往，藉由網路各種平台建立自己的交友圈，也藉由網路平台來維繫朋友間的情感。在網路上的社交活動已然是生活中的大宗，是許多人不可或缺的管道，如此也更相應產生許多的社群交友平台，以滿足眾多使用者的需求。

Dcard 是一個校園社群平台，以學校信箱為帳號，以學生為使用族群，讓校園內許多才華洋溢但在不同生活圈裡的同學，可以在 Dcard 認識。但 Dcard 不同於一般的社群交友網站，其交友規則較為特殊，每天午夜 Dcard 會將使用者配對，每個使用者會收到一封另外一位使用者的自我介紹資料，資料內包含就讀學校、系所、專長、興趣、社團、喜歡的國家等等，而在看過對方的自我介紹後，使用者可以選擇是否要發送邀請給對方。在 24 小時內，雙方皆對彼此發送交友邀請，他們才能成為朋友，並獲得對方的真實姓名、手機、感情狀態等相關更詳細的個人資訊，也才能進行更一步的交流。如果當天沒有成為朋友，下個午夜來臨後對方資料就會消失，不會知道對方有沒有對自己按下成為朋友。大家可以在 Dcard 敘述自己的專長、興趣、可以分享的事物、經驗；透過成為朋友後的聊天交流，這些累積可以讓人真的認識到朋友，真正擴展交友圈。

因觀察到近期 Dcard 使用者人數正逐漸上升，又其新增好友的模式有別於一般的社群網站，十分特殊，我們對於使用者彼此間邀請與否、是否成為好友的選擇模式十分好奇。例如某個使用者 A 會偏好對女生送出邀請，又對於興趣包含音樂領域的人，送出邀請的機率相對較高。

所以我們配合 Dcard 的後台系統，收集 Dcard 內所有使用者的自我介紹內容，從內容抓出關鍵字並加以分類，來建立使用者的群組，再進一步將同一分群內的使用者配對，產生每天使用者收到的新 Dcard，每當有新的使用者加入，也會持續進入各分群。透過我們的 Dcard 使用者配對模型，來增進使用者的配對成功率，讓使用者更容易於 Dcard 上結交到興趣相投的朋友，透過 Dcard，進一步迅速累積自己的交友圈。

<Dcard 網站介面>

2. Related work

本研究主要要先取得使用者的自介內容，並用中文斷詞系統 CKIP 先將每個使用者的文章 normalize 並取的關鍵字，接著我們使用 k-means 將使用者分群，最後我們再用 SVM 將使用者分類。

2.1 CKIP 中文斷詞系統

自然語言處理系統最基本需要讓電腦能夠分辨文本中字詞的意義，才能夠更進一步發展出自然語言處理系統的相關演算法，其中斷詞處理便是一個重要的前置技術， CKIP 中文斷詞服務網站使用了中研院斷詞系統的 Client 端程式，讓有中文斷詞需求的研究者或程式人員可以專注於開發自己的核心演算法。

CKIP 中文斷詞系統[1] 是一個由中研院開發的工具，功能是可以將一整篇的中文文章，把每個詞與字都斷開。基本上自動分詞多利用詞典中收錄的詞和文本做比對，找出可能包含的詞，由於存在歧義的切分結果，因此多數的中文分詞程式多討論如何解決分詞歧義的問題，而較少討論如何處理詞典中未收錄的詞出現的問題（新詞如何辨認）。

此系統可以自動抽取新詞建立領域用詞或線上即時分詞功能，為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統。範例：輸入句子 “我們都喜歡蝴蝶”，斷詞與詞性標記輸出結果為“我們(Nh) 都(D) 喜歡(VK) 蝴蝶(Na)” (詞與詞中間為全形空白)。CKIP 斷詞系統的內部處理採用中研院所編列的簡化詞類，而線上斷詞服務採用之精簡詞類為再簡化的標記，因此最後結果變為“我們(N) 都(ADV) 喜歡(Vt) 蝴蝶(N)”。

