以兩階段資料處理法改善大數據資料不可用的問題 ---大數據的藍海

王淑卿	王順生*	嚴國慶*	邱怡瑄	黃學馴
朝陽科技大學	朝陽科技大學	朝陽科技大學	朝陽科技大學	朝陽科技大學
資訊管理系	工業工程與管理系	企業管理系	資訊管理系	企業管理系

{scwang; sswang; kqyan; s10114028; s9937902}@cyut.edu.tw

*: 聯絡人

摘要

近年來隨著全球資料量呈現爆炸性的增長，大數據(Big Data)已成為當前最熱門話題之一，此一技術的推出可以協助提升決策的效率、因應市場的需求縮短產品上市的時間、提升利潤等。而大數據的上游資料是由網路上的設備物件(物聯網)而來，而這些資料存放在中游的雲端儲存系統，再交給下游的大數據分析將大量的資料即時轉為有用的資訊，最後這些資訊可以立刻被各種設備所存取利用。由於過去對於大數據議題之探討，大都聚焦在資料的分析與處理層面。然而在大數據應用需求中，經常會遇到“6用問題”。其中，“6用問題”的資料不可用問題中之不活躍使用者的資料，需要透過資料儲存和管理技術加以處理。因此，本研究在大數據相關應用中提出以「兩階段資料處理法(Two Phase Data Processing method；TPDP)」，針對在大數據中資料不可用問題之不活躍使用者資料的儲存和管理進行研究。TPDP的第一階段是以本研究所提出的MRFM(Modified-RFM)找出不活躍使用者的資料；接著，在第二階段再以改良後的布隆過濾器(Linked based Bloom Filter；LBF)進行不活躍資料的儲存，以提供後續不活躍使用者資料的維護。

關鍵詞：大數據、6用問題、布隆過濾器。

Abstract

Timely and cost-effective analytics over “Big Data” has emerged as a key ingredient for success in many businesses, scientific and engineering disciplines, and government endeavors. The upstream data of Big Data is generated by the device from objects, and these data stored in the cloud storage system. Then the Big Data will be turned into useful information that can be immediately accessed by the use of a variety of devices. In the past, the study of Big Data mostly focused on the level of data analysis and processing. However, "the six usage issues" with the applications of Big Data are often encountered. And, the information of inactive users in the unavailable data issues need to be addressed through the data storage and management techniques. In this study, "Two Phase Data Processing method (TPDP)" is proposed to solve the management of inactive user's information in the unavailable data issues. In the first phase of TPDP, MRFM (Modified-RFM) is used to identify the information of inactive users. In the second phase of TPDP, Linked based Bloom Filter (LBF) is used to maintain the information of inactive user.

Keywords: Big data, Six usage issues, Bloom Filter.

1. 前言

近年來隨著全球資料量呈現爆炸性的增長，大數據(Big Data)已成為當前最熱門話題之一，此一技術的推出可以協助提升決策的效率、因應市場的需求縮短產品上市的時間、提升利潤等[15]。大數據的應用範圍非常廣，包括科學、天文學、大氣學、網際網路檔案處理、大規模的電子商務等。而在許多領域，由於資料集過度龐大，科學家經常在分析處理上遭遇限制和阻礙。因此過去對於此議題之探討，大都聚焦在資料的分析與處理層面[18]。

依據朱揚勇與熊贇的研究，一個大數據應用需求，通常會遇到“6用問題”[2]。“6用問題”包括：(1)資料不夠用、(2)資料不可用、(3)資

料不好用、(4)資料不會用、(5)數據不敢用、及(6)資料不能用。其中，所謂的資料不可用是指在資料夠用的情況下，還會遇到資料不可用問題。資料不可用是指擁有資料，但訪問不到資料。例如，某個公共決策需要用到教育部、經濟部、警政署、財政部的資料，這些資料在各部門都有，但是資料不在一個系統裡，是資料孤島，並不能用來做大資料決策。又如，一些交易系統只保留活躍使用者的資料，不活躍使用者的資料被備份到備份系統中，然而查詢備份系統資料是一件費時、費力的工作，甚至是不可能的工作。

