2014 數位科技與創新管理研討會
論文編號：35
數位科技組

海量資料研究之智識結構

陳宗天　國立台北大學資訊管理所
鄭宇傑　國立台北大學資訊管理所碩士班

摘　　要

海量資料為 IBM 於 2010 年所提出的新名詞，其於未來扮演著資訊處理領域之關鍵角色，本研究將以引文分析和智識領域視覺化作為基礎，透過技術專家開發的智識建構系統進行「海量資料」議題的因素分析，並產生智識結構圖以表達此議題與相關議題之關聯性，其中列出 19 項「海量資料」的因素分析項目並進行深入探討，描繪因素間之相互關係及延伸應用，並提出「海量資料」未來相關熱門議題趨勢以及建議研究方向，縮短研究人員以往搜尋相關研究領域文獻所花的時間與心血，以兼具效率和效果的方式取代之並輔助研究人員達到理想的研究目標。

關鍵詞：海量資料、智識建構、因素分析

Intellectual Structure of Big Data Research

Tsung Teng Chen, Graduate Institute of Information Management, National Taipei University

Yu Jie Jheng, Graduate Institute of Information Management, National Taipei University

Abstract

Big Data is a new term coined by IBM in 2010 years, it plays a key role in the field of information processing. This study utilizes bibliometrics method to derive the Intellectual Structure of the Big Data research filed. We analyze Big Data related-studies through Intellectual Structure System (ISS), and generate intellectual structure map to show the relationships among popular research themes in the area of Big Data. In this study, the research themes of Big Data are represented by 19 factors, which are discussed separately. We explained the relationship between the factors and discussed their implications. We proposed future research directions based on the analysis to facilitate researchers to have an overall understanding of Big Data research.

Keywords: Big Data, Intellectual Structure, Factor Analysis

1.導論

由於近年來的快速時代變遷，資訊科技的進展好比火箭升空的速度般不斷淘汰舊產品與發展新科技，伴隨的是資訊不斷擴增而產生資訊爆炸時代的來臨，於是「海量資料」的概念由此而生。本研究有鑒於未來的資訊爆炸現象所伴隨著海量資料處理研究需求量大增，進而有潛力成為未來幾十年間的熱門議題，所以本研究欲透過智識建構系統分析「海量資料」的相關議題，主要是透過相關的因素分析方法中找出相關議題，並且根據智識建構圖所產出的議題分佈狀況描繪出各個相關議題之間的分佈關聯性，最終本研究會根據以上分析呈現出目前「海量資料」議題主要關聯性為何，提供未來學者「海量資料」研究發展的具體參考方向與指標。

2.文獻探討

2.1 海量資料

海量資料又稱海量數據、巨量資料〔3〕，為 IBM 於 2010 年所提出的新名詞〔10〕，指的是其所涉及的資料量規模巨大到無法透過人工去統整分析，並在合理的時間內完成資料的輸入、儲存、處理和輸出等等程序。網路的搜尋、交易、輸入都是海量資料的基本元素，透過電腦運用海量資料的運算技術作篩選、整理、分析，所得出的結果不僅克服了過去對於大量複雜運算的障礙並且獲得簡單且客觀的結論。線上科技網站 ZDNET 資深記者 Dan 提到海量資料已逐漸成為世界企業內部 IT 處理大量資料的關鍵技術，不僅提供企業管理者關鍵資訊以利其經營決策，蒐集的相關資料還可以被規劃，創造更多意想不到的附加價值〔9〕，海量資料的主要特點可分為:(1)資料量大 (Volume)、(2)輸入處理速度快 (Velocity)、(3)資料多樣性 (Variety)、(4)價值密度低 (Veracity) 〔10〕。

資料量的大小決定了海量資料的應用價值，由下圖 1 可以明顯發現當資料量超過 Terabytes 時是普通資料處理技術無法負荷的，必須透過海量資料技術來完成，圖中列舉了各項海量資料所需要的關鍵要素與應用實例，清楚呈現海量資料所支援的相關應用與輔助學者對於未來海量資料延伸領域的發掘與開發。

