

大數據研究文獻計量與可視化分析

吳昭彥¹ 梁佳玲² 李健平³

¹義守大學資管系副教授

²陸軍官校資訊系副教授

³義守大學資工系

摘要

在現今資訊爆炸的年代，幾乎很多事情可用資訊與數字來說明，同時隨著目前科技與技術的蓬勃發展，使得資訊的蒐集更加容易且廣泛，藉由廣大訊息蒐集而形成大數據，因此人類可從資訊累積的巨量資料(大數據)中得到許多未知的資訊，並藉由文獻計量與可視化分析，更進一步的找出有價值的資訊，以利於後續決策者能迅速做出判斷。

本研究以「大數據」為關鍵字，並將條件限定為期刊論文，年限則以2003到2018年為止做為蒐集來源，以 WOS 資料庫5628篇文獻資料為主，將 WOS 資料庫收集的5628篇數據透過可視化軟體 Citespace 與 VOSviewer 進行分析比較。

結論得出大數據在近年來研究量下滑，藉由可視化圖形得知，大數據由早期以研究預測分析為主要思維，透過 machine learning 收集大量資料以及 data mining 與 classification 的輔助來找出未來可能的趨勢，中期則是透過 hadoop 與 R 等應用軟體結合雲端運算進行分析，晚期需要大數據的演算法以及成熟的硬體設備，能夠面對各種挑戰能產生十足的影響力。

關鍵字：大數據、big data、Citespace、VOSviewer、可視化分析

1. 緒論

1.1 研究背景與動機

現今社會是資訊爆炸的年代，幾乎很多事情可用資訊與數字來說明，當然每件事都是息息相關，所以可以藉由這些資料找出其中因果關係，並加以利用就能推測出想要的結果[16]，同時隨著目前科技與技術的蓬勃發展，使得資訊的蒐集更加容易且廣泛，並減少了以往需花費大量的人力與物力，藉由廣大訊息蒐集而形成大數據，將改變並運用至各產業的經營[6]，簡單來說，以石油為例，油價的漲跌不單只是金錢上的影響，間接影響到黃金價格的浮動，甚至連運輸業、航空業等需要用到石油的許多產業，也會造成影響，對我們生活上更有極大的影響[10]，資訊科技也是如此，因此我們才須透過分析大數據找出之間的因果。

大數據的分析涵蓋的領域廣泛，已經在不同的領域和研究中都有進展。研究領域如氣候數據分析 [20]，電子商務推薦系統 [5]，社交媒體分析 [22]，醫療數據分析 [23]，智慧城市 [19]等。

對大數據及其分析的研究需求，則是學術平台與工具，這些平台與工具可以利用這一大筆的數據庫來創立可以使整個研究受益的應用程序。如：數據收集、儲存、處理、分析與可視化，以支持研究文獻計量分析等用途[9]。

1.2 研究目的

隨著時代演進，人們需求科學快速發展、電腦科技的進步及分析運算能力增強，而資料存儲與蒐集每日持續在進步並加速累積存取，在面對每日龐大資料的流通，對於資料的應用，在方式上也隨時間而改變，因此人類可從資訊累積的巨量資料(大數據)中得到許多未知的資訊[7]，並藉由文獻計量與可視化分析，更進一步的找出有價值的資訊，以利於後續決策者能迅速做出判斷。

針對大數據在國際學術文獻分析及國內相關的學術文獻研究成果在質與量的差異，台灣目前少有以知識圖譜的文獻計量方式展現相關研究議題，因此，缺乏有效的對比參照，容易導致偏頗現象；故本研究利用文獻計量學的方法，進行 Web of Science(WOS)資料庫長期照護學術研究資料文獻探勘，使用 Citespace 及 VOSviewer 的模形化聚類技術展示分析 (Cluster analysis)，透過聚類算法 (Cluster) 技術與圖譜展示 (Mapping) 技術整合文獻數據。

本研究針對前四個議題，歸納探討的目的如下：

- 一、透過文獻計量方式來達到探討大數據研究領域、作者特性、參考文獻特性、期刊、國家的關聯性以及分布發展的時間。
- 二、根據分析結果，大數據領域書目計量分析比對具代表性之回顧文獻，且是否能對其回顧文獻補足人工不易或無