根據藍海策略一書中之討論[10]，紅海策略是指在現有已存在的市場中進行競爭，企業爭取的是以擴大規模來降低成本，透過降低成本的方式以期在價格中取得競爭優勢。而藍海策略則是在現在的市場中找尋尚未開拓的市場，或是進入較少業者進入的領域，企業強調的是以創新來創造新的市場需求，以差異化來獲取高額的利益。由於，資料不可用問題中的不活躍使用者的資料，就如同大數據相關研究領域的藍海，是需要被研究、使用資料儲存和管理技術，使其具有如藍海中的不完全競爭市場。而在這個市場裡，由於競爭者少或尚未有競爭者，因此企業在這個市場中，只需不斷的創新，就能以差異化的服務來吸引消費者，更能獲得良好的競爭優勢。

因此，在本研究中首先將找出不活躍使用者的資料，接著再以適當的結構儲存不活躍使用者的資料，以利後續這些資料的儲存或使用。在本研究中，將改良RFM(Recently、Frequency、Monetary)資料分析技術以找出不活躍使用者的資料。除此之外，研究中更將使用改良後的布隆過濾器(Bloom Filter)，以較佳的時間複雜度進行不活躍使用者資料的索引查詢並以較少的空間進行設計，來達到動態維護不活躍使用者資料的作業。

本研究的內容共分為 5 節。第 1 節為前言，說明研究背景動機與目的。第 2 節為文獻探討，分別說明大數據、RFM(Recently、Frequency、Monetary)資料分析技術與布隆過濾器(Bloom Filter)。第 3 節說明本研究所提出的改良式的 RFM 及改良後的布隆過濾器。第 4 節說明本研究的執行步驟。最後一節為結論與未來研究。

2. 文獻探討

在本節中將說明大數據、RFM(Recently、Frequency、Monetary)資料分析技術與布隆過濾器(Bloom Filter)。

2.1 大數據

依據資策會於 2013 所做的報導，急遽成長的大數據將帶來結構性變革，預估至 2015 年將會有 80%的可用資料呈現不確定性，如圖 1 所示[23]。大數據具有大量(Volume)、快速(Velocity)、及多樣性(Variety)三項特性[13,22]，其中 Volume 指的是處理的資料量；Velocity 是資料產生與處理的速度；Variety 是資料的多樣性，如圖 2 所示。

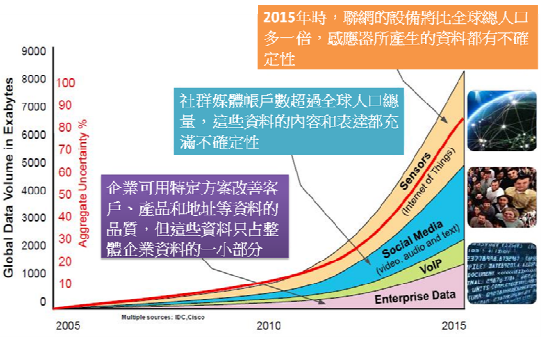

圖 1、大數據帶來的變革[23]

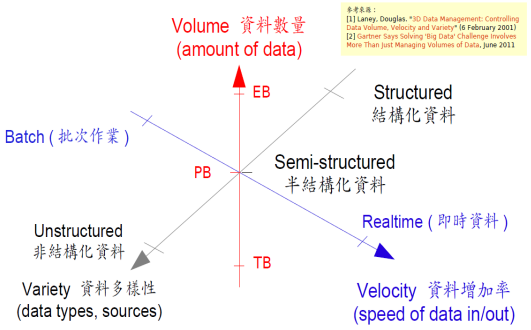

圖 2、大數據的 3V[23]

大數據的挑戰在於如何管理「數量」、「增加率」、與「多樣性」。處理大數據則包括三類主要技術，分別為(1) Data at Rest；(2) Data in Motion In-Memory Processing；及(3) Streaming Data Collection。其中，在Data at Rest中針對靜態資料的處理，可以利用Hadoop開發一個框架。Hadoop包括HDFS和MapReduce兩部分，其中HDFS可用於儲存非結構和結構性的資料，並執行大數據的處理與應用。在Data in