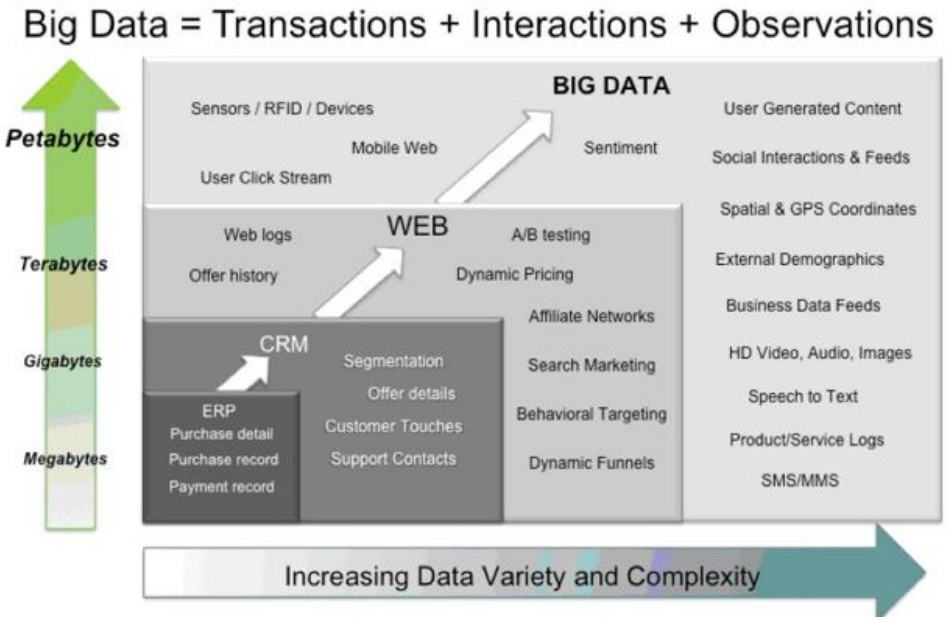

圖 1 : Big Data 發展流程示意圖

(資料來源:Contents of above graphic created in partnership with Teradata, Inc)

Hadoop 為實現海量資料運算技術的實際應用，其創辦人 Doug Cutting 提到其根據 Google 搜尋引擎的相關研究為藍圖，企圖以分散式運算技術為基礎，建立一套如何透過儲存、處理、分析 TB(Tera Bytes)甚至 PB(Peta Bytes)資料量大小的處理方法〔8〕。

Hadoop 是透過數台伺服器連接進行同步平行分散式計算而達到處理海量資料運算的技術應用，另外其可隨著企業的不同需求動態調整任一伺服器設備以達到不同的運算規模需求〔16〕。Hadoop 最大的特色為其本身是由 100% Java 程式語言所撰寫而成的開放原始碼資源，執行 Hadoop 平台時無需透過昂貴的軟體平台，只需使用一般的伺服器群合併達到平行資料處理與分析的目的。目前台灣國網中心提供免費的公用實驗叢集 Hadoop 雲端運算平台，使用者不需自行架設伺服器即可透過 Hadoop 雲端服務實現海量資料分散式運算技術。

處理海量資料最知名的技術為 LDA(Latent Dirchlet Allocation)，為 Blei 等人於 2003 年提出的新技術〔7〕，其為機率理論與圖形理論所產生的模型，主要對於海量資料分群與模組化提供一個可靠的架構，其可透過機率與統計理論的運算規則針對欲分析的海量資料進行機率分佈之相關計算並利用其作相關資料剖析與分群應用。其可應用於資訊檢索〔4〕、語言模型的調整〔13〕與機器學習〔7〕等領域。其藉由擺脫了傳統習慣採用向量空間表示法而採用以訓練模式為主的資訊擷取技術〔12〕，LDA 最大優點在於欲對新文獻群集進行分析時可透過先前的模型參數直接推估出新群集分佈的機率模型。

2.2 引文分析

引文分析是以網路中的鏈結為基礎並呈現，將要分析的單位視作一個節點，並且於每個節點間建立雙邊的鏈結關係，以代表雙邊關係的強度。該領域中的文章彼此之間引用與被引用的關係所結合成的網路圖則稱為主要研究議題的引文網路。如圖 2 所示為直接引用與文獻耦合及共引的相關概念。舉例而言，一篇完整的學術文獻必須具有正文以及相關引

用的參考書目列表，正文本身稱之為引用文獻，參考書目則稱之為被引用文獻，針對兩者關係進行相關研究可以獲取學術文獻間的發展關係與相關學術傳播過程，最終了解該主要議題之目前熱門議題以及未來發展趨勢。

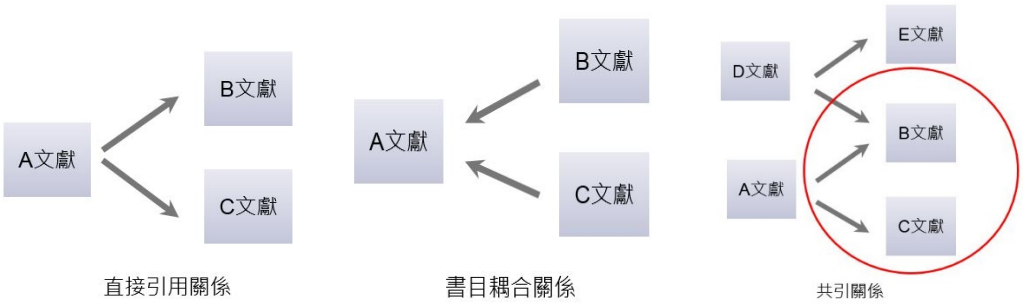

圖 2：引文分析示意圖

(資料來源 ： 本研究)

