# 巨量資料分析導論 Term Project

s101065802 楊庭豪
s103062590 張維元
s100062114 葉安琪

## 主題

Reviewer Recommender Systems
本次計畫預計實作一個 Paper Reviewer Recommender Systems，當一篇新的論文輸入時，本系統會根據該論文的摘要、引用資料等特徵推薦出適合的 reviewer 清單，實做此系統，預計會使用到 Chapter 3 Finding similar document、 Chapter 7 Clustering、以及 Chapter 9 Recommender Systems 的概念以及演算法

## Datasets

Arxiv High Energy Physics theory paper

- https://snap.stanford.edu/data/cit-HepTh.html
- http://www.cs.cornell.edu/projects/kddcup/datasets.html

第一個網址是該資料的 citation graph 資訊，第二個則是包含文章內容的原始資料，資料大小約 1.7 G

| Dataset statistics | |
|---|---|
| Nodes | 27770 |
| Edges | 352807 |
| Nodes in largest WCC | 27400 (0.987) |
| Edges in largest WCC | 352542 (0.999) |
| Nodes in largest SCC | 7464 (0.269) |
| Edges in largest SCC | 116268 (0.330) |

從 KDD 下載回來的資料是以 Latex 格式儲存，透過 Latex 的格式化指令我們可以做前處理找出類似「\author」、「\begin{abstract} 」等指令來幫助我們把資料拆解成各模組所需要的格式

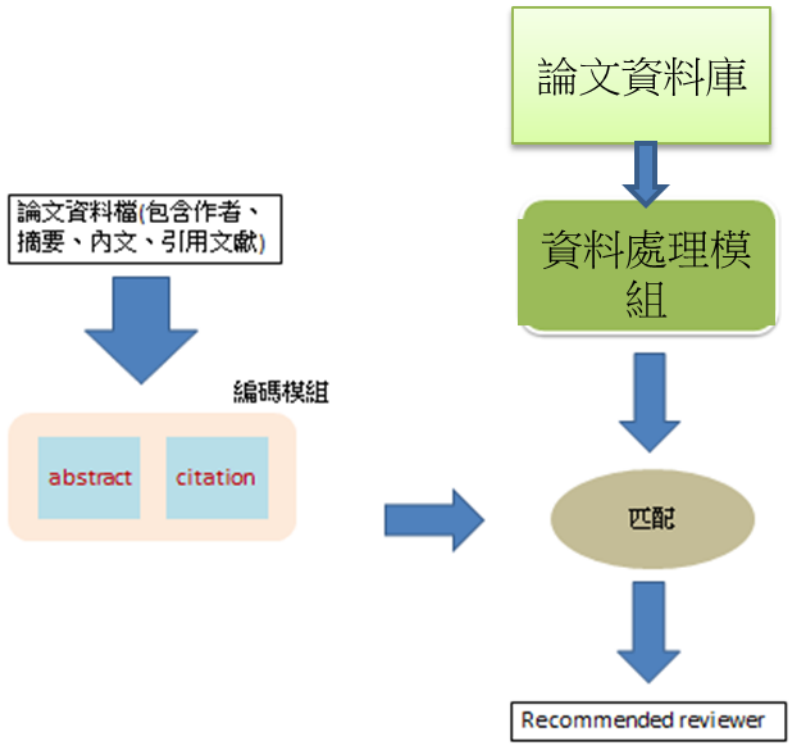
## 系統架構與方法

整個系統大致上可以分成下列幾個部分

1. 編碼模組 ： 這部分會分析輸入的文件，並將其轉換為 vector，此模組會把摘要部分以 TF-IDF 方式轉換成 vector，對於引用關係則是建立是否有引用其他文獻編碼的 Boolean Matrices 後，會透過隨機產生一些線性雜湊函數轉換為 signatureMatrix。
2. 論文: 此部分會將 Latex 格式的資料加以處理，轉換成我們需要的作者、引

用資料、標題等資訊，

3. 匹配模組：將經過編碼模組的特徵資料與整理過的作者資訊做匹配，尋找學術專長相近的作者當作推薦的 reviewer 人選。




在論文資料的處理上，我們把他分成 abstract 包含的文字資訊以及 citation 的引用關係這兩部分，

- Abstract

  abstract 的資料中包含 Title、 Authors、Abstract，
  首先，對每一篇 abstract 使用 TF-IDF 找出重要的字，然後把重要的字當成此作者的特徵向量
  進一步，將這些作者依據字去做 cluster，被分在相同 cluster 的作者表示發表的內容相近，那表示是適合做 reviewer 的

- Citation

  citation 部分預計會利用網站提供的 citation graph 資訊，將一個 paper 有沒有引用其他文獻編碼為 Boolean Matrices，用 Min-hash 跟 LSH 找出有引用相似文獻的作者

最後整合兩部分的推薦人選，有共通的優先推出，不足的人選部分則比較匹配的相似度資訊來挑選。

## 預期結果

我們會將 hep-th-2003 保留為測試資料，當將一份論文的資料檔案輸入時，會輸出系統推薦的 3~5 位 reviewer 人選

## 實驗數據

### 1 Abstract feature

#### 1.1 abstract feature experiment

| dataset size vs time (cluster number = dataset/5) | | | | |
|---|---|---|---|---|
| dataset | 10 | 100 | 1000 | 29554 |
| time(sec) | 0.486594 | 0.905419 | 4.933667 | 383.960394 |

#### 1.2 cluster number vs time

| based on 1000 size | | | | |
|---|---|---|---|---|
| cluster | 10 | 50 | 100 | 500 |
| time(sec) | 4.408036 | 6.700500 | 9.544867 | 32.105099 |

| based on 10000 size | | | |
|---|---|---|---|
| cluster | 10 | 100 | 1000 |
| time(sec) | 45.53684 | 69.36755 | 292.4482 |

### 2. Citation feature experiment

#### 2.1 Number of hash functions

| Number of hash functions VS number of relations of similar documents | | | |
|---|---|---|---|
| Num of hash functions | 3 | 100 | 1000 |
| relations | 28840 | 151908 | 152576 |
| time | 10260 | 13460 | 35480 |

#### 2.2 Number of reducer VS time

| based on 1000 hash | | | | |
|---|---|---|---|---|
| reducer | 1 | 2 | 3 | 4 |
| time(sec) | 35480 | 42930 | 40450 | 44090 |

### 3.推薦實例

#### 3.1

論文編號：9201001

推薦作者：

F.Bonechi

Igor R. Klebanov
Mirjam Cvetic
Satoshi Matsuda
A.Marshakov

3.2
論文編號：9301001
推薦作者：
Jerome P. Gauntlett
Maximilian Kreuzer
Harald Skarke
I. Bars
K. Sfetsos

3.3
論文編號：0301019
Ashoke Sen
Gary T. Horowitz
M. Cadoni
Aoki
Horikoshi

3.4
論文編號：0302077
Xiang Shen
Fiorenzo Bastianelli
Leonardo Castellani
Changhyun Ahn
Soonkeon Nam

3.5
論文編號：0303040
Y. M. Cho
M. L. Walker
D.G. Pak

## 程式檔案說明

本次計畫的程式檔案放在
https://github.com/tinghaoyang/BigDataTermProject

AbstractRecommend 放的是處理 Abstract 這塊的模組以及前處理程式
ReviewerRecommend 放的是處理 citation 這塊的模組