法察覺之處，給予相關研究者日後參考之依據。

三、最終根據回顧文獻與書目分析結果說明其研究限制及探討大數據未來發展之方向。

1.3 研究流程

本研究一共分為五章，分別為緒論、研究背景與動機、研究目的、研究流程、文獻探討、研究方法、可視化軟體的數據分析及結論與研究限制，如圖1所示。

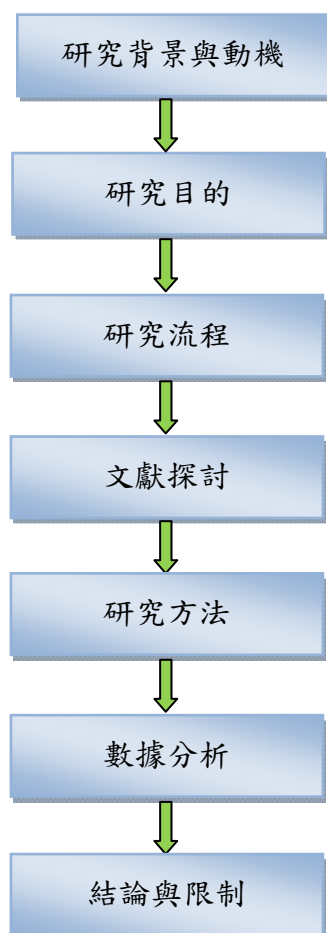


圖1.研究流程圖

2. 文獻探討

2.1 大數據的定義與發展

大數據在文獻中有很多不同的定義。其中有：大數據是能夠執行龐大的技術存儲、管理和有效處理能力[21]。

大數據是一個定義高科技、高速、大容量、複雜和多變量數據的術語，用於獲取、存儲、分發、管理和分析等訊息[24]。

大數據是高容量、高速度和高品種的訊息資產，需要用新的處理方式來呈現，如：加強決策、檢查發現和流程優化[14]& [15]。

大數據技術是新一代技術和架構，旨在通過提供高速獲取、發現和分析，有效地從多變量高容量數據中提取有效的價值[13]。

通過結合文獻中的各種定義來定義大數據如下：將新形式整合在一起以展現各種複雜和高容量數據中隱藏價值的方法和技術群集[17]。

大數據（Big Data）概念是1998年由SGI 首席科學家 John Masey 在 USENIX 大會上提出的。他當時發表了一篇名為 Big Data and the Next Wave of Infrastrass 的論文，使用了大數據來描述數據爆炸的現象。但大數據真正得到業界的關注，則是在其後多年的事情。其中大數據最重要的發酵因素則是2003-2006年 Google 發布的 GFS、MapReduce 和 BigTable 三篇論文[11]。

正如定義所示，在大數據集中時，需要考慮一些要點。例如：數據應該是複雜的，所以無論其大小數據都是重要的。因此傳統方法難以分析大數據，所以需要一套全新的方法和技術。

由於當前數據發展之迅速，人們總是希望將接觸到的訊息與手上的資訊作連結比對，進而產生新的資訊，藉此來實現需要的目的。

國外學者 Özköse, Ari, and Gencer[18]認為隨著大數據每年研究的日益增長，組織和政府越來越關注學術和技術文件的數字化，為數字資料庫的概念增添了新的面貌，而這種生成數據的數量、種類和速度滿足大數據定義；因此，這種學術儲備通常被稱為大數據，國外學者 Elshawi, Sakr, Talia, and Trunfio[12]認為大數據需要可擴展的架構來存儲和處理資訊，因此雲端運算會成為大數據的主流，除了有新的框架，還有相應的軟體資源讓這個領域持續發展。

2.2 書目計量學之發展

書目(文獻)計量學是以文獻體系和文獻計量特徵為研究對象，採用數學、統計學等計量方法，並進而探討科學技術的某些結構、特徵和規律的一門新興分支學科。當書目計量用於探討知識數量成長及其分布，可藉由文獻成長與文獻分布理論，預測社會知識行為模式，並預為規劃因應學科知識行為的發展趨勢。

書目計量學定義如下：以主題、文

獻、目錄為對象，利用數學統計和邏輯方法，對於所涵蓋的特質，作個別或集體的分析和分析[3]。

書目計量學的研究目的既然是在於提供知識結構與傳播的資訊，故其研究類型可分為「描述性書目計量學」(descriptive bibliometric)與「評估性書目計量學」(evaluative bibliometric)，描述性書目計量學重點在於描述文獻特性，用於生產力的計算，例如：描述分析某特定領域論文、圖書、期刊或其他類型資料文獻的數量分布及其研究內容特性；評估性書目計量學係利用參考書目或引文，計算文獻被引用的情形、檢視文件內涵之間的關係，或是瞭解文獻的學術影響力[1]。