引用文獻在傳統上的觀念如下:A 文獻由於具有 B 文獻所需要的學術相關資訊，所以被 B 文獻所引用，例如 A 文獻具有更寬廣的研究內涵、描述所採用的方法或提供相關的數據或討論，故被 B 文獻所引用。A 文獻被引用時並不需具備提供完整參考資訊特性，只需提供 B 文獻關鍵性的學術需求即可，假設所有引用都相等的狀態下，文獻的被引用次數愈高代表其被往後研究相關議題的學者重視的程度越高，也就代表貢獻度越高，影響愈大；反之若文獻被引用次數愈低代表其被往後學者重視程度欲低，將面臨內容老化和不被重視的狀況。

2.3 智識領域視覺化

在大量資料下，對於欲瞭解該領域的學者來說，要能夠辨識理解該領域的重要議題是相當困難的，必須花費大量的時間和精神去完成，但是若透過智識視覺化的呈現方法將複雜的相關議題之文獻關聯作相對直觀呈現，使用者即可以針對圖形進行相當程度的分析來了解研究領域中的相關資訊，甚至可以提供該主題未來的趨勢發展走向，作為未來參考或延伸相關議題的主要方向。其中視覺化研究的主要目的是要藉由圖形或影像的視覺效果，來呈現大量複雜的資料或流程，可使資訊附載更多關鍵資訊並且更容易了解。通常人類對於視覺刺激及圖形理解程度遠勝於接受文字描述，圖形視覺刺激更能直覺性的聚焦想要表達的重點部份。

3.研究方法

本研究主要利用專家所開發的智識建構技術來釐清關於「海量資料」議題的主要研究領域範圍、分佈和主題。其中「智識建構」一詞代表在某一領域內所有相關議題的趨勢及在特定時間點內時間的分佈狀況，並透過智識建構系統所提供的相關議題分析結果中描繪出某一領域內所具有相當顯著性的議題代表。除此之外，智識建構亦可描繪出研究主題與相關議題之間的關聯性，並且透過圖形以資訊視覺化的概念具體實現，方便他人容易了解研究主題與相關議題關聯性之理解。

智識建構技術實現主要是衍生自共引網絡，其為透過研究領域中兩兩文獻之間共引關係的觀察進而推導出來的延伸概念，其中引用的行為決定了共引的誘導關係。關於文獻之間的引用關係，本研究運用微軟學術搜尋網(Microsoft Academic Search)鍵入關鍵字「海量資料」來查詢引用文獻資料庫，查詢結果顯示此議題共有 280 篇的種子文獻。下圖 3 為本研究方法流程圖，首先決定議題關鍵字後，透過這些種子文獻的蒐集，可以將其視為初始文獻查詢檢索目錄，利用種子文獻中引用的連結找出所有關於「海量資料」議題的相關文獻數量共 3848 篇，其中本研究蒐集的文獻為第一層鏈結(1-Level)內之所有文獻資料，且所有文獻都具有引用次數 1 次以上和共引次數超過 1 次的條件。簡潔的引文網絡可推導出共引網絡的相關資訊，本研究透過引用次數門檻值的篩選和共引次數門檻值的設定後，過濾出 543 篇文獻數，並經由共引網絡、因素分析和文獻繪圖檔(PFNET)以產生智識結構相關資訊，智識結構圖則是利用圖形繪製法顯示於 PFNET 圖形中。

因素分析的目的主要是(1)減少變異數(2)檢測結構中的變數關係，故 Stevens(1999)學者表示因素分析可以做為資料壓縮及結構檢測的方法或延伸應用〔15〕。本研究利用因素分析搭配上文獻的相關變量產生出指定議題研究之若干因素群集，以本研究「海量資料」議題為例，其 543 篇文獻區分成 19 個不同的因素群集，舉例來說此議題的第一個因素集群就分類了 82 篇的相關文獻。

