報告題名：

巨量資料視覺化之研究

The study of big data visualization

作者：黃昱霖

系級：都市計畫與空間資訊學系　空間資訊組　三年級

學號：D0033231

開課老師：林峰正

課程名稱：網路原理與應用

開課系所：都市計畫與空間資訊學系

開課學年：　102 學年度 第 1 學期

中文摘要

目前針對 Big Data 的處理，多著重於快速處理三大面向的屬性資料(Volume、Velocity、Variety，簡稱 3 V)，及使用商業智慧分析，來發掘出能夠提高商業價值的策略。2012 年三月美國白宮提出的「Big Data Initiative」想法，就是針對現今的極大量資料，研發出新的儲存、管理、分析等技術。而大資料的視覺化分析（Visual Analytics），更是美國政府部門重點研發項目。資料視覺化 (Data Visualizations) 是一個對於人腦最有效的輸入途徑，進而解讀大數據的模式，眾多呈現方式中，標籤雲或稱文字雲大概是最常見的一種資料視覺化方法，針對使用者輸入的文章或是網頁內容，分析其字詞出現的頻率，以字體大小表達出特定期間內最熱門的關鍵字，頻率越高字詞越大。

分析資料視覺化，它是一種相當有趣的技術，透過特殊的運算模式、演算法將各種數據、文字、資料轉換為各種圖表、影像，使得資料可以比較容易為人所理解，因此本論文將探討不同類型的資料呈現方式，研究人員會依據資料的特性，分為數值資料(numerical data)或數量資料(quantitative data)、類別資料 (categorical data)或定義資料(qualitative data)，藉此完成資料類型的分類，再來將資料對應到視覺屬性 (也就是資料編碼)，決定哪一種視覺屬性來表達資料類型是最有效率的，包含 2D 與 3D 的圖形化資料呈現、即時性的報表產生工具、動態儀表板、資料視覺化動畫模擬等工具，論文最後將列出這些不同工具的優缺點評比。

關鍵字：巨量資料、資料視覺化

Abstract

Big data processing focuses on the three dimension (Volume, Velocity, and Variety, 3V for short), and the usage of business intelligence analysis is to enhance business value. The White House of United States announced the ideas of Big Data Initiative in March 2012 to study the technology of novel storage, management, and analysis in dealing with huge amounts of data, and visual analytics is just the focus of R & D projects in the U.S. government. Data visualization is one of the effective ways to understand for the human brain, and the tag cloud (word cloud) is the common form of data visualization among many presentations. These ways are to collect word and count their frequency of occurrence, and then show the most popular keywords by the font size within a specific period. The higher frequency will be represented by larger word.

Data visualization is a very interesting technology that applies special computational mode and algorithm to transfer all kinds of data into various charts and images; that is, people can understand insights of data more easily. Therefore, this paper will discuss different types of data to render, and researchers can classify data by numerical data (quantitative data) and categorical data (qualitative data) according to their characteristics. By completing the classification of the data type, we map the data and its visible attributes (data coding process) to decide what kind of attributes to express the data type is the most efficient. These presentations include graphics rendering (2D and 3D), real-time reporting tools, dynamic dashboard, and data visualization tools such as simulation. Finally, we list the advantages and disadvantages of these different tools and their evaluation.

Keyword：Big Data, Data Visualizations

目 次

巨量資料視覺化之研究

第一章、前言

隨著科技的發展，資料的數量也跟著大增，而這些數量大到傳統資料庫幾乎無法處理的資料便被定義為 Big Data，也就是巨量資料；巨量資料與傳統資料的差異在於三個層面(3V)：巨量(volume)、即時性(velocity)及多樣性(variety)，即巨量資料不僅數量龐大且不限制在單純的結構化資料上，半結構化和非結構化資料也是巨量資料的一部份，同時資料還具有時效性，需要快速的處理，因此如何達到高效率且高速的處理巨量資料是一個重要議題。

視覺化技術是重要的數據處理後的呈現方法[1][2]：企業上的資料時常以一筆筆紀錄的方式大量地進行儲存，分析資料時單靠人眼無法從如此龐大的紀錄中找出隱藏的知識與脈絡，所以分析師會以企業想要找出的知識為目標對這些資料進行分類，並透過各種視覺化的呈現方式使得一般人員也能夠從不同的圖形中以不同的分析角度找出不同的隱藏資訊。