3. 研究方法

3.1 資料的來源與篩選

本研究以「大數據」為關鍵字，並將條件限定為期刊論文，年限則以2018年為止做為蒐集來源，全部一共調查了IEEE、Wiley Online Library、SDOL、SpringerLink、ACM Digital Library、Taylor & Francis、EBSCO-BSC、WOS等8個資料庫。

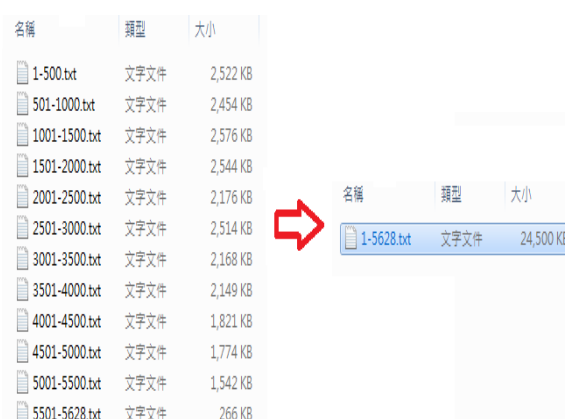
最終結果顯示，由IEEE資料庫6245篇為最高，但因無法提供書目計量程式可使用的格式，所以已由WOS資料庫5628篇的文獻為主，其中WOS(Web of Science Core Collection)涵蓋了Social Science Citation Index(SSCI)、Science Citation Index Expanded(SCI)等數個子資

料庫的資料，包含社會科學學科的3000多種期刊，其涵蓋的範圍是從2003年到2018年，數量總類也都相當齊全，所以才選用WOS 資料庫的5628篇作為本研究之分析數據，如表1所示。

表1.各個資料庫的數量與提供格式

資料庫	數量	提供程式格式
IEEE	6245	不提供
Wiley Online Library	1505	不提供
SDOL	1907	提供
SpringerLink	3579	不提供
ACM Digital Library	2151	不提供
Taylor & Francis	670	不提供
EBSCO-BSC	3056	不提供
WOS	5628	提供

本研究將 WOS 資料庫提供的書目資料進行下載，其中記錄內容選取「完整記錄和被引用參考文獻」，檔案格式為「純文字」(.txt)，因 WOS 下載有500筆之限制，所以將多個檔案進行整合，並使用VOSviewer 與 citespace 分析，如圖2所示。



名稱	類型	大小
1-500.txt	文字文件	2,522 KB
501-1000.txt	文字文件	2,454 KB
1001-1500.txt	文字文件	2,576 KB
1501-2000.txt	文字文件	2,544 KB
2001-2500.txt	文字文件	2,176 KB
2501-3000.txt	文字文件	2,514 KB
3001-3500.txt	文字文件	2,168 KB
3501-4000.txt	文字文件	2,149 KB
4001-4500.txt	文字文件	1,821 KB
4501-5000.txt	文字文件	1,774 KB
5001-5500.txt	文字文件	1,542 KB
5501-5628.txt	文字文件	266 KB

名稱	類型	大小
1-5628.txt	文字文件	24,500 KB

圖2.WOS 資料庫 txt 型態

3.2分析軟體

本研究使用可視化軟體 VOSviewer 與 Citespace 分析 Web of Science Core Collection (WOS) 資料庫的資料，主要是因為 WOS 提供可用於文獻計量程式可使用的格式，且涵蓋社會科學學科的3000多種期刊，其覆蓋的範圍是從2018年到目前為止，說明了種類與數量都相當齊全，再加上 VOSviewer 可以直接對 WOS 產出的資料加以直接分析，且大量筆數也能迅速完成，且其圖像化的呈現方式也相當成熟，以及 Citespace 在文獻計量分析程式中，各方面的分析功能都相當齊全，所以本研究選擇使用此兩套軟體工具來分析資料。

關於軟體 VOSviewer 和 Citespace 兩者之效能比較，已有相關文獻討論，學者張力等人在圖譜、功能和技術層面進行對比研究[4]；學者蕭明等人則分析軟體的界面、操作性、數據處理能力、繪圖功

能、圖譜特色和技術分析[8]；而林曉蔓以長期照護為研究主題，利用 Citespace 和 VOSviewer 可視化軟體，分別對作者、機構、文獻、國家等和期刊共被引用次數與時間群集進行分析[2]。