因素分析方法首先採用最顯著因子做為基準，並依序列出後續擁有較少顯著關係的因子形成共引網絡，進而以因素分析法作處理。因素分析還產生了 Pearson 相關係數矩陣，用來儲存文件之間的相關性測量，其中學者 Schvaneveldt(1990)將其定標，然後透過相關矩陣的探索與應用中，產生出 PFNET 可以描繪出基礎結構概念圖，智識建構系統將多步驟的重要流程電腦化，成功建置出文獻探討輔助系統的智識建構推導流程〔14〕。

為了保留被引用次數較高的重要文獻並使分析文獻數量在可管理的範圍內，本研究將除了直接引用次數門檻值設定為1之外，並將共引次數門檻值設定為1，過濾後得出543篇文獻。下表1列出智識建構系統因素設定相關資訊，其中本研究設定文獻中對各因素負載(Factor Loading)大於0.6的文章作留存，以此做為依據並預設將此議題分類為20個因素來分析研究並繪出其智識結構圖，使研究者能夠更清楚瞭解文獻彼此的關聯性。

表 1: 智識建構系統因素設定資訊 (資料來源 : 本研究)

Factors

Rotation_MaximumIterations	25
Factor Number	20
Factor Loading	0.6

圖 3：研究方法流程圖

(資料來源：本研究)

4.研究結果

4.1 因素分析之變量結果統計

智識建構系統所產生的「海量資料」因素分析結果之項目為 19 項，下表 2 為 19 項海量資料因素分析之相關議題、可解釋變異量和累積變異量，其中 19 項的累積變異量高達 88.09%，證明此因素分析項目群與「海量資料」議題具有高度的相關性。

表 2：因素分析變量結果統計表

(資料來源：本研究)

因　　素	海量資料相關議題	可解釋變異量(%)	累積變異量(%)
1	社群網路　(Social Network)	15.82	15.82
2	視覺化　(Visualization)	8.484	24.31
3	分散式運算　(Distributed Computing)	7.321	31.63
4	序列分析　(Sequence Analysis)	6.882	38.51
5	數目變異性　(Copy Number Variants)	5.561	44.07
6	點對點系統　(Peer to Peer System)	4.593	48.66
7	區隔元素　(Segmentation)	4.190	52.85
8	網路　(Network)	3.886	56.74
9	資料庫系統　(Database System)	3.749	60.49
10	檔案系統　(File System)	3.506	63.99
11	結構化查詢語言 (Structured Query Language)	3.314	67.31
12	自組織映射圖 (Self-Organizing-Map)	3.189	70.50

13	因素分析 (Factor Analysis)	3.189	73.69
14	統計方法 (Statistical Method)	3.050	76.74
15	圖形學 (Graphics)	2.857	79.59
16	極小資訊 (Minimum-Information)	2.509	82.10
17	群集分析 (Clustering Analysis)	2.185	84.29
18	圖片搜尋分析(Image Search Analysis)	2.029	86.31
19	資料集群 (Data Clustering)	1.773	88.09

4.2 個別因素探討

因素一 社群網路 (Social Network)

社群網路為知名的電子商務應用議題，社群透過網路的連結可以使社群溝通變得更為頻繁〔2〕，其中社群網路是建立在數個互相鏈結的節點群並透過多個設備模擬虛擬平台架構，為在虛擬的網路中產生虛擬社會群體的概念。伴隨著社群網路的運作，大量資料如影像、圖像及文字等等於雲端平台互相傳遞，需要的是海量資料相關運算技術支援，是目前最常見的海量資料運用。

因素二 視覺化 (Visualization)

資訊視覺化利用圖像的直觀傳達功能用來表達在科學與社會研究中所發現大量的抽象資訊事件所蘊含的意義，包含前後因果之間的相依關係 提供人們易於理解的解釋資訊。視覺的形象具有吸引注意力、喚起想像、導致聯想、加深印象、易於記憶、引起感情共鳴等功能，加上視覺感官比其他所有感覺都來的有影響力，所以視覺化仍是系統設計者對於傳達資訊方法上首要選擇的方式之一。海量資料的技術困難點就在於如何有效處理輸入、處理、輸出等三大流程，當人們需要針對海量資料進行更深一步的了解時就可以透過視覺化的圖形呈現以傳達相對較易於理解的資訊給需求者。

因素三 分散式運算 (Distributed Computing)

分散式運算的概念是將大量運算內容分割成多個不同的工作區塊，再由多台伺服器分工計算，當運算結束後將資料彙整至服務端進行整合並得出數據理論，目前主流的分散式平台為加州柏克萊大學所發展出來的柏克利開放式網路運算平台(Berkeley Open Infrastructure for Network Computing, BOINC)，故分散式運算可被視為透過大眾合作力量完成對海量資料複雜運算的概念技術。

因素四 序列分析 (Sequence Analysis)