國網中心發表於〔iThome〕全球企業迎接「資料視覺化」世紀來臨的文章中[8]，提到分析一個科學視算的專案有三個流程，包含資料的產生、資料的轉換與視覺化的呈現。

- 資料的產生：對於企業所提供的資料，需挖掘與分析其隱含在數據背後的知識。
- 資料的轉換：必須與科學視算的專家群進行深入討論處理細節
- 視覺化的呈現：瞭解轉換細節並進行相關程式的撰寫，並與該企業密集討論，確保最後視覺呈現可符合最初討論之目標，而將之轉化為有用知識。

製作一個高品質的資料視覺化呈現方式[7]時必須考慮到四點：

- 深度：把專家的角度融入資料呈現中，使大眾能用專家的角度，更深的理解資訊內容
- 極簡：只呈現有用的資料，以及必要的符號及文字，去除雜訊
- 好用：好的資料呈現會說故事，用圖像方式正確表達出知識，更深的理解資訊的內容
- 忠實：能夠呈現資料的正確程度

分析師在進行資料視覺化時有幾個步驟，首先研究人員會依據資料的特性，分為數值資料(numerical data)或數量資料(quantitative data)、類別資料 (categorical data)或定義資料(qualitative data)，細節可參考機率與統計概論一書[5]，接著將資料對應到視覺屬性(也就是資料編碼)，藉此完成資料類型的分類，決定哪一種視覺屬性來表達資料類型是最有效率的。

目前於網路上已提供多種 Open source 的視覺化工具可使用，我們將對於一些常見的視覺化工具進行效能評比，並列出各自的優缺點與適用場合。本論文後續內容如下，第二章將說明我們整個巨量資料視覺化的流程，第三章介紹常見的視覺化呈現圖形的優缺點，第四章列出目前較常被使用的視覺化工具，接著第五章我們將以一份車隊資料來實作三種視覺化工具，並比較使用的視覺化工具產生之圖形的優劣，接著第六章將對於幾種視覺化工具進行一些比較，並以表格的方式列從而看出各種工具的適用場合，最後第七章為結論。

第二章、視覺化步驟

進行一個巨量資料視覺化的分析主要為四個階段：資料前處理、資料分類、對映視覺屬性、挑選視覺化工具。

巨量資料不只數量龐大，其中也可能包含非結構化資料，我們必須先將其進行前處理，也就是解析資料的內容並將有用的部分進行結構化儲存以方便後續的應用；資料前處理進行完後，我們將資料分類為四種類型：數值資料、數量資料、類別資料與定義資料，將資料分類完後，我們將依照資料的類型對

映到適合的呈現圖形上[3][4]，決定最終將產出那些圖形；確定完呈現的圖形後，我們必須選擇適合的視覺化工具來產出我們的圖形。

第三章、視覺化圖形評比

底下將介紹三種常見的視覺化圖型種類與其優點，並根據其特性做適當的資料處理的範例呈現：

一、Table

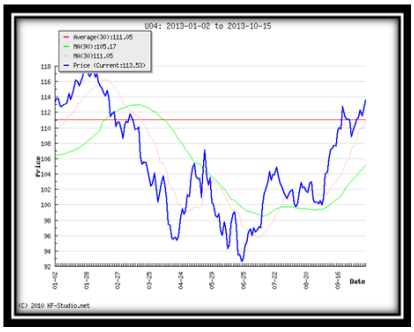
以表格方式記錄或處理數據，參考表一為例，以規律性排列的方式，簡單呈現出各類資料的時間、種類、數據等。由於兼容性佳，其數據的呈現簡潔而規律，可應用在動態網頁 HTML 的系統建置中，使代碼易處理與呈現。多維表格更能以層級結構呈現各資料的相關性。

表一表格範例

降雨微粒水平		
日期	每日降雨量	微粒
2007/1/1	4.1	122
2007/2/1	4.3	117
2007/3/1	5.7	112
2007/4/1	5.4	114
2007/5/1	5.9	110
2007/6/1	5	114
2007/7/1	3.6	128
2007/8/1	1.9	137
2007/9/1	7.3	104

