Lecture 2 : Web Mining

楊立偉教授

wyang@ntu.edu.tw

本投影片修改自Introduction to Information Retrieval一書之投影片
Ch 20~21

Co-occurrence and Association

基本原理：共現分析

Database 結帳紀錄

TID	Itemset
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

L_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

Support ≧ 2

L_2

Itemset	Support
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

L_3

Itemset	Support
{2 3 5}	2

最常被一起購買的產品組合
組合數為2時 : {2,5}
組合數為3時 : {2,3,5}

wyang@ntu.edu.tw

關聯規則 Association rules

- 尋找每筆交易中被同時購買之商品的關聯性

  Buy (milk) → Buy (bread)　　信心度 80 %

- 尋找消費者與商品之間關聯性

  iPhone7 Plus → 男性、上班族、年收入80-120萬　　信心度 60 %

- 亦可尋找任何人、事、物彼此間同時出現之關聯性

  Ex. 找文件中重要字詞之間的關聯性

  或用出現字詞找文件之間的關聯性

wyang@ntu.edu.tw

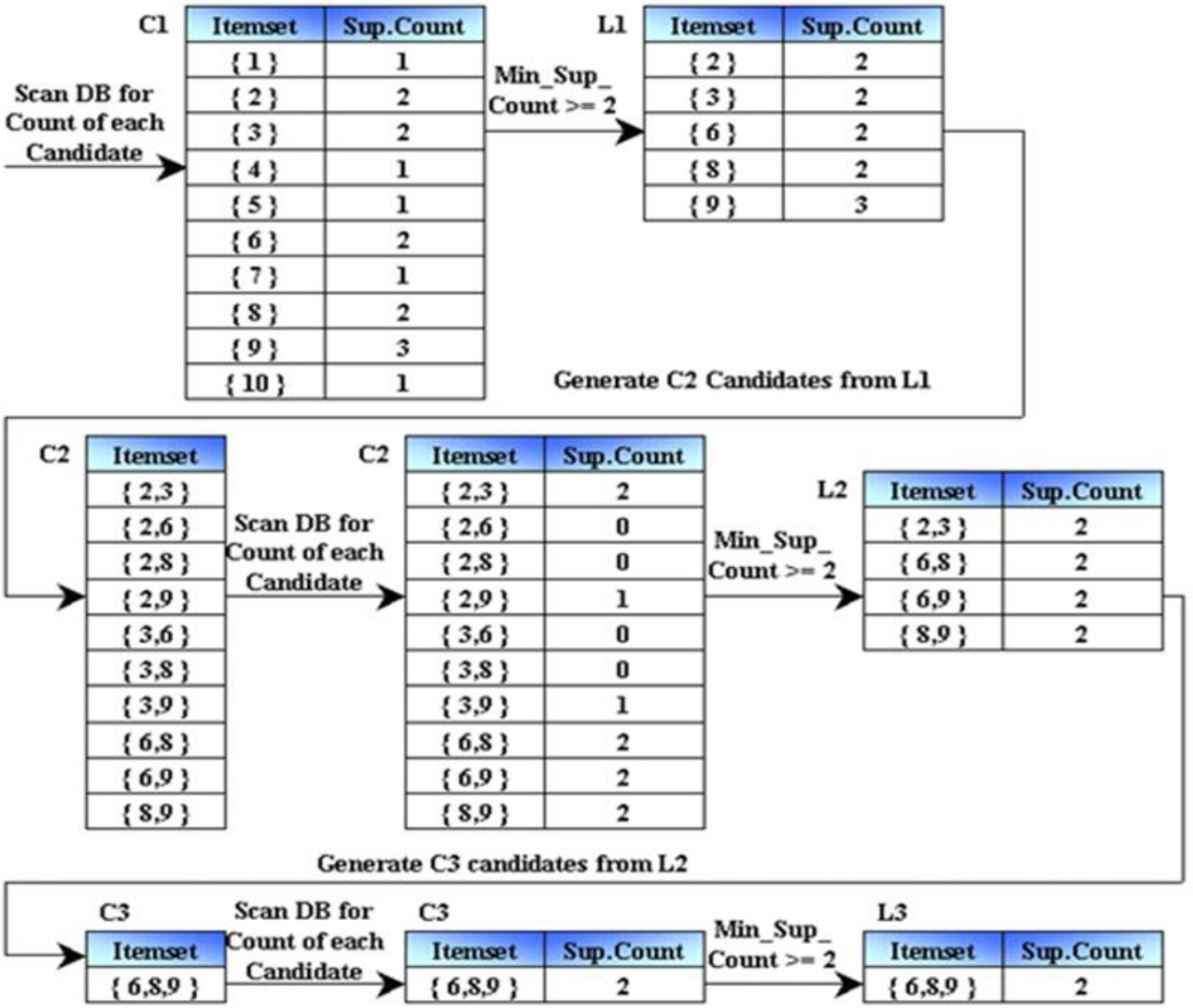
常見有：

Apriori

演算法、

FP growth

演算法



wyang@ntu.edu.tw

- FP-growth 演算法
  - Han, Jiawei, et al. "Mining frequent patterns without candidate generation: A frequent-pattern tree approach." Data mining and knowledge discovery 8.1 (2004): 53-87.
  - Some slides from Internet

wyang@ntu.edu.tw

關聯規則 Association rules (續)