3.3 項目研究

在書目計量分析項目，同時設定描述性與評估性書目計量分析，並以大數據領域之文獻根據下列條件進行分析：

- 一、文獻來源出處國家(VOS 與 CS)：分析文獻出版國家及目前文獻數量與成長幅度。
- 二、文獻內容特性(VOS 與 CS)：分析文獻關鍵字、與研究領域的發展。
- 三、文獻期刊與相關機構 (VOS 與 CS)：分析計算文獻期刊被引用的情形、以及相關機構的分布。
- 四、參考文獻(CS)：分析被引用參考文獻，並透過分群找出目前研究的時間趨勢。

4. 可視化分析

4.1 國家分析

依照項目研究所提及大數據研究的四種條件分析，首先分析的 Citespace 軟體與 VOSviewer，結果差異性不大，因為 Citespace 可將國家數量透過可視化分布，圖形相較 VOSviewer 明顯，所以取 Citespace 的分析圖與數據為主。如圖3。

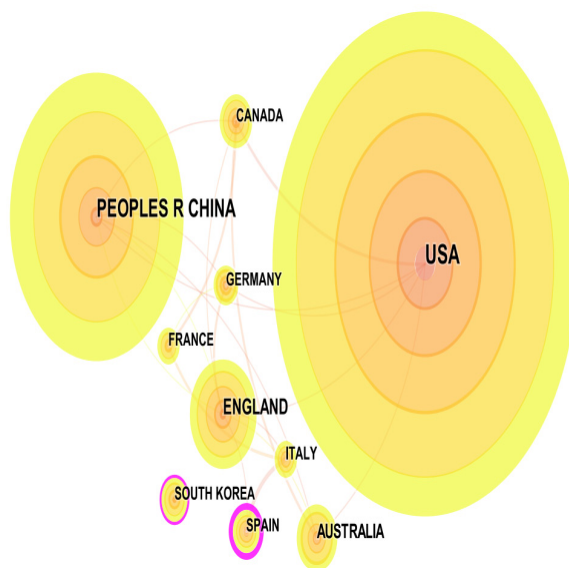


圖3. 國家分析圖(Citespace)

由此可得知文獻分布前三名的國家，首先以美國(1940)篇最多，第二是中國大陸(1117)篇，其次是英國(449)篇；文獻年份由2003到2018年之變化如表2，顯示大數據在2012年以前關注度趨近於0，明顯發展於2012年到2017年，接著2018年呈現負成長。如下圖4與表2所示。

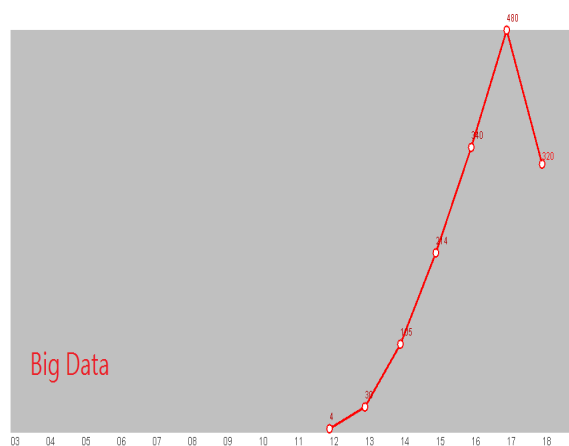


圖4. Big Data 歷年成長圖(Citespace)

表2.大數據文獻年代篇數量表

年代	文獻篇數	與前一年的變化量
2003	0	0
2004	0	0
2005	0	0
2006	0	0
2007	0	0
2008	0	0
2009	0	0
2010	0	0
2011	0	0
2012	4	4
2013	30	26
2014	105	75
2015	214	109
2016	340	126
2017	480	140
2018	320	-160

4.2 關鍵字分析

接著一樣透過VOSviewer與Citespace軟體，經由關鍵字下去做分析，由上述兩種分析軟體得知的結果大多相同，但因VOSviewer關鍵字以年份分群以及圖形的呈現，相較Citespace來說較為明顯且容易辨識，所以取VOSviewer的分析圖為主，如下圖5之VOSviewer關鍵字分析。