二、Graph

由數個點（ vertex ）以及數條邊（ edge ）所構成，可以曲線或直線顯示資料的趨勢，一般用來顯示波動程度。與 Chart 不同的是，若要顯示出完整波形曲線資料，數據資料必須以陣列的方式輸入。可以 *f(x)* 函數製作出函數圖形，如下圖一所示。

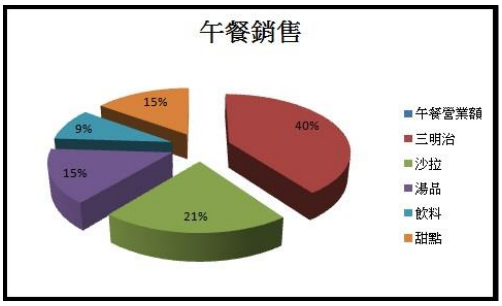


圖一股票走勢圖

Source: Retrieved OCT 18, 2013, from http://fund.kf-studio.net/fund_details.php?aHdgcGJhdmE

三、Chart

Chart 擁有資料保留的特性，能於畫面中新增資料點，以顯示歷史趨勢。以視覺化圖形呈現統計數據中的變化性。通過 Chart，能在看到歷史數據的情況下，有關新數據的分布情況。一般來說可以分為圖表及地圖等呈現方式，如下圖二所示。



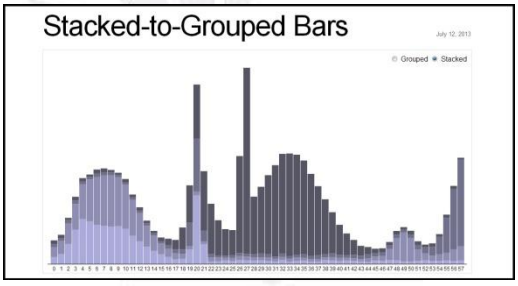
圖二圓餅圖

第四章、視覺化工具

底下將介紹常見的視覺化網站工具的種類與其優缺點，並根據其特性做適當的資料處理的適用場合與範例呈現：

一、D3(Data-Driven Documents)

D3 是一個 JavaScript 資料庫，可用來操作數據文件。D3 沒有使用新的表示法，而數據的呈現使用符合 Web 標準：HTML，SVG 和 CSS，如下圖三所示。



圖三 D3 圖表範例

Source: Retrieved JUL 17, 2013, from http://bl.ocks.org/mbostock/3943967

二、Google chart API

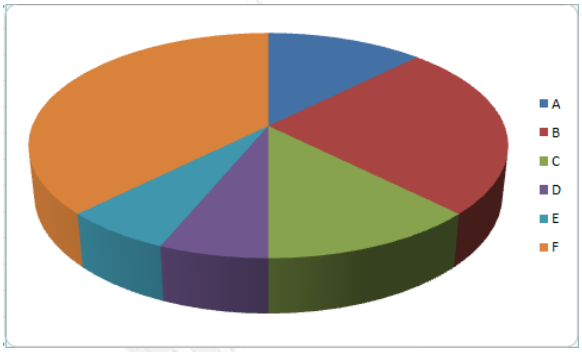
Google Chart API 是 Google 提供的線上製作圖表的工具，它可以讓使用者動態產生圖表。Google Chart API 可以繪製非常多種圖表，且語法較簡單，假如不熟悉語法，也可以使用 Chart Wizard 來快速建立圖表，Google Chart API 所提供的圖表種類如下圖四所示。

圖四 Google 所提供的範例圖表

Source: Retrieved JUL 17, 2013, from http://coopermaa2nd.blogspot.tw/2011/01/google-chart-api.html