- 檢驗方式

  若 X→Y

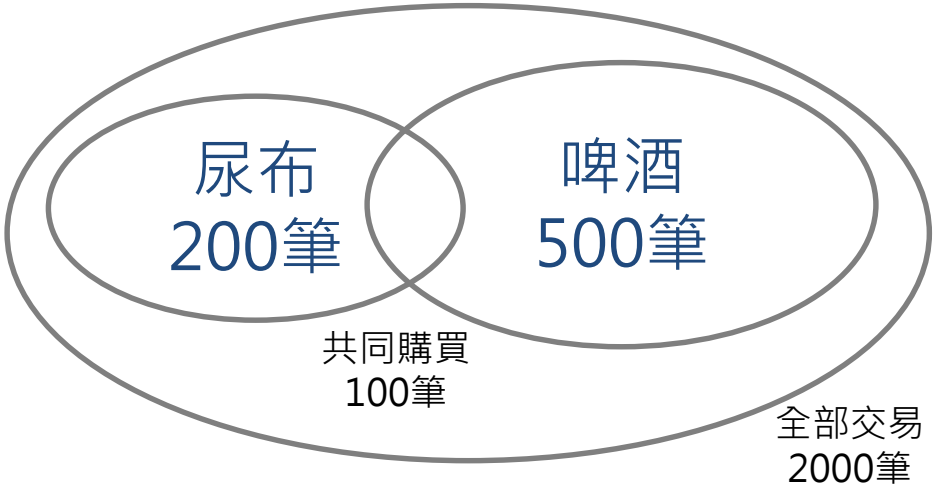
  支持度 Support = P(X∩Y) = 包含X及Y的筆數 / 總交易筆數

  信心度 Confidence = P(Y | X) = 包含X及Y的筆數 / 包含X的筆數

  提升度 Lift = P(Y | X) / P(Y) = 信心度 / (包含Y的筆數 / 總交易筆數)

  三者代表不同意義，越高實用價值越大

wyang@ntu.edu.tw



Association 檢驗方式

尿布→啤酒

支持度 Support = 100/2000=0.05

代表是否重要

信心度 Confidence = 100/200=50%

代表是否準確

提升度 Lift = (100/200) / (500/2000) = 200%

代表是否特別

wyang@ntu.edu.tw

Lift的說明

- **假設2000筆資料中，500筆有買啤酒(4個中有1個)；今200筆買尿布的資料中，若有接近50筆有買啤酒 (4個中也有1個)，則接近於原本之機率(密度)，故稱找到的是"common sense"。**
- **笑話一則**
  - 「報告，經過分析，發現有鼻子的人都有眼睛，confidence和support都是1，超準、超重要的」，「孩子，這是廢話」
  - 以上敘述，重點在於要知道有眼睛的機率為何

wyang@ntu.edu.tw

綜合案例 – 台灣最大實體書店

- 台灣地區大型書籍零售賣場領導品牌，擁有數十萬會員資料，每年會員交易紀錄超過數百萬筆
- 分析目標
  - 目標 1 : 尋找會員購買商品之間的關聯性
  - 目標 2 : 尋找會員基本資料、與購買商品之間的關聯性
- 樣本資料
  - 20萬筆會員資料
  - 10萬筆行銷活動收集之名單
  - 二年度的會員交易資料明細

wyang@ntu.edu.tw

綜合案例 – 台灣最大實體書店 (續)

- 針對目標 1，使用關聯分析 (Association) 模組，自動尋找出最具關聯性的購買商品
- 發現：
  - 購買 休閒娛樂 類商品的會員, 同時會再購買 旅遊 類商品
  - 購買 乾隆相關 書籍商品的會員, 同時會再購買 雍正王朝 DVD
- 意義：
  - 可以針對上述具高度關聯性的商品進行搭售與聯合促銷
  - 可以寄送另一商品之促銷訊息予只購買單一商品之會員
  - 賣場動線設計 : 具高度相關之商品應陳列在同一鄰近區域

wyang@ntu.edu.tw

綜合案例 – 台灣最大實體書店 (續)

- 針對目標 2，使用主力客群 (Clustering) 模組，自動尋找出會員資料中與商品特性關聯性最高的欄位
- 發現：
  - 旅遊 類商品與會員資料中的 性別 與 年齡 欄位有高關聯性
    - 顯著區間：(Female, 30~40)
  - 財經 類商品與會員資料中的 職業 與 收入水準 欄位有高關聯性
    - 顯著區間：(Employee, 500K~800K yearly)
- 意義：
  - Direct Marketing : 可以將促銷商品 DM 只寄給最具關聯性的潛在客戶。可大幅降低行銷成本，並提高回應率與成交率

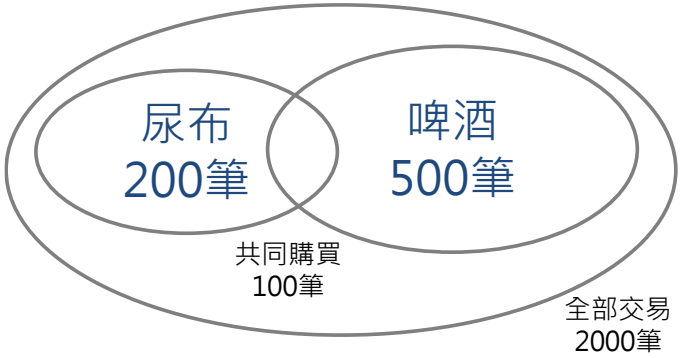
wyang@ntu.edu.tw

綜合案例 – 台灣最大實體書店 (續)

- 專案導入：
  - 該專案執行期間, 由總經理指派專案小組負責
  - 部份商品為少量多樣，如建築類、藝術類、國外進口書等
  - 原先每年寄送的會員 DM 約 100 萬封，每封 DM 成本約 10-12 元，但平均回應率低於2 %
- 效果：
  - 經過資料分析後，了解客群分布，可進行精準的目標行銷
  - 每年寄送的會員 DM 降為 20 萬封，回應率提高為 8-10 %
  - 可以更準確地開發新客群，以及進行存書控制

wyang@ntu.edu.tw

Comparison: Association and MI



Association 檢驗方式

尿布→啤酒

Association
有方向性

MI
無方向性

- Mutual Information

$$MI = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = \log \frac{f(x,y)}{f(x)f(y)}$$

P : probability

N : size of the corpus

f(x) : the occurrences of term x in the corpus

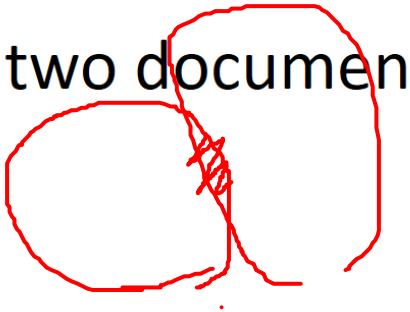
f(y) : the occurrences of term x in the corpus

f(x,y) : the co-occurrences of term x and y in the corpus

wyang@ntu.edu.tw

Comparison: Association and JC

- A commonly used measure of overlap of two sets
- Let *A* and *B* be two sets from n-gram of two documents
- Jaccard coefficient:



$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

用來表示正規化後的集合重疊程度

- JACCARD(*A*,*A*) = 1
- JACCARD(*A*,*B*) = 0 if $A \cap B = 0$
- *A* and *B* don't have to be the same size.
- Always assigns a number between 0 and 1.