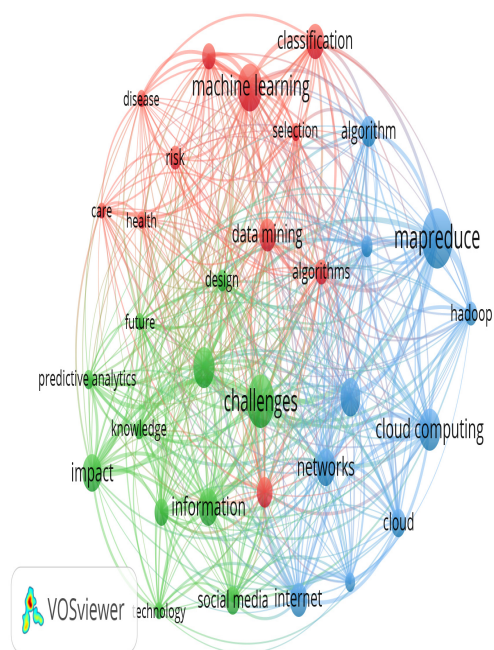


圖5. VOSviewer 關鍵字分析

由此可以推斷出大數據從以前到現今的研究領域變化，得出 mapreduce 為最多搜尋的關鍵字，其次是 challenges，再來才是 machine learning，接著則是以年代做區分可以分為三個群集，分別代表三個不同時期大數據的研究方向，第一個時期是1980~2010之間，主要關鍵字包含了 machine learning、data mining、classification，故推斷早期研究以預測分析為主要思維，透過 machine learning 來找出未來可能的趨勢，因為 machine learning 需要大量的資料，以現有的資料來建檔，並分類及歸納，整理出一套公式，透過現有可行之通用公式來預測未來走向，當然具備這樣的條件需要 data mining、classification 來輔助。

市場的顧客分析也是那時期重要的情報，所以也是大量透過顧客購買之資料彙

整並做出行為模式加以分析，以便將最需要或相關之物品做集中擺放，所以便可利用 data mining、classification 的技術和方法來獲取。

第二個時期是2011~2016之間，主要關鍵字包含了 mapreduce、cloud computing、networks，經觀察發現與大數據發展期間相符，透過 mapreduce 的衍生出 hadoop 也就是大數據的基礎，透過網路(networks)結合雲端運算(cloud computing)的分析，使得資料的儲存得到解決，並透過4G 網路讓使用者能更迅速取得資訊。

第三個時期是2017~2018之間，關鍵字包含了 challenges、information、impact，本時期是大數據的分水嶺，當中面臨了許多挑戰，不免涵蓋了虛擬機器、醫學、地質學等，這些都是我們需要用數以萬筆的資料去做分析，無法從肉眼輕易去做判斷的，所以需要大數據的演算法以及成熟的硬體設備，能夠面對各種挑戰，產生十足的影響力。

4.3 期刊分析

透過 VOSviewer 與 Citespace 軟體，經由期刊下去做分析，兩者結果差異不大，所選擇由在時間圖形表現上比較好的 Citespace 進行分析，如圖6。

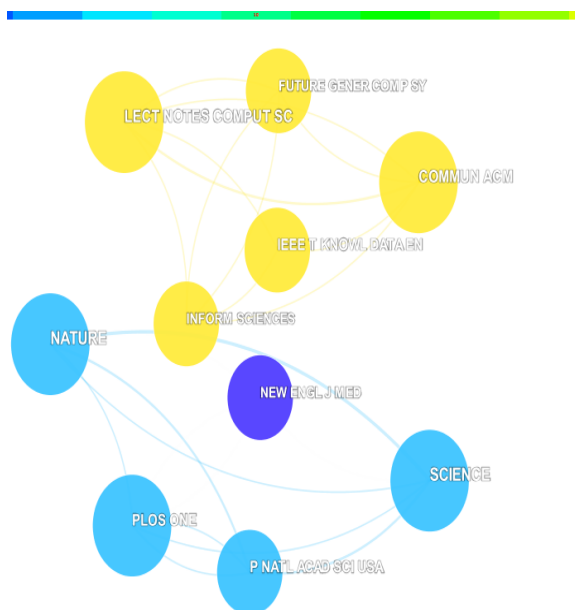


圖6. Citespace 期刊分析

由此圖得知被引用的前三期刊，首先最多的是 SCIENCE(762) 篇，其次是 NATURE (719) 篇，再來是 COMMUN ACM (632) 篇，所以大數據被引用最多的期刊以 SCIENCE 為主。

