實作估計總作業時間的網路服務

基於郵件狀態改變及其改變的時間

團隊名稱：bigdata最後希望
帶隊教授：趙逢毅博士
成　　員：胡芯瑜
陳博文
石振琳
鄭龍森

摘要

信件(包裹)從離開寄件人手中開始，到達收件人手上結束，在這過程當中存在許多變數，例如氣候、路況、送件量、路途遠近、路線規劃方式等；但基於確切的郵件狀態變化加上狀態變化的時間差距，雖然在巨觀上具有一定程序，但在細節上卻存在著不確定性，藉由馬可夫鏈（英語：Markov chain）的計算推估郵件在離開寄件者手中完成資料登錄之後，需要多少的時間能送達收件者的手上；而在送件過程中，郵務中心也可以即時監控運送過程的異常狀態，以縮短異常排除時間，提高郵件運送效率。

提案動機

「XXX！掛號！掛號！」

曾幾何時，粗獷嘹亮的掛號唱名聲響遍大街小巷，而時光荏苒，曾經的掛號唱名因電鈴而不再常聽，但等待信件的心卻是不變。工業化的發展加速了人們的腳步，也壓縮了等待的心；尤以近年電商發展迅速，甚有電商喊出6小時到貨服務，而郵件呢？擁有最悠久歷史、最熟悉街頭巷尾的郵務系統，有沒有可能發展出幾點收信(包裹)的服務呢？

答案是可以的，大量的歷史資料、健全的作業流程，只要搭配一套完整的系統，讓寄件人在郵局收到信件(包裹)建檔開始，就能知道對方什麼時候可以收到；也許是愛情的思念，又或許是親情的掛念，因為郵務系統的精準送達，讓懸著的一顆心不再因為未知的等待而焦躁不安。

選用理論架構模型

觀察
- 郵件追蹤查詢資料
- 各欄位的關係

轉換
- 網路模型的有向路徑圖

發現
- 狀態和狀態之間為隨機過程
- 郵件主要的狀態改變路徑

選用理論架構模型

網路模型(network model)

定義

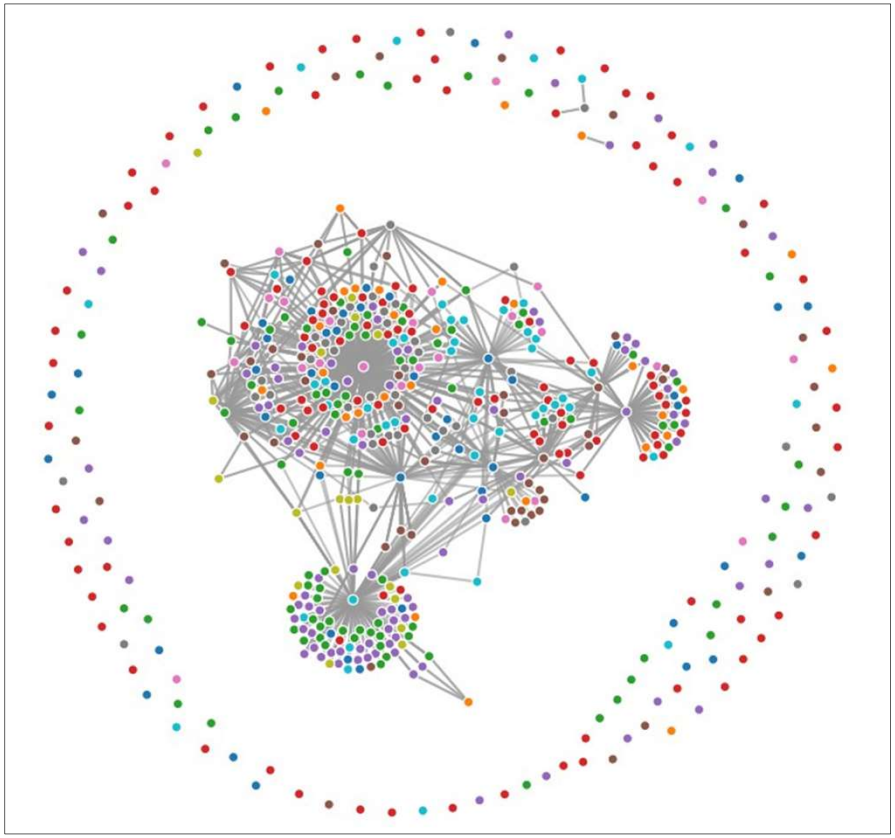
1. 點:郵件的狀態代碼、處理局號
2. 線:狀態的改變
3. 依據狀態、郵件號碼、時間決定線的連結