Motion中針對動態資料的處理，需先進行資料的蒐集(Data Collection)，再在分散式資料庫中資料處理(Data Processing)，最後再進行事件處理(Event Handling)。至於Streaming Data的蒐集，則可以使用Message Queue進行[20,24]。

由於現今終端設備的處理能力與儲存容量不斷提升，以及設備有越來越小化的趨勢，加上資訊技術與網路型態不斷的改良與創新，人們的生活正逐漸進入一個始終連接(Always Connect)的網路世代，而網路型態已從過去以人為導向(Internet of People)衍生至以物(Thing)為導向的物聯網(Internet of Things；IoT)環境的網路型態[17]。在物聯網的環境中，整合許多資訊科技，包括無線感測網路、RFID技術、行動通訊技術、即時監控、普及運算及IPv6等，透過這些技術的使用，將可獲得每一個智慧物件(Smart Object)的相關資訊，並應用這些資訊進行相關服務。然而在物聯網複雜的環境中，所提供的資訊量非常龐大，因此運算資源與模式已從過去的網際網路衍生至現在的雲端運算(Cloud Computing)[8]。而隨著時間的發展，未來人們將進入物聯網的生活環境中。物聯網不僅是網際網路或通訊網路的擴展，更由於是透過物件間連結的能力，可搜集到更豐富的資料，因此可提供更為智慧的服務環境。

大數據近年來吸引許多人的注意，其根本原因在於科技的進步，尤其是網際網路應用服務的興起。這些新興的服務大幅豐富了資料的內容，也大幅增加資料分析的困難度，使得傳統的資料處理工具與方法不再完全適用。而大數據的產生及後續的研究與物聯網及雲端儲存間息息相關，在實務上是屬於同一線上的上中下游。上游資料是由網路上的設備物件(物聯網)而來，而這些資料存放在中游的雲端儲存系統，再交給下游的大數據分析將大量的資料即時轉為有用的資訊，最後這些資訊可以立刻被各種設備所存取利用。物聯網、雲端儲存、與大數據三者的關係如圖3所示[23]。

由於大數據所具有的大量、快速、及多樣性的特性，因此傳統的關聯式資料庫系統及在此系統上運作的資訊探勘、統計分析等各種技術，無法直接因應上述幾項特性。是以面對大數據需要設計能夠大量儲存並實行平行運算的各種新方法來處理、利用大數據，以期找出大數據中蘊含的巨大價值(Value)。

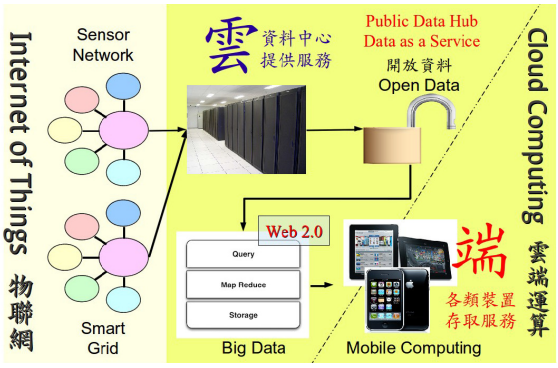

圖 3、物聯網、雲端儲存、與大數據三者的關係[23]

2.2 RFM 資料分析技術

在過去相關的顧客關係管理之研究領域中，已知有許多方法可以幫助企業瞭解其顧客的相關資訊[6]，運用這些相關資訊進而了解顧客的消費行為並制訂滿足其顧客需求的行銷策略。其中，Kahan 主張 RFM(Recently、Frequency、Monetary)資料分析技術由於能提供企業每個顧客的交易資訊，因此是比一般認知分析(Cognitive Analysis)更為有用的行為分析(Behavioral Analysis)技術[14,16]。並且在許多藉以瞭解顧客相關資訊的方法中，以 RFM 資料分析技術最為被廣泛運用[16]。