三、Excel

Excel 是微軟公司所開發的 office 系列工具之一，Excel 是最基本的視覺化工具，可以很容易的將資料轉換成常見的圖形呈現，如下圖五所示。

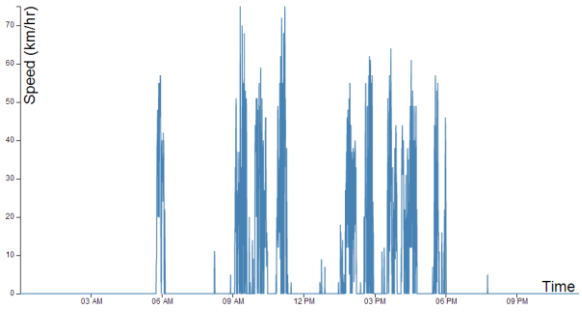


圖五 Excel 圖表範例

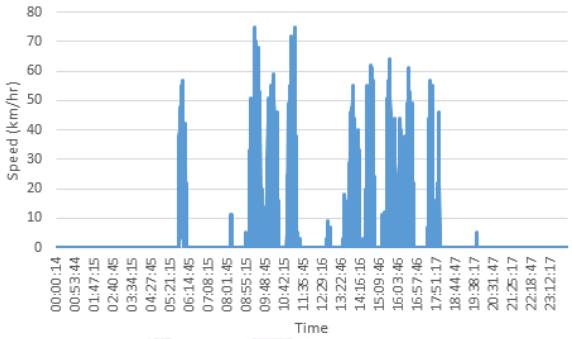
第五章、視覺化呈現與效果評比

本章節我們使用一份現有逢甲大學地理資訊系統研究中心所服務的“車隊資料”，資料來源為 2013 年 10 月部分資料，來進行視覺化呈現，這份資料是由散佈於屏東市區約一千五百台汽車的 sensor 資料，在一天內各不同時間點的位置、車速，我們將利用時間與車速的資料以折線圖呈現，再以時間和位置在地圖上呈現出行駛路徑。

在折線圖的呈現上我們選擇以 D3 與 Excel 兩種工具來呈現，呈現的結果如圖六、圖七所示，Excel 的折線圖上所呈現的細部資訊比起 D3 所呈現的折線圖更為完整，但並不代表 D3 的呈現效果一定比 Excel 來的差，因為 Excel 的視覺化功能是制式化的，使用者無法進行太大的更改，而 D3 則能較為自由的對程式進行修改已達到不同的呈現效果。

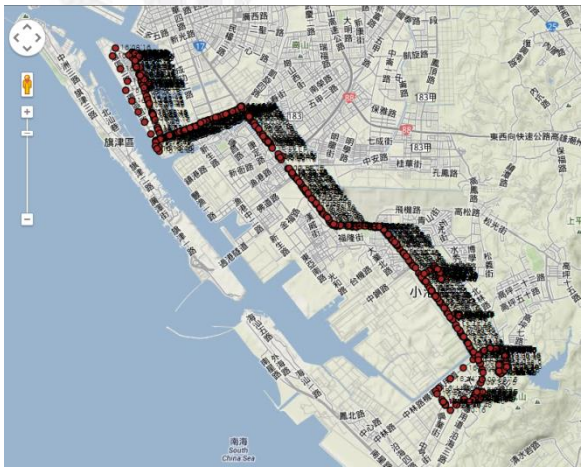


圖六 D3 折線圖(x 軸：時間，y 軸：速度)

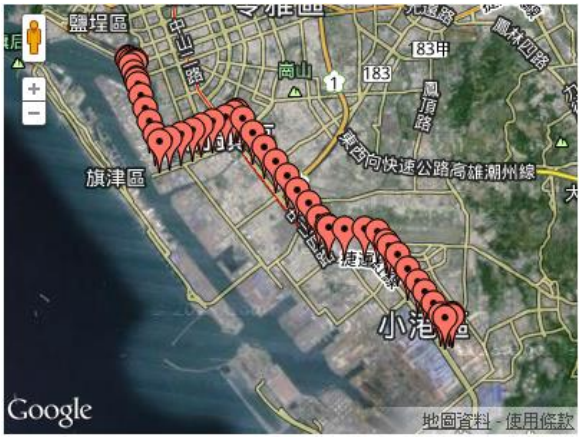


圖七 Excel 折線圖(x 軸：時間，y 軸：速度)

以地圖做為呈現方式的方法上我們選擇的是 D3 和 Google chart API。圖八所呈現的是 GoogleMap API 與 D3 搭配使用的方法，先以 Google Map API 產生地圖後再由 D3 繪製上標記從而形成這輛汽車一天中的行駛路徑；圖九使用的工具是 Google Chart API，在呈現的效果上與圖八差異不大。