選用理論架構模型

狀態-狀態之間的關係(10000筆資料)

選用理論架構模型

局號-狀態之間的關係(10,000筆資料)



選用理論架構模型

馬可夫鏈模型

馬可夫性質:

在目前以及所有過去事件的條件下，任何未來事件發生的機率，和過去的事件是不相關的(獨立的)，而僅和目前的狀態相關。

具備馬可夫性質的隨機過程則稱為馬可夫鏈。

原理:

利用歸納事件的所有狀態，統計出事件的狀態轉移的機率，表示成轉移機率矩陣來進行模擬分析，參數可隨時間具有系統性，顧客用來預測未來事件狀態的轉移或是空間擴張的趨勢。

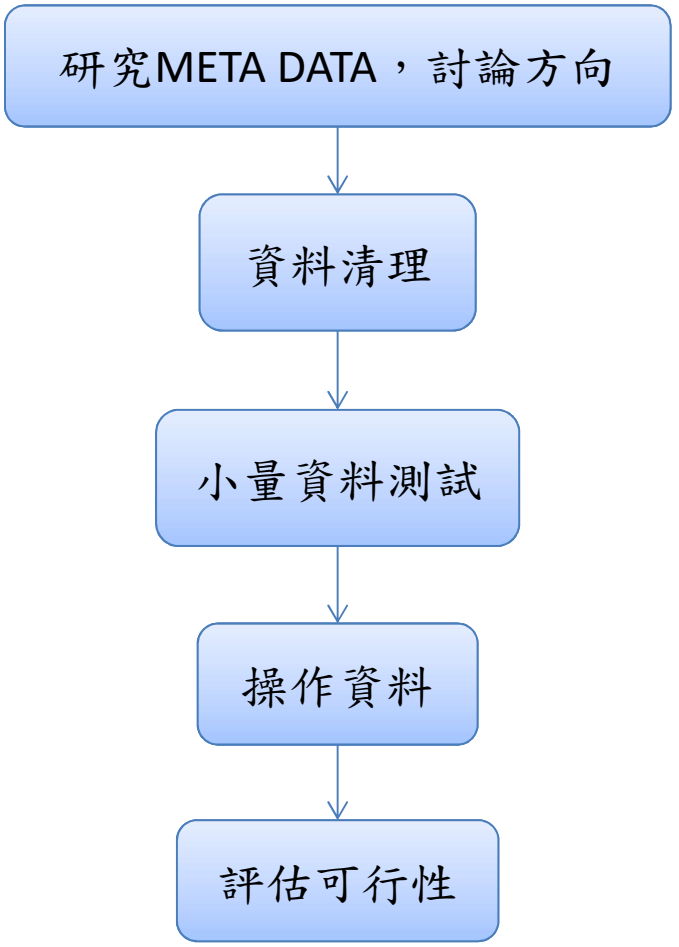
選用理論架構模型

處理環境

目前處理環境：
硬體：個人PC
作業系統：WINDOWS 10專業版
軟體：Anaconda3(64-bit)、EXCEL
程式語言：PYTHON 3
資料庫：無