序列分析是一種行為分析的工具,當研究者做研究記錄的時候，其可以將「事件」進行「編碼」〔5〕，然後依照時間順序排列，就可以得到一連串的事件序列觀察樣本。海量資料可以透過序列分析工具進行更進一步的事件分析，包含透過不同維度觀點進行不同需求的相關研究，並且針對其前後關係推測資料間具有的關聯性、時間性或排序性。

因素五 數目變異性 (Copy Number Variants)

拷貝數變異又稱為拷貝數多型性(Copy Number Polymorphisms, CNPs)，是刪除、插入、覆寫，以及複雜多位置變異(Complex Multi-Site Variants)的合稱，在所有人類以及其他已測試的哺乳動物中皆可發現。基因組拷貝數變異是基因組變異的一種形式，通常使

基因組中大片段的 DNA 形成非正常的拷貝數量。例如人類正常染色體拷貝數是 2，有些染色體區域拷貝數變成 1 或 3，該區域發生拷貝數缺失或增加則位於該區域內的基因表達量也會受到影響。

因素六　點對點系統 (Peer to Peer system)

對等網路（Peer-to-Peer,P2P），又稱作點對點技術，是無中心伺服器且依靠使用者群（peers）交換資訊的網際網路體系。點對點系統的特點為存在於對等網路中的每個使用者端既是節點，也有伺服器的功能，任一節點無法直接找到其他節點，必須依靠其用戶群進行資訊交流。P2P 節點能遍佈整個網際網路，也包括開發者在內的任何人、組織或政府組織，點對點系統在網路隱私要求高和注重檔案分享的網路時代中，存在許多廣泛的應用。

因素七　區隔元素 (Segmentation)

區隔元素經常被用來當作區分領域之用，舉例來說：確認目標市場後,行銷人員必須以策略行銷(Strategic Marketing)發展行銷目標與計劃,策略行銷的精髓是運用 STP 公式,選定目標市場進行區隔或是區隔市場再訂定目標對象。海量資料即需要透過定義明確的區隔元素將大量運算資料進行分類進而優先針對順位較高的資料進行分析以提高數據分析之成效。

因素八　網路 (Network)

網路意指將一部以上之電腦周邊或資訊系統透過傳輸媒介連結在一起，並且能彼此交換訊息、資料與共享資源，即為電腦網路(Computer Network)；區域網路(Local Area Network)意指以上述方式於短距離區域內(通常為 2km 範圍內)連結而成的電腦網路，而廣域網路(Wide Area Network)則意指在長距離或透過公眾電信網路將區域網路予以連結，因此 LAN-WAN 之連結我們便稱為網際連結(Internetworking)。以下為區域網路主要的三大功能性連結，分別為(1)連結性(Connectivity)、(2)網際連結(Internetworking)、(3)互動連結性(Interoperability) 。

因素九　資料庫系統 (Database System)

簡單來說資料庫系統可以被視為電子化的文件櫃，用於儲存電子文件的位址，其扮演的角色為處理與儲存資料，尤其是大型資料庫更是充分發揮了資料整合與資料共享的特色，資料庫系統的優勢則為(1)避免資料重複、(2)避免資料不一致、(3)資料共享、(4)提供異動管理、(5)建立資料標準、(6)確保資料安全性〔1〕。資料庫指的是以一定方式儲存在一起、能為多個用戶共享、具有盡可能小的冗餘度、與應用程序彼此獨立的數據集合。資料庫系統是海量資料技術的基礎設施之一，支援其能夠備份及儲存最新的修改資訊。

因素十　檔案系統 (File System)

檔案系統為一種儲存和組織電腦資料的實際應用，其讓使用者存取和尋找資料變得十分便利，檔案系統使用檔案和樹狀目錄的抽象邏輯概念代替了硬碟和光碟等物理裝置必須使用資料儲存區塊的傳統概念，使用者不必關心資料實際保存在硬碟的位址，只需記住檔案所屬的目錄和檔案名。嚴格地說，檔案系統是一套實作了資料的儲存、分級組織、存取和獲取等操作的抽象資料型別。

因素十一　結構化查詢語言 (Structured Query Language)

SQL 全名是結構化查詢語言，為用於資料庫中的標準數據查詢語言，IBM 公司最早將 SQL 語言導入至其開發的資料庫系統中。美國國家標準學會對 SQL 語言進行規範後以此作為關聯式資料庫管理系統的標準語言。不過各種通行的資料庫系統在其實踐過程中都對 SQL 規範作了某些編改和擴充。所以實際上不同資料庫系統之間的 SQL 並不能相互通用。當海量資料技術需透過資料庫系統進行存取動作時就必須透過結構化查詢語言進行實作。SQL 包含 3 部分：(1)資料定義語言(DDL : Data Definition Language)、(2)資料操縱語言(DML : Data Manipulation Language)、(3)資料控制語言(DCL : Data Control Language)。