RFM 資料分析技術是由 Arthur Hughes 在 1994 年所定義的，所謂的 RFM 資料分析技術分別為最近購買時間(Recently；R)、購買頻率(Frequency；F)與購買金額(Monetary；M)的縮寫。其中，R 是指顧客最近的購買時間，即所謂顧客最近一次購買的時間與現在時間的距離天數，用來衡量顧客再次購買的可能性。當購買時間距離愈近，則表示該顧客再次購買程度愈高；若最近購買日期距離愈遠，則表示該顧客購買意願降低、或購買行為改變、或是因其他因素而導致至他處消費。F 是指在某段期間內購買該企業產品的總次數，此期間可定義為一個月、一季、或是任何可衡量的時間長度，可用來衡量顧客在購買行為中與企業的互動程度。當購買頻率愈高，則表示顧客的熱衷程度愈高。M 是指在某段期間內購買該企業產品的總金額，可做為用來評價顧客對該企業的貢獻度及顧客價值。當顧客的購買金額愈高時，則表示顧客的價值較高。RFM 資料分析技術主要是在分析及衡量顧客的消費行為，透過顧客過去的歷史交易資訊進行顧客的區隔，以作為衡量顧客的忠誠度與貢獻度之依據。

Arthur Hughes 認為 RFM 資料分析技術在衡量一個顧客的重要性程度是一致的[3,4]，透過顧客最近一次消費可以呈現顧客最近才來購買商品或者服務，很有可能近期再來購買。因為吸引最近才來購買的顧客，比吸引很久沒有來購買的顧客來的容易。消費頻率是顧客在同一期間內購買的次數，也可以說最常購買的顧客，也就是對商品或者服務是最滿意的，因此相對的消費頻率高也就是忠誠度高。而在消費金額則是透過顧客在同一期間內消費的總金額，透過這三個要素分析消費者的價值。

傳統使用於顧客關係管理的 RFM 資料分析技術，可以以顧客過去的歷史交易資訊進行顧客的區隔，透過衡量顧客的忠誠度與貢獻度後，提供深度經營之參考依據。而由於大數據具有的大量、快速、及多樣性的特性，為提供大數據相關的應用與服務，一個快速的服務回應機制，是必須探討的重要議題之一。而由於在大數據相關的應用與服務中，檔案被存取的行為資訊與顧客的消費行為之交易資訊有類似的特性，因此本研究將改良傳統之 RFM 因子，並運用於資料儲存時資料的價值分析，以找出不活躍使用者的資料。

2.3 布隆過濾器

在布隆過濾器的相關研究中，針對不同的應用也有許多學者提出改良的布隆過濾器。Antichi 等學者透過改良布隆過濾器的機制，能夠將其結合索引值，雖然過濾器判斷存在後能夠快速的找到索引位置，其時間複雜度為 O(log(n))[7]。但 Antichi 等學者的機制必須在完美雜湊函數(Perfect Hash)下才能夠實現，在實際應用上，完美雜湊函數的設計式非常不容易，且其研究也沒有考慮到刪除檔案時，布隆過濾器該如何處理。

Papadopoulos 等學者，利用 R-tree 的資料結構與布隆過濾器的資料結構進行範圍查詢(Range Query)與單點查詢(Point Query)的應用[19]。在其實驗中可發現，布隆過濾器在實現單點查詢有非常優異的表現。

Wei 等學者，提出了動態布隆過濾器架構，當檔案系統容量增大時，不必修改現有映射位置，而是新增新的陣列來進行映射放置，藉此能夠有效降低錯誤率，以滿足其動態性。而在進行刪除動作時則採用批次處理，利用兩個計數器記錄，一個計數器紀錄目前該位置共有幾個映射值指向這個位置，另一個計數器紀錄刪除幾個檔案，當第二個記數器到達一定數量時，則會更新目前狀態[21]。雖然該方法能夠解決布隆過濾器刪除的問題，但透過兩個計數器記錄，卻會浪費許多空間。

雖然目前已有許多學者提出布隆過濾器與索引值結合的方法[7,19]，但還是有其限制，如必須在完美雜湊函數下進行，且目前所提出的方法之時間複雜度仍是 O(log(n))。在一般大數據的環境中，勢必會經常進行檔案的維護，由於過去學者所提出的方法也不盡理想。因此在本研究中提出了鏈結式布隆過濾器(Linked based Bloom Filter；LBF)，透過鏈結結構與雜湊函數的整合，以較佳的時間複雜度進行索引查詢並以較少的空間進行設計，來達到動態維護不活躍使用者資料的作業。