Exercise

- 語料
  - 產業新聞(industry)內容全文
  - 給定5個字詞: 鴻海, 郭台銘, 夏普, 蘋果, 手機
- 次數統計

	鴻海	郭台銘	夏普	蘋果	手機
次數	475	193	263	484	874

	鴻海&郭台銘	鴻海|郭台銘	鴻海&夏普	鴻海|夏普
次數	192	476	248	490

	鴻海&蘋果	鴻海|蘋果	鴻海&手機	鴻海|手機
次數	131	828	128	1221

wyang@ntu.edu.tw

Exercise

Jaccard coefficient

J.C.	郭台銘	夏普	蘋果	手機
鴻海	192/476=0.40	248/490=0.51	131/828=0.16	128/1221=0.10

Mutual Information

M.I.	郭台銘	夏普	蘋果	手機
鴻海	log(192/(475*193)) =-2.68	log(248/(475*263)) =-2.70	=-3.24	=-3.51

wyang@ntu.edu.tw

Exercise

Association

↓這邊假設總數就是聯集數

Rules	confidence	support	lift
鴻海→郭台銘	192/475=40%	192/476=40%	(192/475)/(193/476) =1.00
鴻海→夏普	248/475=52%	248/490=51%	(248/475)/(263/490)=0.97
鴻海→蘋果	=28%	=16%	=0.47
鴻海→手機	=27%	=10%	=0.38

反向比較

無方向性(就是J.C.)　無方向性(類似M.I.)

Rules	confidence	support	lift
郭台銘→鴻海	192/193=99%	192/476=40%	(192/193)/(475/476) =1.00
夏普→鴻海	248/263=94%	248/490=51%	(248/263)/(475/490)=0.97
蘋果→鴻海	=27%	=16%	=0.47
手機→鴻海	=15%	=10%	=0.38

wyang@ntu.edu.tw

Discussion

- Co-occurrence 共現是一種相關 (Relevance)
  - 共現是一種不隨機的信號
  - 可以由此推論出因果嗎？

    找出方向性及時序上的相關性
    還需要理論支持，及通過操作性驗證

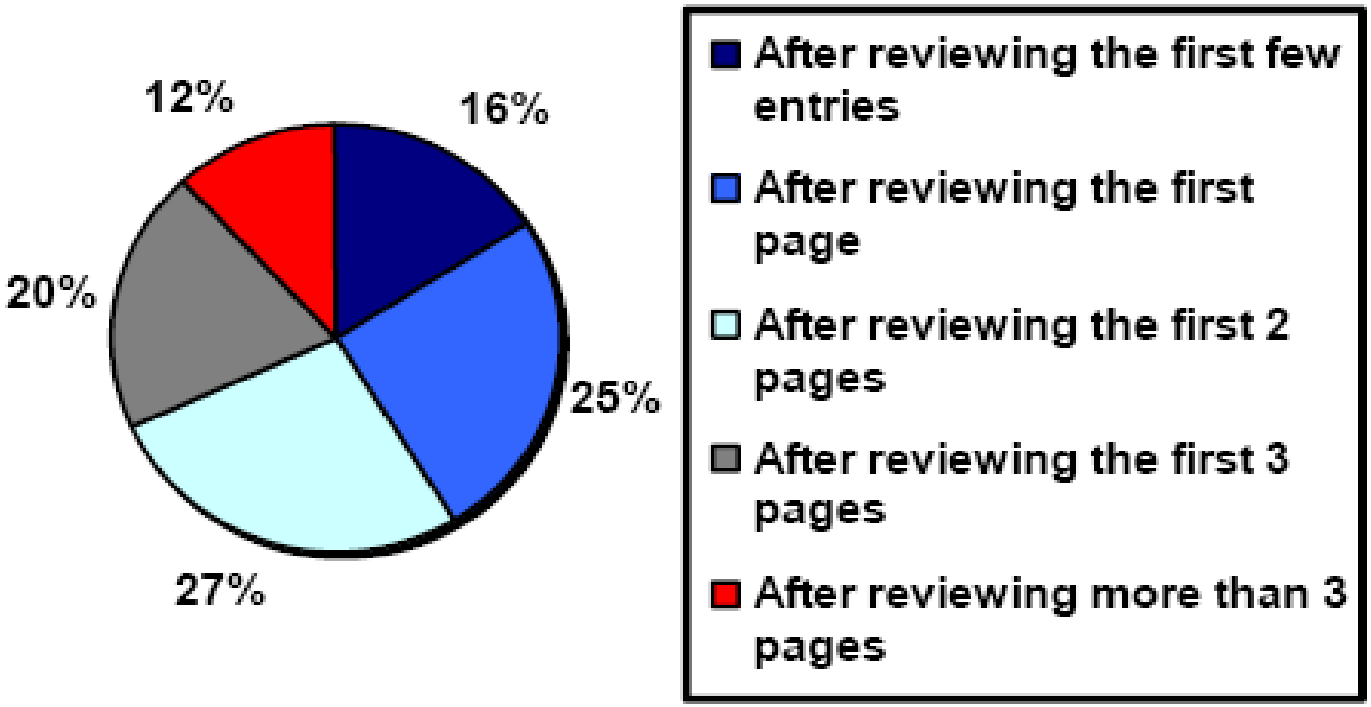
- 問題: 請找出與鴻海在過去產業新聞中，最相關的詞，並依相關性進行排序
  - →夏普、郭台銘、蘋果、手機
  - 或有更好的方法？

wyang@ntu.edu.tw

Link analysis

前言：How far do people look for results?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

How Google rank a website (or a page) ?