4.4 機構分析

經由 Citespace 與 VOSviewer 的比較，發現 VOSviewer 在時間分群上較 Citespace 詳細，所以由 VOSviewer 進行機構分析如圖7，由此圖得知被引用的前三期刊，首先最多的是 Chinese Acad Sci (106) 篇，其次是 MYU (54) 篇，再來是 Tsinghua Univ (53) 篇，所以大數據被引用最多的期刊以 Chinese Acad Sci 為主。

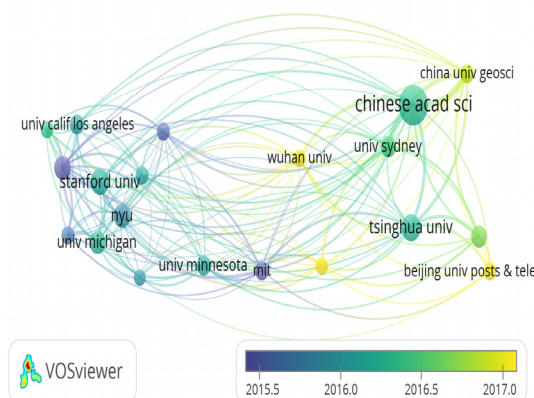


圖7. VOSviewer 機構分析

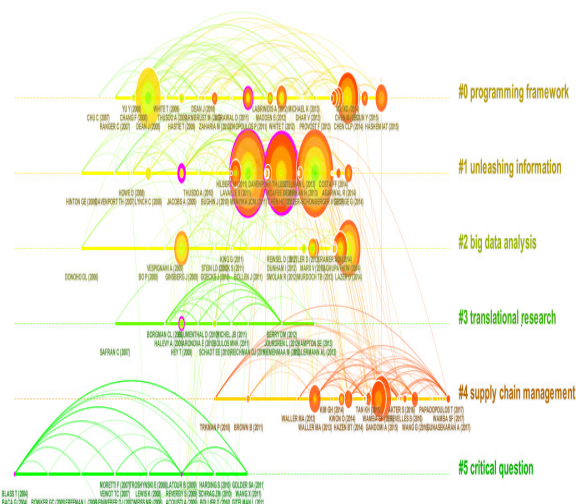


圖9. Citespace 時間群組分布圖

4.5 參考文獻分析

此分析採用 VOSviewer 與 Citespace 分析，因解析領域不同故一起取樣，分析結果如圖8。

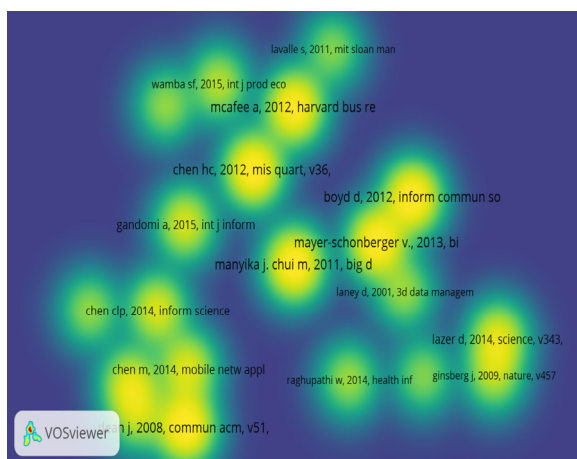


圖8. VOSviewer 參考文獻引用熱點分析圖

由此圖8得知被引用參考文獻主要集中於 2011~2014 年，只有少數是在 2008~2009 年之間。

由此圖9得知可分六群，但第二群與第三群為大數據字彙相關不採用，因此一共分為四群如下：

第一群為 programming framework 時間分布於2008-2014年，其含意為這段時間被引用參考文獻偏重於大數據的程式架構等方面研究(舉例來說:如構成大數據其中的 mapreduce 框架以及基礎 hadoop，再加上 R 軟體應用相關的研究等)。

第二群為 translational research 時間分布於2009-2013年，其含意為這段時間的被引用參考文獻偏重於轉化醫學的研究(簡單來說：此醫學研究需藉由大數據分析來整合基礎醫學和藥物開發以及臨床上等三大領域得以快速連結)。

第三群為 supply chain management 時間分布於2010-2017年，其含意為這段時間被引用參考文獻偏重於大數據在供應鏈管理的運用(例如:亞馬遜利用大數據來確保、監控、追蹤，並利用預測分析技術可以實現預期發貨的情況。對供應鏈管理的

優化大幅度的提升了其客戶的體驗)。

第四群為 critical question 時間分布於 2004-2009 年，其含意為這段時間被引用參考文獻偏重於早期大數據在各種環境下如何解決重大問題(簡單來說:早期大數據研究尚未成熟，所以透過資料探勘的輔助與機器學習來完成許多重大的問題)。