圖八 D3 & Google map 呈現汽車路徑



圖九 Google Chart API map 呈現汽車路徑

表二視覺化工具適用場和比較表

Tools	UIs	Category	Platform	Graphics Category
D3	No	Library	Code editor &browser	Multiple
Google chart API	YES	Library& Visualization app/service	Code editor &browser	Normal
Excel	YES	Visualization service	Windows	Normal
Many Eyes	YES	Visualization app/service	Browser	Multiple
Datawrapper	YES	Visualization app/service	Browser	Normal
Flot	No	Library	Code editor &browser	Normal
Raphaël	No	Library	Code editor & browser	Normal
Leaflet	No	Library	Code editor & browser	Only Map

第六章、視覺化工具適用場合

本章節將列出數種視覺化工具的適用場合(表二)，我們考慮四個項目分別為使用者介面(UI)、工具種類(Category)、支援平台(Platform)與圖形種類(Graphics Category)，部分結果整理自線上資料分析視覺化工具線上資源 [6]。

現代的視覺化工具多數以函式庫的方式提供基礎服務，但對於非程式設計師的使用者來說是非常困難的，因此部份視覺化工具會提供使用者介面來幫助非程式設計師的使用者；支援平台是很重要的一項因素，使用者必須考慮到視覺化呈現的環境來挑選適合的工具，而挑選的工具提供的圖形種類越多，我們越能找出最適合資料呈現的方式。

現在越來越多的模組結合 Hadoop 生態系統，而針對視覺化未來在巨量資料上的呈現，我們認為有幾點是可以繼續發展的: 1. 發展出開放式 Hadoop 生態系統的資料視覺化展示引擎，利用 HDFS 的儲存資料集，產製多維度的圖表; 2. 從眾多圖表中自動化根據資料特性，推薦合適的數個圖表讓決策者選擇，最後輸出圖表資料; 3. 與 Hadoop 生態系統密切的整合與介接，並整合 eclipse 系統開發工具，讓有志一同開發者站在此基準上，繼續擴充相關圖表模組，成果分享於技術交流社群相關網站上。

第七章、結論與未來工作

巨量資料分析的視覺化分析是一門新興的領域，許多新的技術一直被開發出來，也應用在新的領域。在現在資訊爆炸的時代，會是一個協助分析與決策的好工具。一位會根據數據圖表說故事的專家，涉及多學科領域，包括統計學、資訊檢索、資料庫技術、機器學習、模式識別、知識庫建置、和資料視覺化等技術。本研究透過實際的運用視覺化工具呈現圖形的實作過程得知幾種常用的視覺化工具之優缺點與適用場合，藉以提供未來進行視覺化分析時能更快速且呈現出更高品質的視覺化圖形；未來的研究將著眼在如何提高資料前處理的效率，透過自動化的分析非結構化資料來減少時間成本，使得巨量資料能在第一時間被處理，達到巨量資料的最高價值。

致謝

本研究經費由國科會計畫 NSC101-2119-M-035-001, NSC101-2625-M-035-002, NSC101-2119-M-035-003 所提供，特此致謝。

參考文獻

[1] Murray, Scott.*Interactive Data Visualization for the Web*. O'Reilly Media, 2013.

[2] Ware, Colin.*Information visualization: perception for design*. Elsevier, 2012.

[3] Gray, Jonathan, Lucy Chambers, and Liliana Bounegru.*The data journalism handbook*. O'Reilly, 2012.

[4] Steele, Julie, and Noah Iliinsky.*Beautiful visualization*. O'Reilly Media, Inc., 2010.

[5] Mendenhall, Beaver, and Beaver, *Introduction to Probability and Statistics*, Fourteenth Edition, Brooks/Cole, Cengage Learning, 2013.

[6] Leonard Murphy.(2013,August 4).*50 New Tools Democratizing Data Analysis & Visualization*[Online]. Available:http://www.greenbookblog.org/2013/08/04/50-new-tools-democratizing-data-analysis-visualization/

[7] 康仕仲. (2012). *決策者的偏頭痛-談資料視覺化* [Online].Available:https://docs.google.com/presentation/d/1B2DFcFKTK4mCQOz9iwBJX_We9Vq4bd8CpNVCGWbNaRo/edit#slide=id.p

[8] 全球企業迎接「資料視覺化」世界來臨 iThome NO. 370 期刊 [Online]Available:http://www.nchc.org.tw/tw/about/publication/special_scientific/2008ithome_2.php