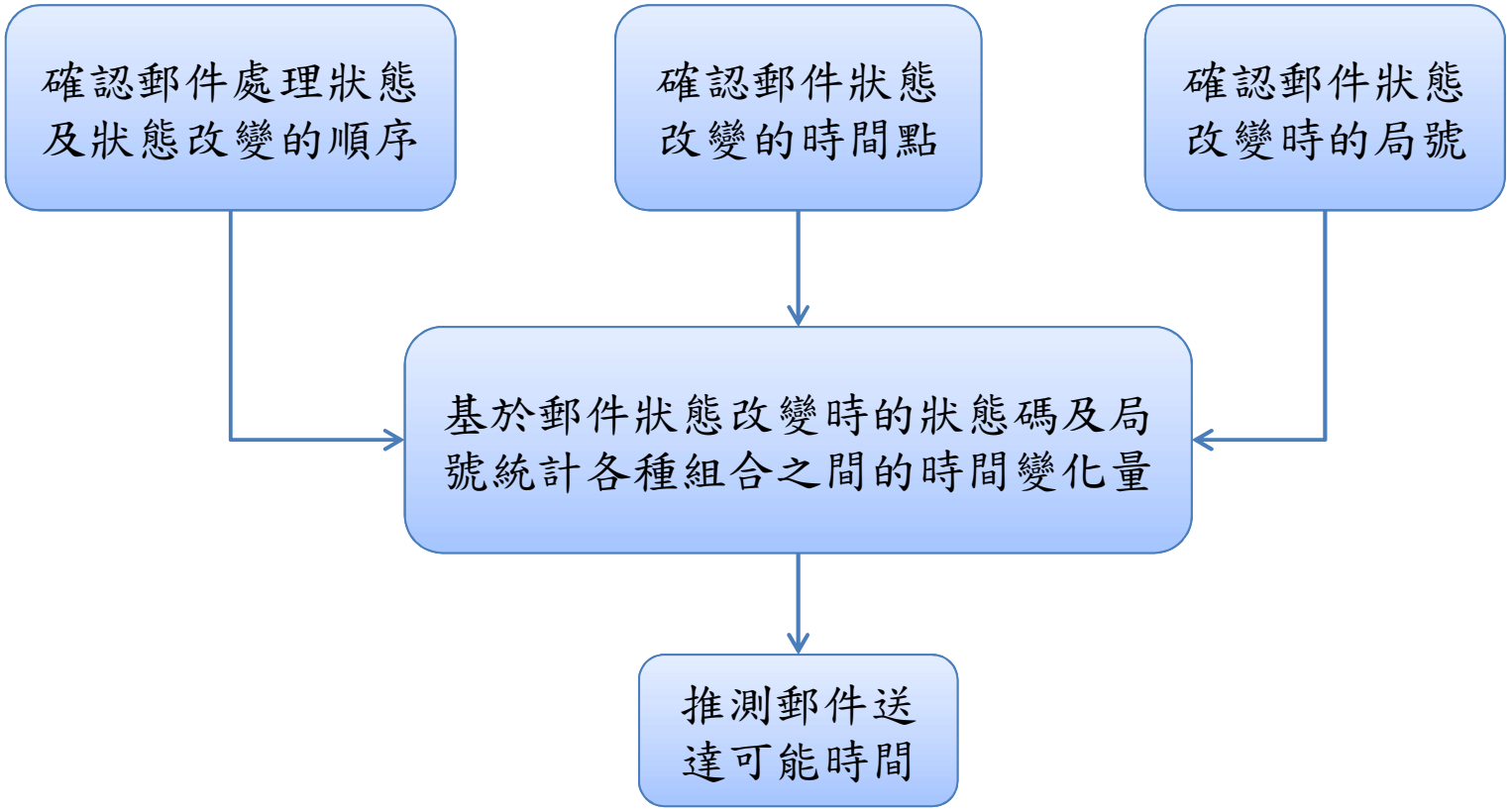
分析流程說明

分析流程



分析流程說明

分析架構


確認郵件處理狀態
及狀態改變的順序
確認郵件狀態
改變的時間點
確認郵件狀態
改變時的局號
基於郵件狀態改變時的狀態碼及局
號統計各種組合之間的時間變化量
推測郵件送
達可能時間