5. 結論

本研究運用 Web of Science 資料庫，收錄從 2003 年至 2018 年之間的大數據期刊論文，利用 Citespace 及 VOSviewer 的可視化圖譜技術，以及聚類演算法的分析整合相關文獻數據資料，並後製成共被引用關係圖譜及共現圖譜，藉此釐清大數據的領域中最具影響力的核心文獻及其演進過程中的關鍵轉折點，並進一步探測近年來新興研究熱點與前沿的發展趨勢，了解大數據學術研究的最新動態。

本研究已對 Özköse 等[18]與 Elshaw 等 [19]多位學者的文獻驗證，比較得到下列結果：

- 一、大數據主要研究偏向美國，研究領域多元，但歷年文獻成長的總量變化並不是一直是正成長，在 2018 年開始走下坡，相較於前者兩位學者有不同的結果。
- 二、在關鍵字分析的第二個時期，即 2011~2016 年之間，其中的主要雲端運算逐漸融入大數據，與前兩位學者

研究不謀而合。

- 三、目前大數據的研究趨勢由此研究分析軟體得知，在不同時期有不同研究領域。

- 四、研究領域可分為以下四點：

1. 在 2004-2009 年之間，也就是早期的大數據在各種環境下如何解決重大問題(簡單來說：早期大數據研究尚未成熟，所以透過資料探勘的輔助與機器學習來完成許多重大的問題)。
2. 在 2008-2014 年之間，其含意為這段時間研究偏重於大數據的程式架構等方面研究(舉例來說：如構成大數據其中 mapreduce 框架以及基礎的 hadoop，再加上 R 軟體應用相關的研究等)。
3. 在 2009-2013 年之間，其含意為這段時間研究偏向於轉化醫學的研究(簡單來說：此醫學研究需藉由大數據分析來整合基礎醫學和藥物開發以及臨床上等三大領域得以快速連結)。
4. 在 2010-2017 年之間，主要是說明大數據在供應鏈管理的運用(例如：亞馬遜利用大數據來確保、監控、追蹤，並利用預測分析技術可以實現預期發貨的情況。對供應鏈管理的優化大幅度的提升其客戶的體驗)。

5.1 研究限制

整體而言，本研究限制有四：

- 1.因 WOS 資料庫僅提供2003年之後的文獻統計資料，故本研究受限於資料的年限限制，僅能以2003~2018年之間所得數據做為研究分析的資料來源。
2. Citespace 在建構可視化圖形整個流程中，各個步驟處理比較瑣碎，且時間區域內的群集標籤堆疊現象不易辨識。反觀 VOSviewer 操作性佳，可視化圖形美觀，直接辨識度優，可惜分析功能選擇太少。
- 3.Citespace 與 VOSviewer 在字母縮寫部分存在誤判，在 Citespace 方面：同個名字縮寫方式不一樣會無法合併，使結果產生誤差，而 VOSviewer 採取一致的格式(以縮寫為主)，在全名不同但縮寫一樣情況下，會誤判成同筆資料。
- 4.Citespace 在編群號碼是由0開始編排，而 VOSviewer 的編群號碼則是由1開始，由於演算法的不同，兩套軟體在文獻共被引用分群時，結果有所落差。

參考文獻

- [1]. 林敏，臺灣地區資訊組織文獻書目計量分析，圖書資訊學刊，7(1/2)，pp. 101-123(2009)。
- [2]. 林曉蔓，長期照護研究文獻計量與可視化分析，義守大學資訊工程學系博士論文(2017)。
- [3]. 何國光，「圖書資訊組織原理」，三民書局(1993)。
- [4]. 張力、趙星、趙星、趙星葉鷹，信息可視化軟件 Citespace 與 Vosviewer 的應用比較，信息資源管理學報，1，pp. 95-98 (2011)。
- [5]. 連柏鈞，RapidMiner 大數據挖掘應用與案例分析，朝陽科技大學碩士論文(2017)。
- [6]. 曾至浩，大數據發展對保險監理影響之研究，國立中正大學法律學系研究所碩士論文 (2018)。
- [7]. 蔡松翊，大數據應用於商辦大樓之中央空調系統性能預測，國立交通大學碩士論文(2016)。
- [8]. 蕭明、邱小花、黃界、李國俊，馮召輝，知識圖譜工具比較研究，圖書館雜誌，pp. 61 -69 (2013)。
- [9]. Burnap, P., and Rana, O., and Williams, M., and Housley, W., and Edwards, A., and Morgan, J. "COSMOS: Towards an integrated and scalable service for analysing social media on demand", International Journal of Parallel,