因素十二　自組織映射圖 (Self-Organizing-Map)

自組織映射圖網路 (Self-Organizing Map Neural Network, SOMNN)為無監督式學習網路，它是由 Kohonen 於 1998 年提出〔11〕。所謂的無監督式學習，是從問題領域中取得僅有輸入變數值的訓練範例，從中學習範例的內在集群規則、資料分佈或相似性，以應用於新的輸入案例，據以推論此案例屬於哪個集群之應用。以下為自組織映射圖的特色：(1)有效地處理資訊、(2)加速對傳入訊息的辨識速度、(3)易於存取資訊、(4)易於系統的交互作用。

因素十三　因素分析 (Factor Analysis)

因素分析是一種用來簡化變項、分析變項間的主軸關係，尋找共同潛在構念的統計方法，亦即以少數幾個因素來解釋一群有相互關係存在的變數之模式，又能保有原有資料結構所提供的大部份資訊。因素分析假定樣本單位在某一變數上的反應（即觀察值）是由二個部分所組成：(1)各變數共同變異的部分，稱為共同因素（Common Factor）。(2)各變數所獨有的部分，稱為獨特因素（Unique Factor）。

因素十四　統計方法 (Statistical Method)

統計方法是以少量的資料(稱為樣本)所提供的資訊來推斷欲研究對象(稱為母體)特徵的方法。如何蒐集有價值的資料？如何組織、解釋所蒐集的資料？如何分析並因素十八　圖片搜尋分析 (Image Search Analysis)

網際網路的日漸發達導致網路上的文字、圖片和影片的流量大增，當使用者必須透過網路尋找相關的圖片資訊時就必須運用圖片搜尋分析技術來實現快速搜尋的功能。圖片搜尋分析是屬於海量資料運算的應用實例，相較於文字搜尋分析，圖片具有檔案相對複雜、龐大的特性，所以更需要海量運算技術的支援才可以實現圖片搜尋分析的技術。

因素十九　資料集群 (Data Clustering)

資料集群(Data Clustering)即是資料分群，看似與資料分類(Data Classification)類似，其實資料集群的主要目的是分群，即是將一整群的資料，依照某些樣本來分成個別群體，使得在各個群組中的資料與其樣本的相似性達到最大，而與其他樣本的相似性達到最小。

4.3 因素關係探討

下圖 4 為智識建構系統透過視覺化技術產生出來的因素分析集群智識建構圖，其中以藍色圓圈框起的部分為「海量資料」議題中前三大因素集群分佈，因素一的社群網路議題於圖中分佈非常集中，並沒有混雜任何其他因素集群；因素二的視覺化議題於圖中與其他

的因素集群交錯分佈，此現象代表視覺化議題時常被其他因素集群議題所涵蓋，屬於跨領域之因素集群議題；因素三的分散式運算議題於圖中分佈較為集中，且距離前兩大因素集群較遠，此現象代表分散式運算議題為較獨立領域，提供其他議題領域之基礎技術支援。下圖 5 為針對不同因素群集對於「海量資料議題」的核心重要程度作排序，愈接近核心的文獻代表愈是海量資料議題的核心重要文獻。

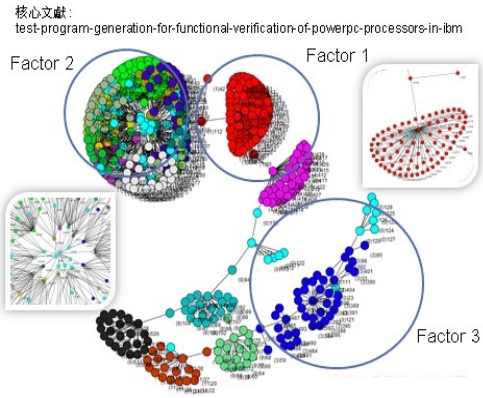

圖 4：知識管理領域智識建構圖

(資料來源 ： 本研究)

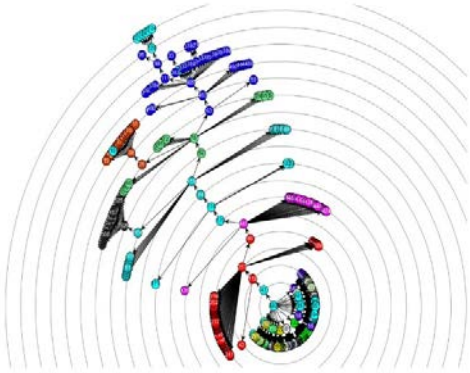

圖 5：逕向圖

(資料來源:本研究)