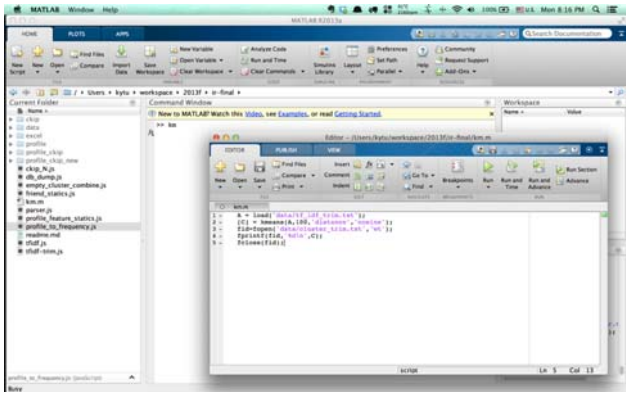
分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。除了基本詞彙庫外，使用者可依需要附加領域專屬詞庫。詞類標記為選擇性功能，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類。分詞系統採用之詞典俱可擴充性，使用者可依據不同領域文件，補充以領域詞典做為分詞之用。

2.2 K-means

k-means 是一種分割式群集法，其分割演算法主要目的是將資料分為由使用者指定的 k 個群集，由使用者決定要將資料分為多少群集。

演算法首先隨機地選擇 k 個 document，每個 document 初始地代表了一個群集的平均值或中心。對剩餘的每個 document 根據其與各個群集中心的距離，將它賦給最近的群集，然後重新計算每個群集的平均值。這個過程不斷重複，直到準則函數收斂。

而我們系統中所用到的 k-means 則是採用 Matlab 的軟體來進行實作，版本為 R2013a，其函式 IDX = kmeans(X,k)，輸入資料矩陣 X，以及要分成的個群集數量 k，即可得到 IDX 的分群。

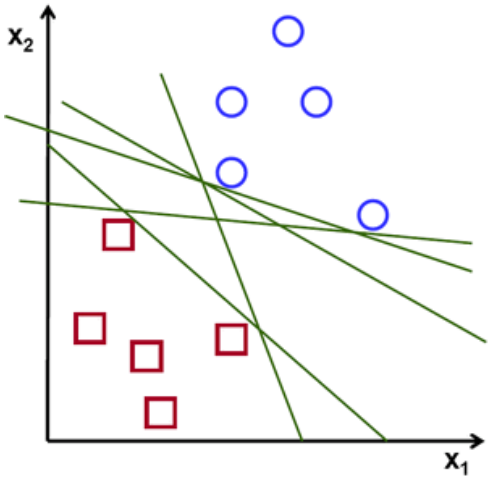


2.3 SVM

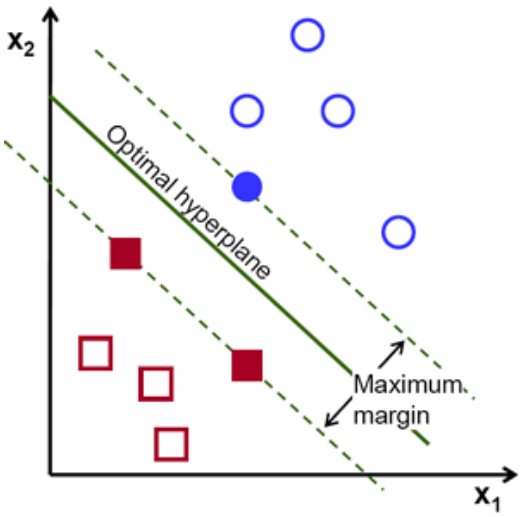
Support Vector Machine，常簡稱為 SVM，是一種監督式學習(supervised learning)的方法，可廣泛地應用於統計分類以及回歸分

析，一般是應用於分類(Classification)等相關議題上。SVM 基本運作模式如下：在給定一群訓練樣本之下，每個樣本會分別對應至兩個不同的類別(Category)，SVM 會嘗試從建構一個模型(Model)，並利用此模型將每一個樣本分配到一個類別上。