- Emergent And Distributed Systems, 30 (2) , pp. 80-100(2014).
- [10].Chiang, H. E. and Hughen, W. K. “Do oil futures prices predict stock returns? ”, *Journal of Banking & Finance*, vol. 79, pp. 129-141(2017).
- [11].Elshaw, R., Sakr,S., Talia,D., Trunfio,P. “Big Data Systems Meet Machine Learning Challenges:Towards Big Data Science as a Service” . *Big Data Research* 14, pp. 1–11(2018).
- [12].Elshawi , R., Sakr S., Talia, D., Trunfio P. “Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service”, *Big Data Research*, 14, pp. 1-11(2018).
- [13].Gantz, J., & Reinsel, D. “Extracting value from chaos.” IDC iview, 1142, pp. 9-10(2011).
- [14].Gartner IT Glossary, What is Big Data?(2014), URL: <http://www.gartner.com/it-glossary/big-data>.
- [15].Gürsakar, N., Büyük Veri, Dora Kitabevi, Bursa(2014).
- [16].Hammond, K., and Varde., “A. Cloud based predictive analytics: text classification, recommender systems and decision support”, *Proceedings of IEEE 13Th international conference on data mining workshops* (2013).
- [17].Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. “The rise of big data on cloud computing: Review and open research issues”. *Information Systems*, 47, pp. 98-115(2015).
- [18].Khan, S., and Liu, X., and Shakil, K. A., and Alam, S. M. “A survey on scholarly data: From big data perspective”, *Information Processing & Management* Volume 53, Issue 4 ,pp. 923-944(2017).
- [19].Khan, Z., and Anjum, A., and Soomro, K., and Tahir, M. “Towards cloud based big data analytics for smart future cities”, *Journal of Cloud Computing*, 4 (1) (2015).
- [20].Lu, S., and Li, R., and Tjhi, W., and Wang, L., and Li X., and D, Ma. “A framework for cloud-based large-scale data analytics and visualization: case study on multiscale climate data”, *Proceedings of the IEEE third international conference on cloud computing technology and science* (2011).
- [21].Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. “Big data: The next frontier for innovation, competition, and productivity” (2011).
- [22].Özköse, H., and Ari,E.S., and Gencer, C. “Yesterday, Today and Tomorrow of Big Data”, *Procedia - Social and*

Behavioral Sciences, 195, pp. 1042-1050 (2015).

- [23].Raghupathi, W., and Raghupathi, V.
“Big data analytics in healthcare: Promise and potential”, Health Information Science and Systems, 2 (1), pp. 3(2014).
- [24].TechAmerica Foundation’s Federal Big Data Commission, Demystifying Bigdata: A Practical Guide To Transforming The Business Of Government(2014), URL:
https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf.

Bibliometrics and Visualization Analysis of Big Data

Chao Y. Wu¹ Jia-Ling Liang² Chein P. Li³

¹ Department of Information Management, I-Shou University, Taiwan

² Department of Computer & Information Science, ROC Military Academy

³ Department of Information Engineering, I-Shou University, Taiwan

ABSTRACT

In an era of information explosion, almost everything can be explained using information and figures. Meanwhile, the booming development of current technologies and techniques makes information collection easier and also extends its scope. By widely collecting information to create big data, humans are able to learn information hitherto unknown. With bibliometrics and visualization analysis, valuable information can further be found to subsequently help decision makers make a judgment quickly.

This study used “big data” as a keyword to do keyword research, and set “journal papers from 2003 to 2018” as a matching criterion for data collection, with a focus on the 5628 papers found in the WOS database. An analysis was conducted on, and a comparison made between, these 5628 papers collected from the WOS database using the visualization software: citespace and VOSviewer.

The conclusion suggested that there has been a decline in the number of studies on big data in recent years. It can be learned from visualization images that big data focused on prediction and analysis of studies in the early stage, where a large amount of information was collected via machine learning and supplemented by data mining and classification, to find potential future trends. Application software, such as hadoop and R, were combined with cloud computing to conduct an analysis in the middle stage. Algorithms for big data and mature hardware and equipment were needed in the late stage to create an adequate impact when facing challenges.

Keywords: big data, Citespace, VOSviewer, visualization analysis