- For a good search engine, important websites or pages should be displayed first in the search result.
- The goal is the same for an analyst, to rank the importance of websites or pages.
- …but how to do so ?

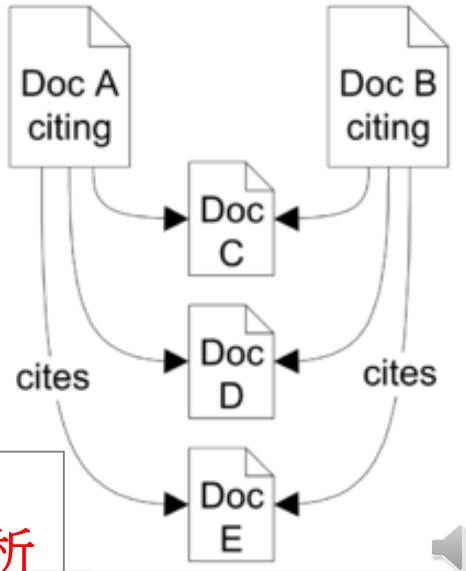
Some techniques from Google

- 與查詢相關的內容
- 獨特、明確的內容； 不重覆、不是垃圾的內容
- 大家喜歡的內容
- 經常更新的內容
- 被具有影響力的網站引用的內容 (PageRank algorithm)
- 其他：有提供行動版、網頁回應速度快等

註: 針對搜尋引擎的排序配方，以不付費的方式，努力提升自身網站之排名，在商業界稱為搜尋引擎優化 (search engine optimization, 簡稱SEO)，是一種商業模式，屬於行銷範疇之一 (提升曝光及開拓資訊通路)，有龐大的相關產業

Origins of PageRank: Citation analysis (1)

- Citation analysis: analysis of citations in the scientific literature.
- Co-citation analysis and Bibliographic coupling analysis
  - articles that are cited together are related. Ex. C, D, E
  - articles that co-cite the same articles are related . Ex. A, B
- Citation analysis works for scientific literature, patents, web pages, and directed documents.
  - Google use co-citation similarity on the web for "find pages like this" feature.



Co-citation正式翻譯為「共被引」，故可稱共被引分析
Bibliographic coupling正式翻譯為「書目耦合」，故可稱書目對分析

Origins of PageRank: Citation analysis (2)

- Citation frequency can be used to measure the impact of an article .
  - Ex. Google Scholar, CiteSeer
- On the web: citation frequency = inlink count
  - Simplest measure: Each article gets one vote
  - A high inlink count mean high quality.
  - … but not very accurate because of link spam.
- Better measure: weighted citation frequency or citation rank
  - An article's vote is weighted according to its citation impact.

Ex. NY Times inlink is much more important than a nobody's inlink.

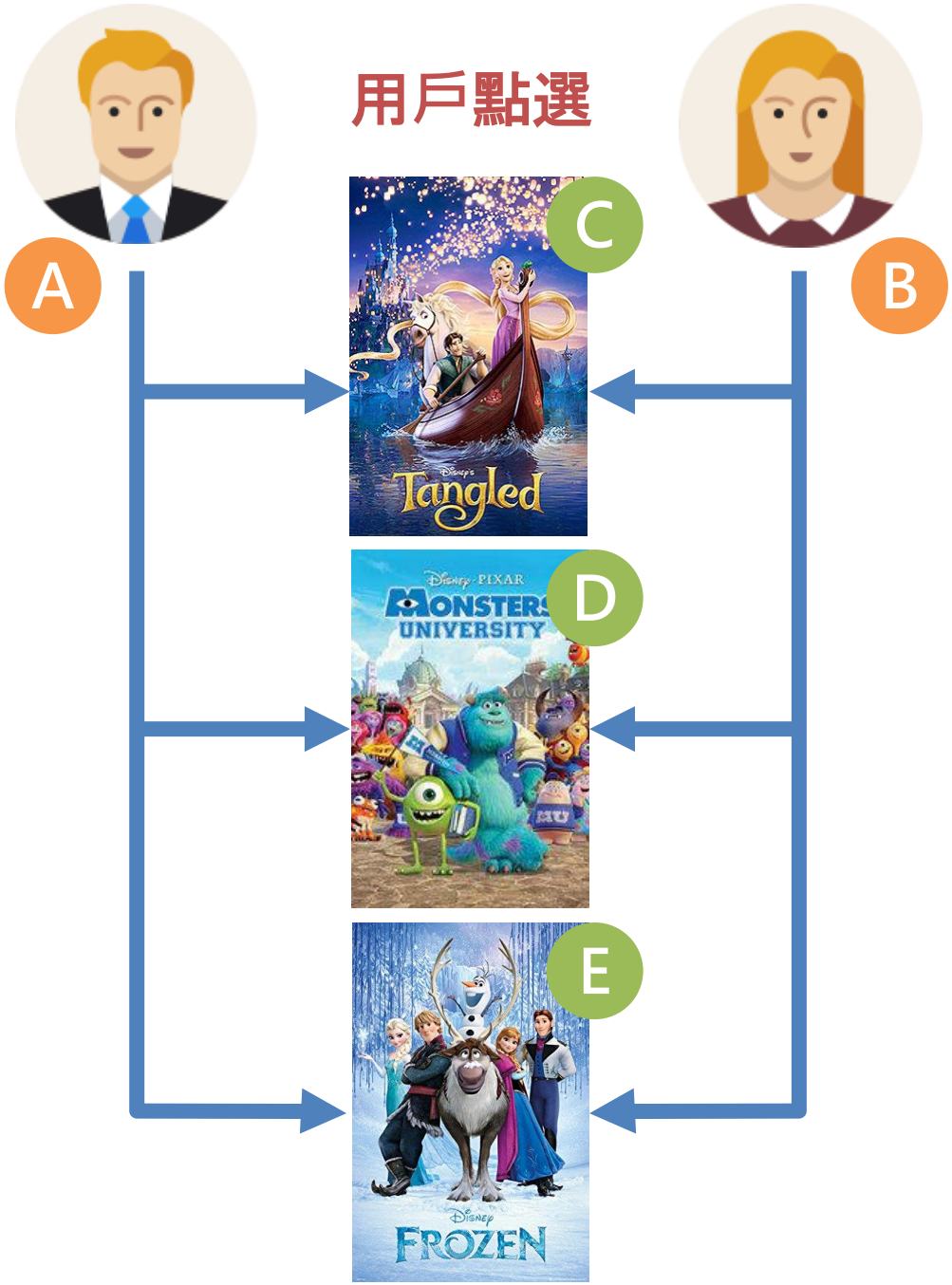
Origins of PageRank: Citation analysis (3)