典型的 SVM 是一種二元分類器(Two-class classifier)，因此先針對典型 SVM 來進行說明:



在二元分類中，SVM 嘗試在訓練資料(training data)所構成的空間中，尋找一個超平面(hyperplane)能將不同類別的資料完美的分開，而且，希望此超平面與不同的類別的距離愈大愈好。如圖所示，藍色圓形為第一個類別(標記為+1)，紅色方形為第二個類別(標記為-1)，而 SVM 則想要找出的超平面即是 wx+b=0，此超平面可以使得兩個類別(class)的距離最大。



3. Methodology

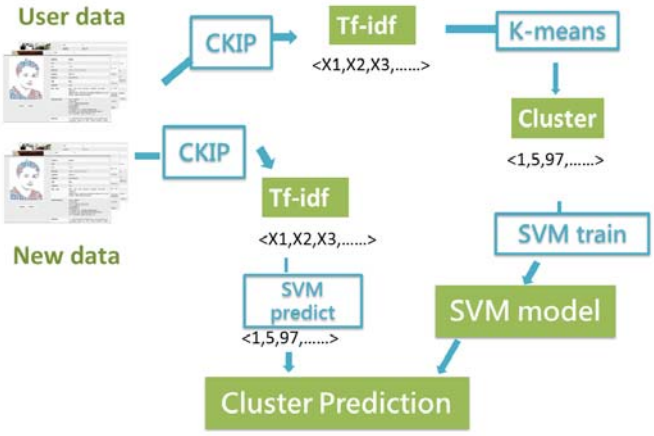
3.1 System Overview

本系統為配合 Dcard 的管理後台，利用使用者的自我介紹內容建立使用者分群，成為系統的配對模型，並進一步透過系統來為使用者配對，以增進 Dcard 使用間的配對成功率。

我們系統分成兩大部分，一部分是利用所有使用者的自我介紹資料，來建立使用者的分類群組，也就是使用者基本的配對模型；另一部分為當 Dcard 有新使用者，將其加入已經建立好的分類組別，來持續改進配對模型。每天皆可於當前的選擇模型下，產生新的配對組合。

關於建立基本使用者配對模型部分，我們先從 Dcard 後台獲取所有使用者在交友成立之前，會被看到的自我介紹資料。將這些資料透過 CKIP 中文斷詞系統來進行斷詞，經過濾掉 stopwords 並計算其 tf-idf 值，再利用 matlab 的 K-means 方法，將這些自我介紹分成 K 組(K = 20 & 100)。利用這些分好的 cluster 資料進行 SVM train，完成 SVM model，便可利用此 SVM model 進行預測，把同一群組的使用者利用 DFS 來進行配對。

但因以上的 SVM model 建立的過程耗時，需要花很多時間在處理既有使用者的自我介紹資料，所以在 SVM model 建好後，新使用者加入時我們並不會重跑一次整個流程。每當有新使用者加入 Dcard，一樣會先將該使用者的自我介紹資料送入 CKIP 進行斷詞，移除 stopwords 並計算 tf-idf，之後便直接利用已有的 SVM model 進行 SVM predict，將新使用者歸納進入適當的分群，以便繼續進行下一次午夜的使用者配對。



<System Overview>

3.2 CKIP 字詞處理

一開始我們從 Dcard 中拿到 14,000 多筆使用者的自我介紹資料，資料內包含 5,000,000 多字。

專長、興趣	魔術、手語、排球、程式語言、網頁製作、影片後製、上台報告:目 喜歡參與活動，總是有許多奇妙的點子!覺得自己的人生相當充實，喜歡把想到的東西實踐!
曾參加過的社團	台南一中魔術社 台大南友會 台大不一樣思考社(設計思考) 台大管理顧問社 創意創業學程 台大國商營（第一次接觸行銷還蠻有趣的） 外交部青年大使（最近被戰很兇但是個好活動XD） ATCC 全國季軍（很累人的比賽 Q_Q）
喜歡的課	大一的國文跟英文課~學到很多很多人生的大道理!國文課上的是史記，頗析人性、局勢，中國史上的鉅著！也很喜