4.4 因素分析模型

本研究將根據智識建構系統產生的十九個因素集群作分析，針對其不同屬性與其他因素集群的關聯進行進一步了解，企圖建構出海量資料議題中細部相關熱門主題之關聯情況，並透過下圖 6 的因素分析模型作呈現，往後即可藉由此圖針對海量資料之相關研究趨勢或關鍵跨領域議題進行研究，省去大量探索相關議題的時間花費並藉由此因素分析模型構思出未來更有價值貢獻的跨議題應用。

海量資料的因素集群之熱門議題我們大致上可分為(1)資料來源、(2)資料儲存、(3)資料傳輸、(4)資料處理、(5)資料呈現，舉例來說：當社群網路、DNA 分析技術與自組織映射圖產生的海量資料如文字資料、排序資料與數位資料時，可以透過資料庫或是分散式檔案儲存系統以相關資料庫操作語言進行管控與保存，或是在網際網路的支援下透過點對點伺服器連接以分散式運算技術建立可傳輸海量資料的平台架構；資料處理則是先透過區隔元素根據海量資料的重要程度排序出資料處理的優先順序，再透過資料處理內部所提到的多種方法進行資料處理與運算，最後透過資訊視覺化與電腦圖形化技術將處理運算結果作呈現。

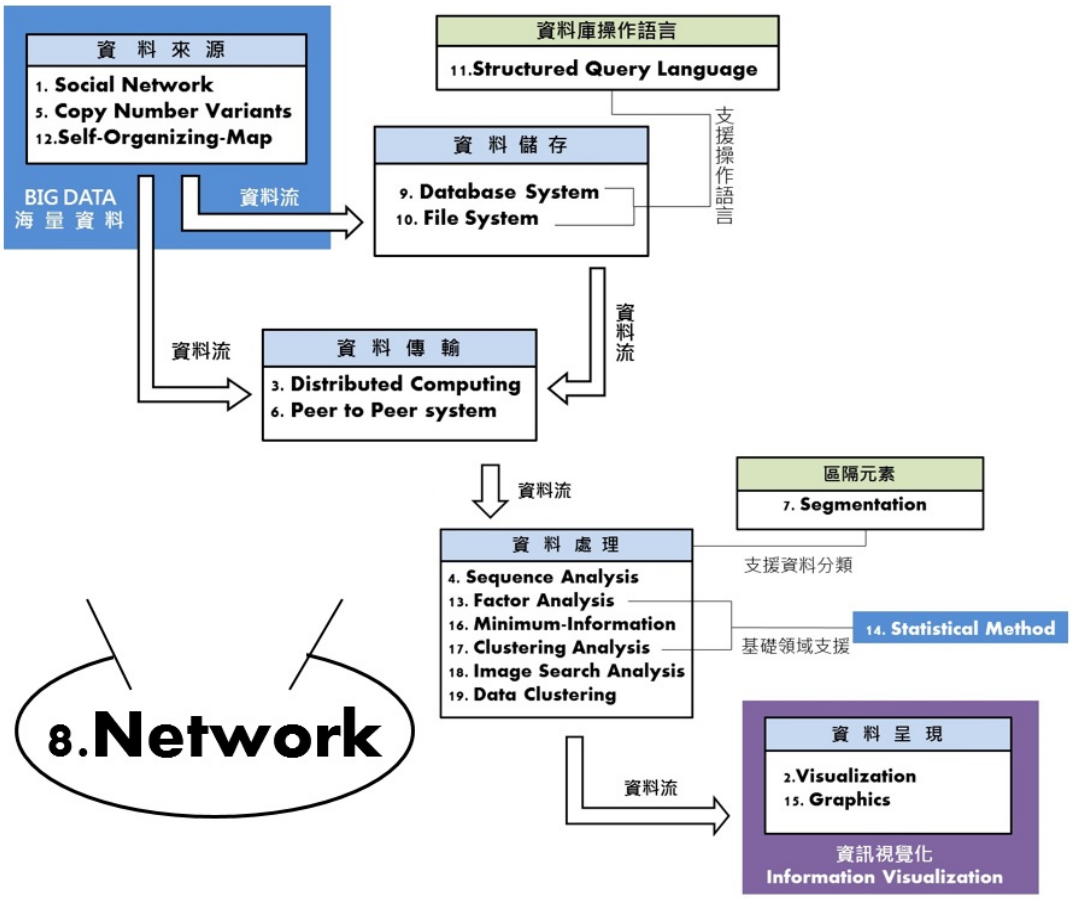

圖 6: 因素分析模型

(資料來源:本研究)