- Weighted citation frequency or citation rank is basically PageRank
  - invented in the context of citation analysis by Pinsker and Narin in the 1960s.
  - Google uses it and other heuristics for web page ranking. (independent from query)

討論：擴大應用

- Co-occurrence may be seen as a kind of link
  - The authors who post articles in the same board
    - Ex. John, Mary都在PTT Pet板貼文; John及Mary有關係
  - For a specific author, the boards where he/she posts articles
    - Ex. John在PTT Pet板及Dog板貼文; Pet板及Dog板有關係
- It can be extended to a network analysis
  - Ex. Bob在Dog板貼文, Mary及Bob有第二層關係
- It can be generalized for many applications
  - User-User, Item-Item, User-Item, and so on.

wyang@ntu.edu.tw



綜合應用：

依照影片描述內容做相似性計算(例如用VSM) 後由C推薦D、E，業界常稱相似推薦

依照A及B兩人觀影紀錄，由C推薦D、E (Link analysis之共被引分析，為Collaborative filtering協同過濾之一種)，業界常稱They also like，可作為銷售擴展

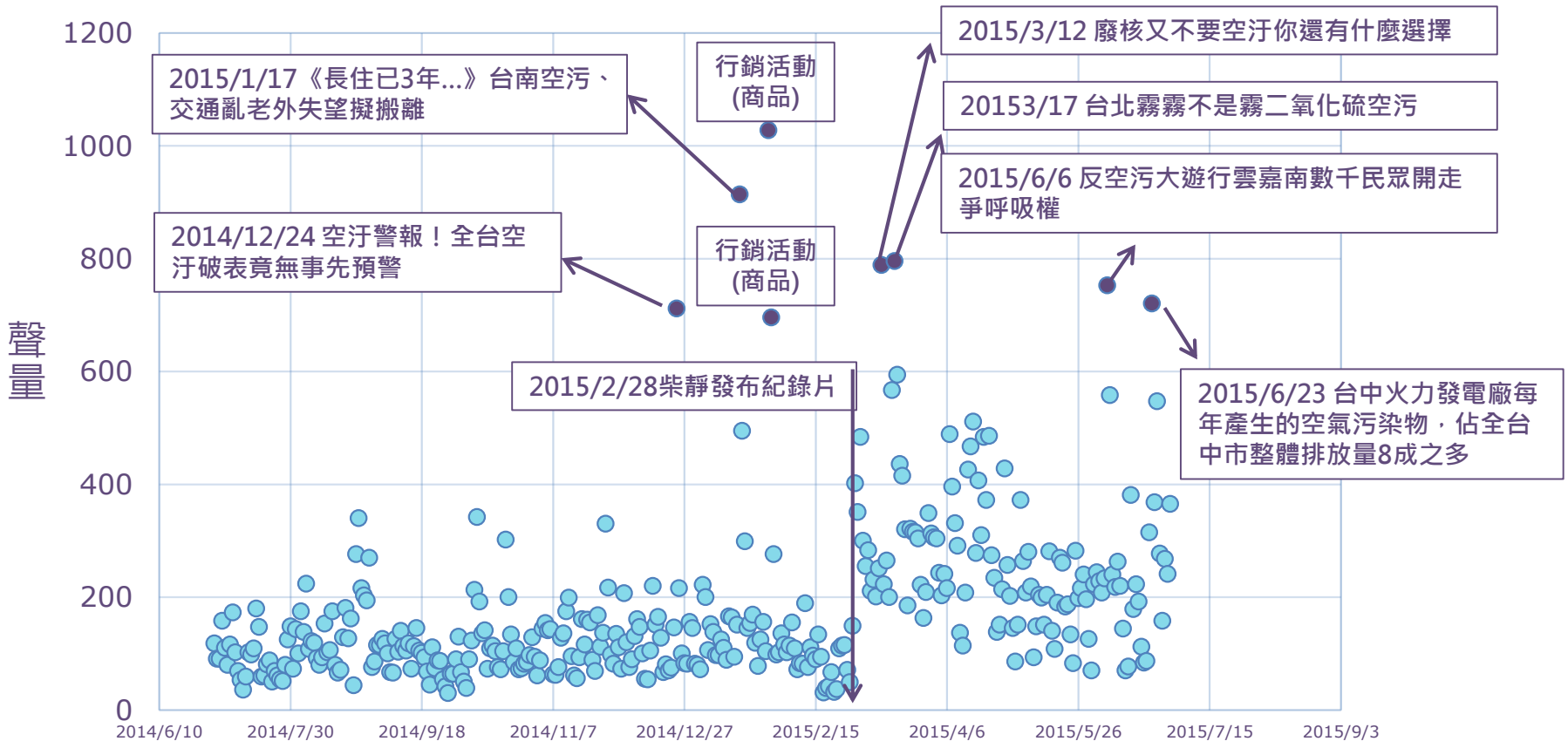
依照A及B兩人觀影紀錄，由A找到相似觀眾B (Link analysis之書目對分析)，業界常稱為Look alike，可作為用戶擴展

wyang@ntu.edu.tw

綜合案例：空污議題之看法

- 觀察網路上的民眾對於空污意識的變化程度
- 觀察期間為一年，從2014/07/01至2015/06/30
  - 原始關鍵字：民眾在討論空氣品質時，可能提及的關鍵字詞，包括空氣髒、不乾淨、汙染、污染，簡稱為空汙或空污，或該紀錄片中特別提及的「PM2.5」等。
  - 人名及社群網絡圖：採用文字探勘中的命名實體識別（named entity recognition）技術，以意藍公司的Tornado Text Miner實作，擷取出民眾討論內容中提及的人名，再對比新聞內容中前50位最常被提及的人名後。將人名做為節點（node），不同人名同時出現在同篇新聞中做為接線（link），以建立社群網絡圖。