<使用者自我介紹範例>

將這些資料送入 CKIP 斷詞，中途我們也嘗試了其他的斷詞系統，但還是 CKIP 的結果最精準。另外，CKIP 會將結果分為動詞、名詞等等詞性，而因為我們觀察後，認為其中名詞較其他詞性具代表性，例如「籃球」、「吉他」會比「打」 「彈」更清楚代表個人特色。所以我們只採用斷詞後的名詞，共 15913 詞作為資料繼續處理。

去(ADV) 觀摩(Vt) 一下(N) 顆顆(DET)
短暫(Vi) 的(T) 籃球(N) 經理(N)
很少(ADV) 去(Vt) 的(T) 羽球社(N)
學生會(N)

<CKIP 部分斷詞結果>

雖然 CKIP 結果相較其他斷詞系統精確許多，但還是包括不少不具意義或對群組分類可能沒有幫助的詞，例如「事情」、「恩」、「隔天」，其中也有出現頻率很高的字詞，例如「我」對於使用者分類較沒有顯著幫助，但卻是第一多的字詞。所以我們將這 15913 字詞中，先刪除出現次數低於 4 次的，留下 4147 個詞，再用人工的方式來做 feature selection，刪除無意義的字詞，最後留下 2277 個詞。

3.3 Cluster 分群與建立 SVM model

再來計算這些自我介紹的 tf-idf，然後利用 matlab 的 k-mean 進行分群。一開始嘗試分成 20 組，後面步驟拿 20 群下去測試也會有不錯的使用者配對成功率提升，但將每群的前 10 個關鍵字印出，我們卻發現很多關鍵字會混雜在一起，分群結果不夠細緻，例如有一類可能同時包含「系隊」、「排球」、「壘球」、「羽球」等。所以我們將 K 從 20 調整成 100 群，從各群前 10 關鍵字來看也有滿明顯的特色。

2	環島	台灣	日本	風景
3	羽球社	羽球	校隊	音樂
4	廣東話	港澳	港劇	追星
5	滑翔翼	高空彈跳	風帆	降落傘
6	攝影社	攝影	相機	日本
7	論文	碩士	電影	運動
8	天文社	天文	星星	觀星
9	學生會	活動	電影	福爾摩沙社
10	澳洲	紐西蘭	日本	加拿大
11	法國	義大利	英國	德國
12	口琴社	口琴	帝國	國二
13	歷史	社會	中國	文化
14	程式	設計	網頁	遊戲
15	自然	保育社	科學	研究社
16	熱音	樂團	吉他	音樂
17	吉他社	吉他	音樂	電影
18	壘球	系隊	棒球	系壘
19	攝影	底片	製片	電影
20	音樂社	西洋	音樂	樂團

<K =100 部分分群關鍵字(以列為一群)>

在 cluster 分類後，即可利用 SVM 來進行 train。如上述所提到，最基本的 SVM 需採用 expert 分好的 training data 進行 train 後，完成 SVM model，而我們利用 cluster 分類資料作為 training data 來進行 SVM train，來完成我們的使用者自我介紹 SVM model。