分析流程說明

分析報表結果

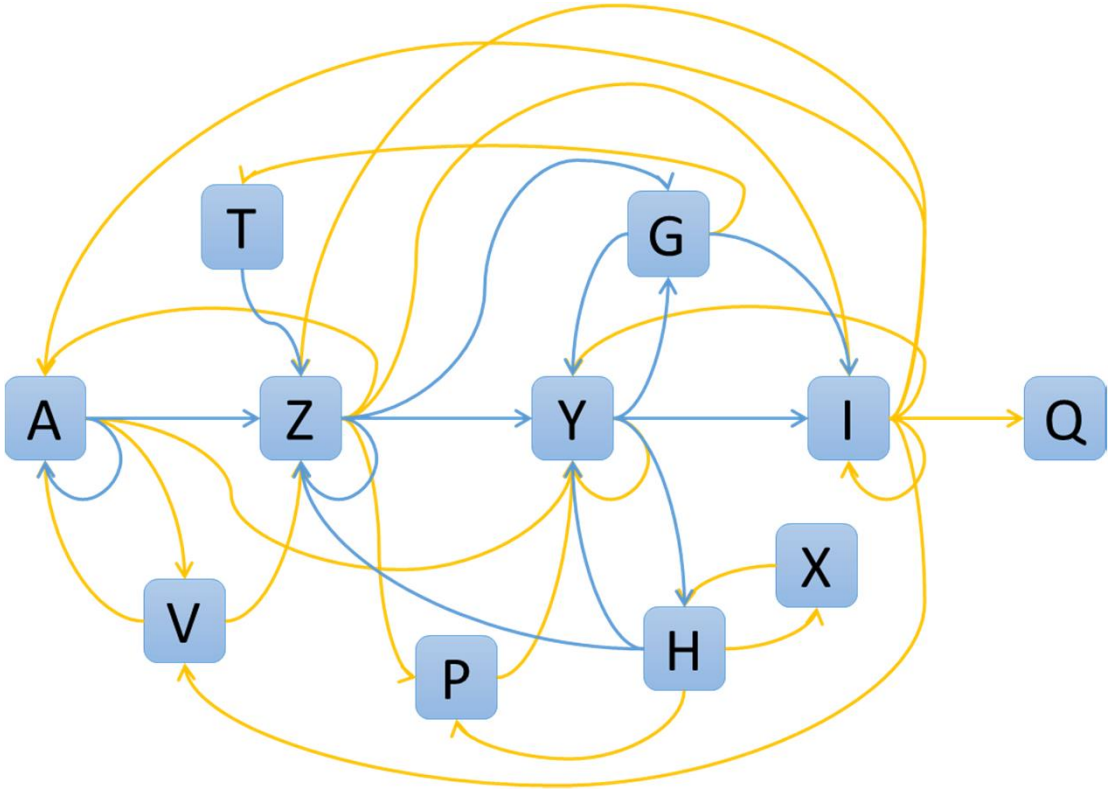
狀態變化	計數	平時時數	百分比	累計百分比
Z -> Y	75,647,648	21.20	26.00%	26.00%
Y -> I	74,231,948	6.11	25.51%	51.51%
A -> Z	61,845,982	7.88	21.25%	72.76%
Y -> H	17,046,769	5.12	5.86%	78.62%
Z -> Z	13,297,445	12.92	4.57%	83.19%
H -> Y	12,348,865	25.20	4.24%	87.44%
Z -> G	6,987,350	15.92	2.40%	89.84%
G -> I	5,335,784	60.16	1.83%	91.67%
H -> Z	4,378,472	8.19	1.50%	93.18%
Y -> G	2,363,505	20.20	0.81%	93.99%
G -> Y	1,306,300	0.83	0.45%	94.44%
T -> Z	1,283,155	3.86	0.44%	94.88%
A -> A	1,192,365	542.03	0.41%	95.29%
A -> Y	1,182,150	26.91	0.41%	95.69%
P -> Y	1,155,135	12.34	0.40%	96.09%
I -> Q	1,013,289	15.41	0.35%	96.44%
G -> T	891,498	452.02	0.31%	96.75%
Z -> P	885,811	18.32	0.30%	97.05%
Z -> I	836,587	31.69	0.29%	97.34%
X -> H	731,999	4.81	0.25%	97.59%
I -> A	627,866	478.43	0.22%	97.80%
I -> Z	530,107	287.71	0.18%	97.99%
A -> V	463,495	20.59	0.16%	98.15%

資料筆數： 381,043,734
有效筆數： 290,977,795
郵件筆數： 89,227,622
平均狀態變化：4.27

由於數據過多，故取累計百分比至98%。

分析流程說明

狀態分析圖



藍色線為狀態變化取累計統計至2σ，橙色線為2σ至接近3σ之間的狀態，由圖可知，在2σ之前的狀態較為穩定(藍色線13條)，而在2σ之後狀態變化大增(橙色線19條)。

分析流程說明

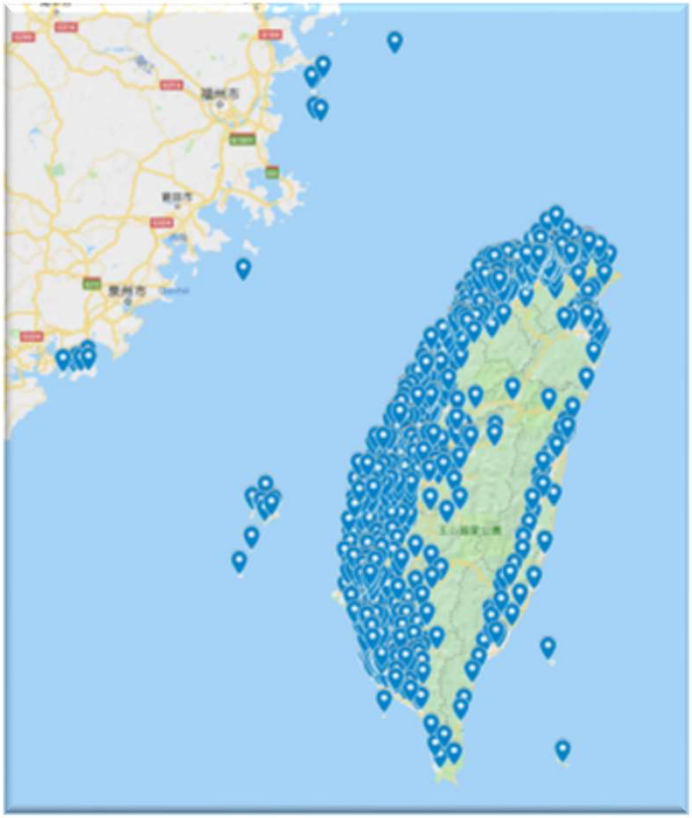
分析結果

以目前一季的資料量(381,043,734)、郵件數量(89,227,622)、平均狀態變化量(4.27)的資料來看，平均每秒需處理49筆的郵件查詢量，而以某電商雙11時以七台伺服器處理每秒約300筆客戶查詢需求推估，服務上線之後約需一台伺服器即可處理每日郵務查詢的需求。