綜合案例：空污議題聲量散布圖



每日的平均聲量為178則，並且有8天單日超過600則，其中有2天(1/28、1/29)是Facebook粉絲團的行銷活動造成的大量聲量。

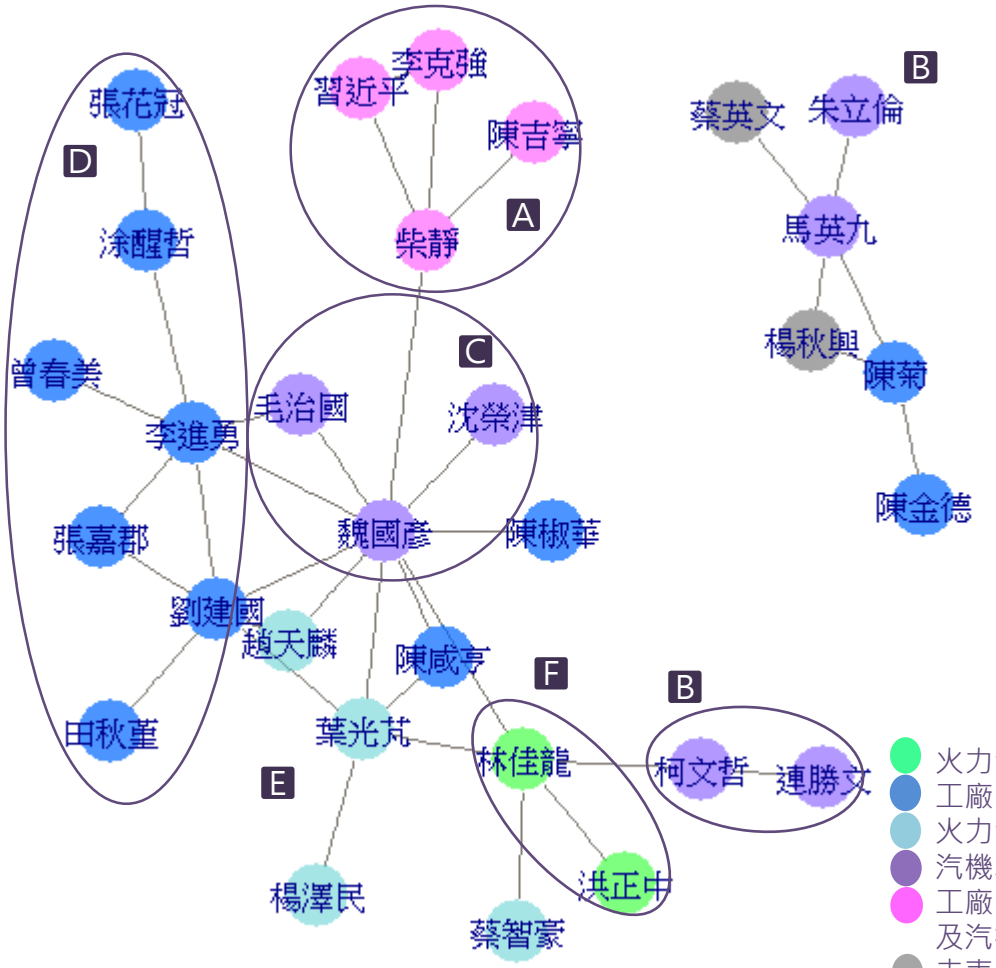
另外，在2/28的柴靜發布紀錄片過後，聲量明顯上升，平均聲量從129則提升至276則。