最後，用此來進行 Cluster Prediction，將同一組別內的使用者兩兩配對，採用 DFS 的方式進行配對，此外加入的篩選條件包括有：

- 配對過的不能再進行配對
- 把組別內為奇數個的抓出來湊成一對。
- 若最後還剩奇數個，則管理員不配對。

如此便可完成利用使用者的自我介紹，來進行配對。

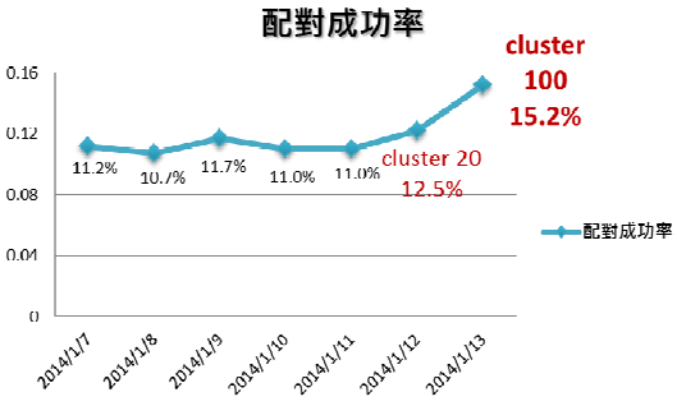
3.4 New Data 的處理

因在將所有使用者資料送入 CKIP 的處理十分耗時，第一次使用時約需 6 個小時光處理斷詞工作，所以我們為提升系統效率，而對於新加入的使用者有不同的處理步驟。

每當新加入一名使用者，同樣會送入 CKIP 進行斷詞，並計算其自我介紹的 tf-idf 值。因 SVM 可透過已知的 model 進行未來預測，所以我們利用舊有 data 已經產生的使用者分類 model，來預測此新的使用者是屬於哪一組別，並將之歸入該組。如此一來除了還是能擁有不錯的分類成果，兼併有適合實際應用的效率。

4. Conclusions

4.1 實驗結果



我們目前拿兩天的配對及送出邀請比率來做實驗，當我們分成 20 組的時候，當天的送出邀請比例為 12.5%，能看出有稍高於平日；但當分的更細，將 Dcard 的使用者分成 100 個 clusters 後，第二天的的配對率竟然有 15.3%，和平日的配對率有顯著的差別！

4.2 結論

從這兩天的實驗結果顯示，將使用者的自介分群後，對同個群的人進行配對有助於提升配對成功率。這是因為有部分人們較容易對和自己有共同興趣的人有好感；但同時，肯定也有人並不想只認識有共同興趣的人，而是想要認識各式各樣，與自己有所差異的人。因此這樣的做法只是短期的實驗階段，未來在做了更多的實驗有了更明確的結論後才有可能會在真實系統中完全套用新的配對演算法(目前配對為完全隨機)。

另外，我們的實驗也指出，將使用者分為 100 群的效果將比 20 群顯著，這是因為分成 20 群時，其關鍵字的分類還不夠細緻的緣故；而當我們將使用者分成 100 群後，其各群的差異就更明顯了，每群的特色也更為鮮明，配對成功率也顯著的提升。

透過這個系統，Dcard 又讓世界多了一千份友誼！

4.3 未來展望

1. 潛在語意分析 (Latent Semantic Analysis)

   因為使用者的自介其實也是短篇的文章，未來可以考慮實作潛在的語意分析，分析使用者的自介所包含的語意。

2. 文字分類

   由於目前的文字還是全部混在一起的，像是國家、運動、社團等類別的文字還無法明確的分類，未來若是能實作這一部份，相信在使用者分群上一定也會更為精細。

3. 參照過去配對記錄

根據個別使用者以往送出邀請的偏好，不再只是同類相吸。若每個使用者都有其差異化的配對模型，相信一定能更大幅度的提高配對成功率。

5. Member's Workload

字詞處理：林裕欽、孫至緣、劉盈秀

Clustering, SVM modeling：林裕欽

Powerpoints and Presentation： 林裕欽

Term Project Report：孫至緣、劉盈秀

6. References

[1] CKIP 中文斷詞系統

http://ckipsvr.iis.sinica.edu.tw/

[2] Support Vector Machines Intro, 林宗勳

http://www.cmlab.csie.ntu.edu.tw/~cyy/learning/tutorials/SVM2.pdf

[3] OpenCV SVM 實驗

http://cg2010studio.wordpress.com/2012/11/12/opencv-svm-%E5%AF%A6%E9%A9%97-2/#more-7778

[4] http://tm.itc.ntnu.edu.tw/CNLP/?q=node/6

[5] K-means clustering, Matlab

http://www.mathworks.com/help/stats/kmeans.html