面臨問題

資訊不足

由於無法取得完整郵務處理局號及其地址，原本計劃配合圖資系統進行流程可視化的作業告停，待取得完整資料之後，可考慮重啟可視化作業。



後續計劃

處理環境

目前處理環境：
硬體：個人PC
作業系統：LINUX + Hadoop cluster
軟體：Anaconda3(64-bit)
程式語言：PYTHON 3

後續計劃

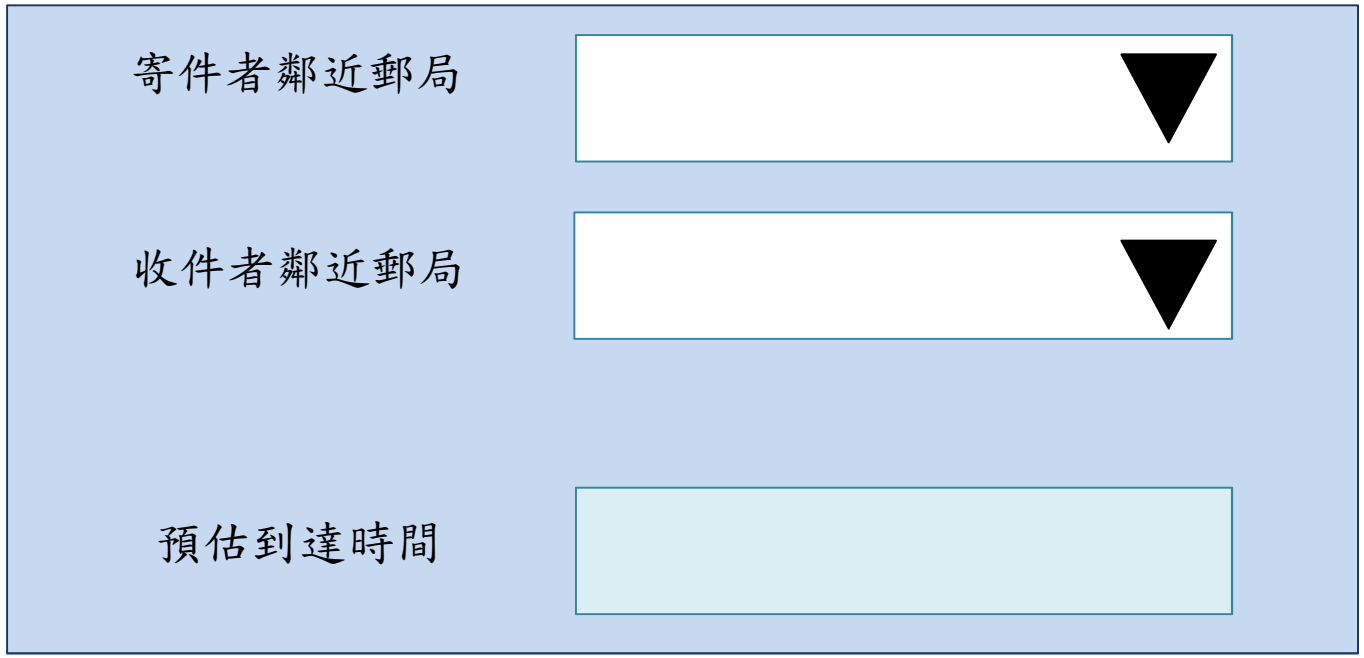
處理方式

重新規劃處理環境，以叢集方式處理整個年度的資料量，甚至經由新資料的加入，重新計算推估時間，以符合環境的變化。

而在最終結果，希望能在郵局窗口完成資料登錄之後能立即知道信件(包裹)送達的時間，並經由界面的運用，讓郵務中心及送(收)件人能查詢所需相關訊息。

後續計劃

用網頁呈現(預估郵件送達時間的網路服務)



研究限制：

無法以 **寄件者地址** 或 **收件者地址** 為輸入欄位，是因為資料無法取得，故為 **研究限制**