5.結論

從因素分析中我們可以了解到有關於「海量資料」議題目前研究現況及未來較有可能的發展走向，其中包含海量資料的處理平台、儲存技術語言、運算技術、分析技術、區隔概念、社群網路應用、DNA 序列分析等相關議題都有可能被視為未來「海量資料」的熱門研究目標，下列幾點為本研究推薦之未來針對海量資料議題之研究方向:

(1)資料優先處理順序之資料剖析應用

(2)大量學術文獻集群分類、因素分析與應用

(3)海量資料透過 LDA 機率統計模型進行分析、運算與應用

(4)優化海量資料儲存與傳輸流程以達一定之穩定性與效能

(5)海量資料分析結果呈現之資訊視覺化回饋與應用

以海量資料及資料儲存兩者跨領域議題為舉例，O'Reilly 線上科技論壇記者 Barry 就於報導中提到被公認能夠支援決策制定之資料倉儲技術如何面對海量資料時代的來臨並提出適應的辦法〔6〕，故學者針對此跨領域議題進行進一步的研究發現資料倉儲在處理海量資料時必須回歸根本的理念為建立企業資訊的一致性與信任，才能配合海量資料的運算提供最有效的決策制定支援。

研究議題與方向對於研究本體是至關重要的，故本研究期望提供未來的學者進行海量資料研究時依循正確的研究方向，直接切入其關鍵議題執行進一步的探討與研究，大大縮短了以往學者為了尋找相關議題文獻時所付出的大量時間，提供兼具效率及效果的方式使學者選擇符合其研究需求的相關領域進行研究。

6.參考文獻

1. 范士展，關係式資料庫:觀念解析實務大全，五南圖書出版股份有限公司，民國 93 年 6 月。
2. 孫治本(2005)，網路社群概論【線上資料】，(取得日期：2013 年 12 月 25 日)，國立交通大學通識教育中心人文社會學系副教授，民國 94 年 5 月，來源：http://web.it.nctu.edu.tw/~cpsun/sun-internet-community.pdf【2005,March】。
3. 維基百科，大資料【線上資料】，(取得日期：2013 年 11 月 23 日) ，民國 102 年，來源：http://zh.wikipedia.org/wiki/%E5%A4%A7%E6%95%B8%E6%93%9A【2013,December】。
4. Azzopardi, L., Girolami, M., Van Rijsbergen, C.J. “Topic based language models for ad hoc information retrieval”, Proceedings of the International Joint Conference on Neural Networks, pp. 3281-3286, 2004.
5. Bakeman, R., “Observing interaction: an introduction to sequential analysis.”, Cambridge; New York: Cambridge University Press, 1986.
6. Barry, D., “Will data warehousing survive the advent of big data”, O’Reilly Media, July 2011 (available online at http://strata.oreilly.com/2011/01/data-warehouse-big-data.html).
7. Blei, D. M., Ng, A. Y., & Jordan, M. I., “Latent dirichlet allocation”, J.Mach. Learn. Res., 3, 993-1022. doi: 10.1162/jmlr.2003.3.4-5.993, 2003.
8. Doug, H., “8 Big Data Deployments In Detail”, InformationWeek: Connecting The Business Technology Community, August 2010(available online at http://www.informationweek.com/database/8-big-data-deployments-in-detail/d/d-id/1091732).
9. Dan, K., “What is Big Data”, ZDNET Technology News Website, February 2010 (available online at http://www.zdnet.com/blog/virtualization/what-is-big-data/1708).
10. IBM, “Massive data for the enterprise give meaning”, IBM Official Website,June 2010 (available online at http://www-01.ibm.com/software/tw/data/bigdata/).
11. Kohonen, T., “The self-organizing map” Neurocomputing 21(1): 1-6, 1998.
12. Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., and Baldi, “P. Mining concepts from code with probabilistic topic models”, Proceedings of the twenty-second IEEE/ACM international conference on automated software engineering, November 05-09, 2007.
13. Salton, G., McGill, M. J., “Introduction to Modern Information Retrieval”, New York: McGraw-Hill, 1983.
14. Schvaneveldt, R. W., “Pathfinder associative networks : studies in knowledge organizations”, Norwood, N.J.: Ablex Pub. Corp, 1990.
15. Stevens, J., “Applied Multivariate Statistics for the Social Sciences”, New Jersey: Lawrence Erlbaum Associates, 1999

16. Shvachko, K., Kuang, H. et al., “The hadoop distributed file system. Mass Storage Systems and Technologies (MSST)”, 2010 IEEE 26th Symposium on, IEEE, 2010.