社群網絡圖 各群集及人名認為之空汙原因

D 雲嘉南
以李進勇為首擴及雲嘉南地區相關政治人物，認為空污最大污染源為**工廠**。

E 環保組織
葉光芃、楊澤民長期在中彰投地區活動，認為境內最大空污來源為**火力發電及工廠**。

F 台中
林佳龍與洪正中認為中部境內空污最大污染源為**台中火力發電廠**。



A 中國
柴靜《穹頂之下》及中國當局針對**工廠、火力發電及汽機車**皆有譴責及政策實施。

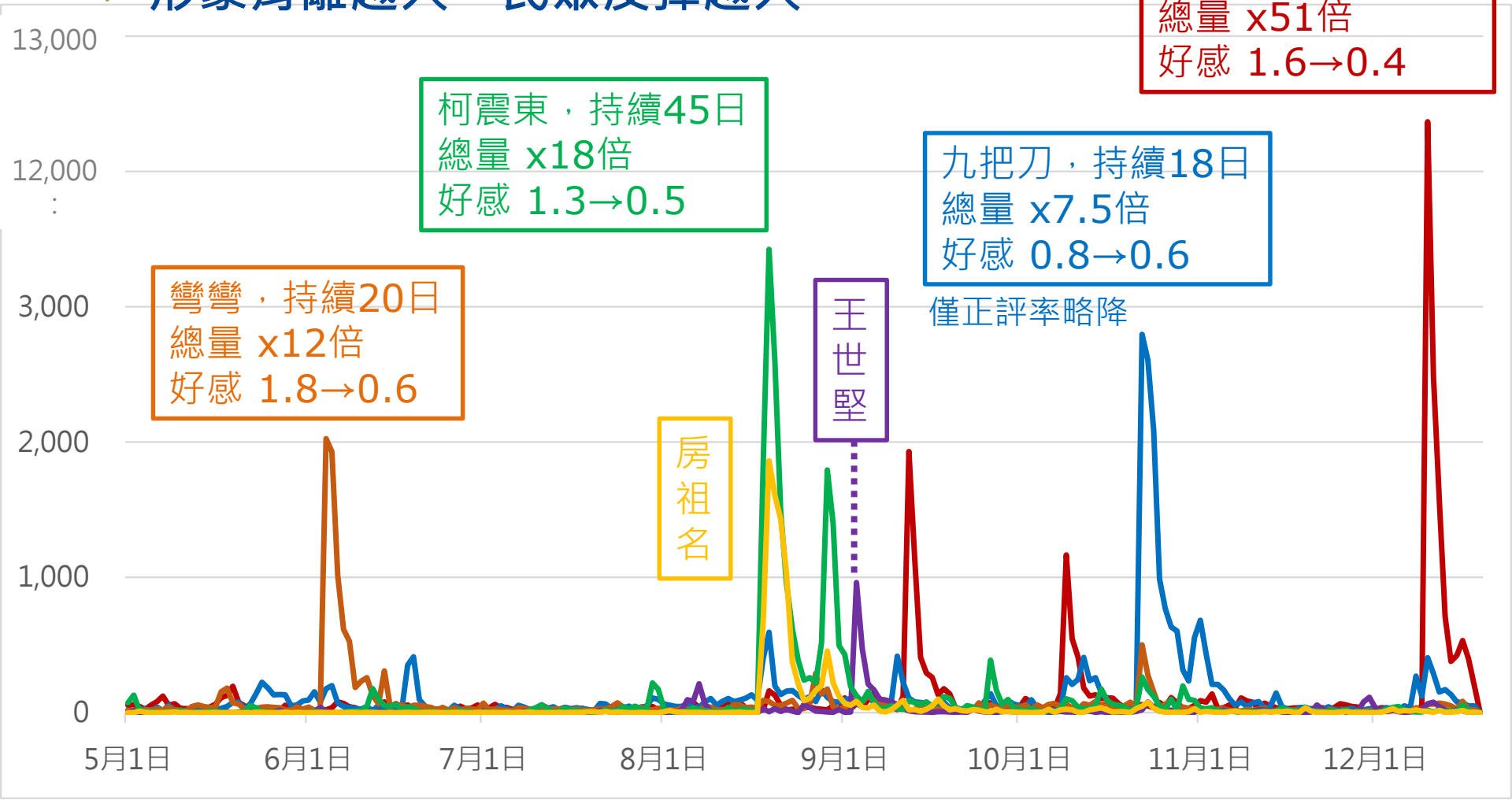
B 雙北
北部地區包含馬英九、柯文哲等主要領導人，認為北部空污最大污染源為**汽機車**排放廢氣。

C 中央政府官員
毛治國、沈榮津、魏國彥站在經濟層面立場，認為應從**汽機車**改善。

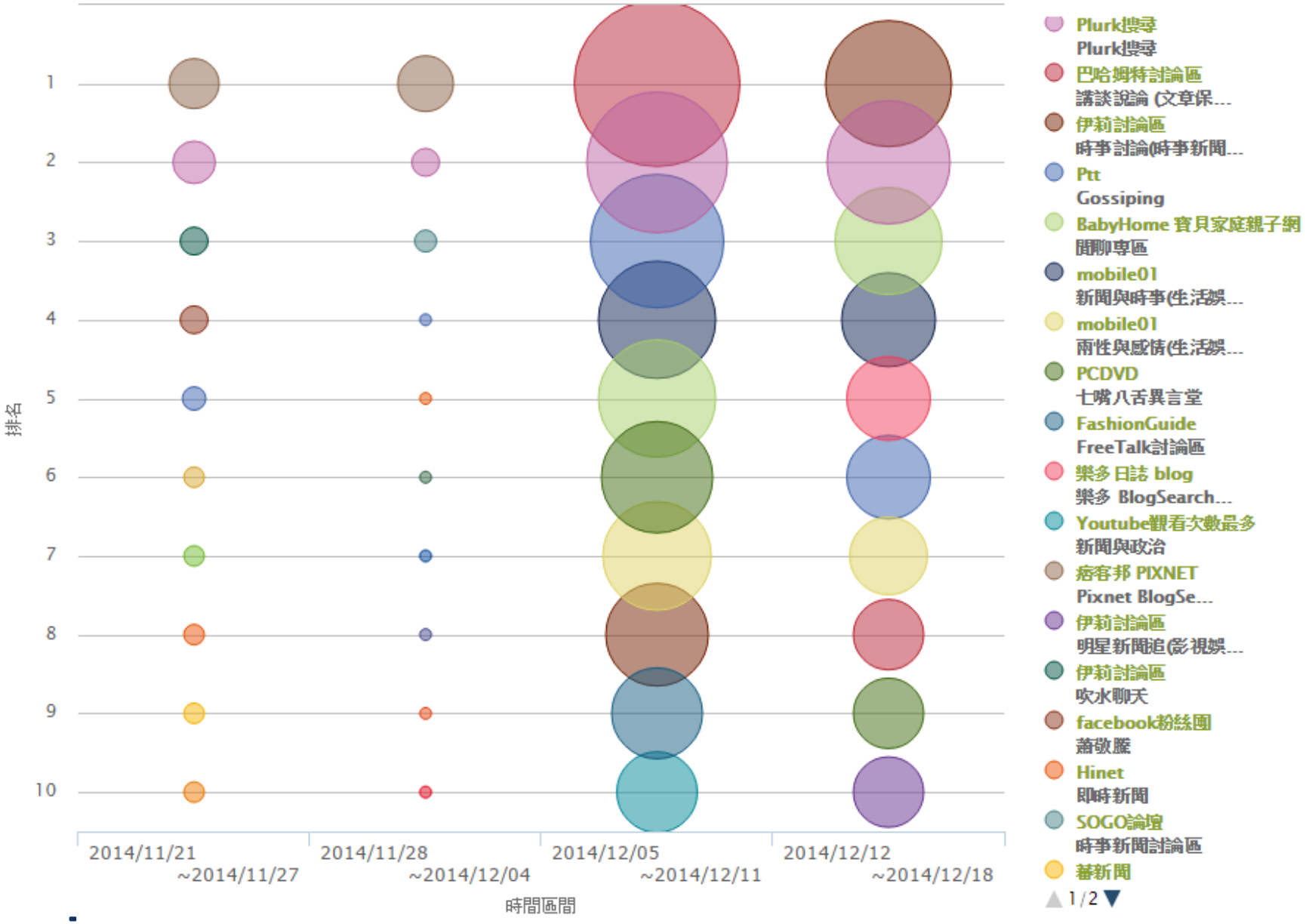
權重>20
表示兩位關鍵人物在超過20篇以上的文章之中被同時提及。

綜合案例: 名人外遇劈腿醜聞多，引發熱議

- 形象背離越大，民眾反彈越大



「阿基師事件」社群傳播路徑圖



「阿基師事件」對消費品牌的影響

前五大討論區		好評度變化
1	mobile01	1.0→0.4 ▼0.6
2	伊莉討論區	2.3→0.3 ▼2.0
3	巴哈姆特	2.0→0.4 ▼1.6
4	PTT	1.1→0.6 ▼0.5
5	Babyhome	**3.0→0.5** ▼2.5

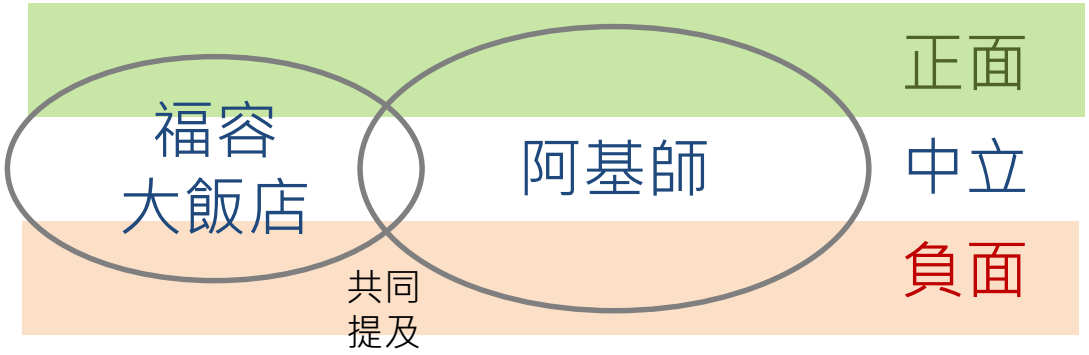
媽媽們的破滅

相關代言商品	受影響提及程度	整體品牌好評度變化
福容大飯店	54%	2.3→1.1 ▼1.2
桂格 / 得意的一天	7%	無變化
福樂 / 一番鮮	6%	無變化
牛頭牌	11%	1.7→1.3 ▼0.4
型男大主廚	73%	1.7→**0.7** ▼1.0
五月花	12%	1.5→1.2 ▼0.3
7-Eleven 年菜	38%	1.9→1.5 ▼0.4
百略醫學	17%	無變化
義廚寶	52%	2.9→1.2 ▼1.7

拆解子議題

使用AND OR NOT

關鍵字篩選



主題	總聲量數	正面情緒聲量數	正面情緒比例(P)	負面情緒聲量數	負面情緒比例(N)	中立情緒聲量數	顯著情緒比例	P/N比
阿基師	36719	6849	18.7%	16162	44%	15987	56.5%	0.42
福容	716	349	48.7%	307	42.9%	235	67.2%	1.14
阿基師及福容	392	155	39.5%	225	57.4%	116	70.4%	0.69
阿基師(不含福容)	36327	6694	18.4%	15937	43.9%	15871	56.3%	0.42
福容(不含阿基師)	324	194	59.9%	82	25.3%	119	63.3%	2.37

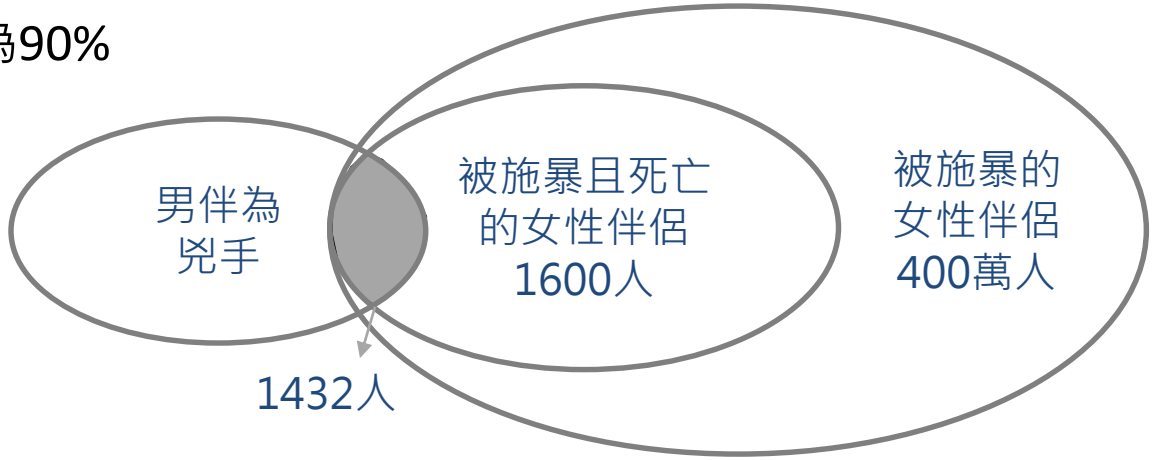
衝擊分析

- 討論福容時，有54.7%提及阿基師
- 福容負評中，有73.3%提及阿基師 (lift 1.34)
- 單獨提及福容之PN值，與共同提及之PN值，由2.37→0.69 ▼1.68

wyang@ntu.edu.tw

Case Study : Conditional Probability

- 辛普森(O.J. Simpson)是當年著名的美國足球明星，因為涉嫌殺害自己的前妻被起訴，引起軒然大波，
- 辛普森的律師辯述：美國400萬被施暴的女性伴侶中只有1432名被其男伴殺死。所以得出，辛普森可能殺死前妻的機率只有1432/400萬，大約為1/2800
- 但在美國400萬被施暴的女性伴侶中，死亡的人數是1600人，其中被男伴殺害的是1432 。因此被施暴且死亡的女性伴侶，男伴是兇手的機率是1432/1600，約為90%
- 應採何種算法？



Conclusion

- Co-occurrences and Link are important signals for data mining, and can be generalized to many applications.
  - terms and documents
  - users and items
  - and many.
- Some analytical techniques are introduced.
  - Conditional probabilities, Venn diagram, Association
    - to compare with other data / benchmarks
  - Co-citation and Bibliographic coupling analysis

Discussions