3.研究方法

在本節中將說明本研究所提出的改良式 RFM 及改良後的鏈結式布隆過濾器(Linked based Bloom Filter；LBF)。

3.1 改良式的 RFM (Modified-RFM；MRFM)

Arthur Hughes 學者認為 RFM 資料分析技術在衡量一個顧客的重要性程度是一致的[3,4]，因此採用相對的分級，將 R、F、M 各分為五等分，以 1~5 來做為區分，數字越大則代表價值越高，即每一等分將是整個資料庫的 20%。因此，每位顧客在紀錄中將會有三個數字，組成方式為 555、554、553 到 111，共有 125 種，而 555 即代表最近曾經購買產品、且購買次數頻繁、以及消費的金額很多的顧客。

此外，平均購買時間間隔是學者沿用以久，用來計算購買期間最簡單的方法。Goodman 學者透過計算每次顧客購買的間隔時間[12]，利用 MLE (Maximum Likelihood Estimation)計算出該顧客的平均購買時間間隔，MLE 計算公式如(1)所示，t_1 係指第一次購買時間，t_2 為第二次購買時間，以此類推。

$$\text{MLE}=(t_1+t_2+t_3+\ldots+t_n)/n, \text{ 即 } \sum_{i=1}^{n} t_i / n \ldots(1)$$

最後將依據最大概似估計值，對顧客購買的間隔時間來做加權，計算出加權平均購買期，簡稱 WMLE(Weighted Maximum Likelihood Estimation)，WMLE 計算公式如(2)所示。

$$\text{WMLE}=(t_1+2t_2+3t_3+\ldots+nt_n)/(1+2+3+\ldots+n),$$
$$\text{即} \sum_{i=1}^{n} it_i / \sum_{i=1}^{n} i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

根據以上敘述，透過 MLE 與 WMLE 之差異，可用來作為判斷該顧客的價值及分類，如表 1 所示[12]。

除了根據上述的分析來做為判定顧客價值之外，Goodman 學者也提出透過(MLE-WMLE)/MLE 以獲得顧客變動比率，再配合運用平均購買次數或平均購買金額與判定值之差異，來做為顧客價值之判定，如表 2 所示[12]。

表 1、顧客價值趨勢分析表[12]

MLE 與 WMLE 之差異	**顧客價值趨勢分析**
MLE-WMLE>0	表示顧客的購買間隔時間短，為企業主要的獲利顧客，即具有較高的顧客價值。
MLE-WMLE=0	表示顧客的購買間隔時間穩定，為企業一般獲利客顧客，即顧客價值趨向穩定。
MLE-WMLE<0	表示顧客的購買間隔時間長，為企業公司較少的獲利顧客，即具有較低的顧客價值。

表 2、顧客價值趨勢及平均判定值分析表[12]

	(MLE-WMLE)/MLE>0	**(MLE-WMLE)/MLE<0**
平均購買次數-判定值>0	忠誠顧客	消極的顧客
平均購買次數-判定值<0	機會顧客	(即將)流失的顧客
平均購買金額-判定值>0	忠誠顧客	消極的顧客
平均購買金額-判定值<0	機會顧客	(即將)流失的顧客

而在本研究所提出的改良式的 RFM(Modified-RFM；MRFM)，針對使用者的資料以活躍度 RFM 分析因子進行活躍度的分析。本研究為因應大數據的特性，因此改良過去熱門度 RFM 分析因子，使之成為適用於大數據環境下的使用者資料活躍度分析因子，其中 R 為 Recently 是資料最近一次被存取的時間；F 為 Frequency 是資料被存取的頻率；M 為 Monetary 是資料大小。

透過活躍 MRFM 分析因子對使用者的資料進行分析時，當使用者的資料透過 MRFM 分析後，將可以分析出每一個資料的活躍度。而針對每一個資料的 R、F、與 M 都會有一個介於 0 至 1 的活躍度，且此活躍度將會影響資料的更新速度。在 MRFM 的分析中，也採用相對的分級，將 R、F、M 各分為五等分，以 1~5 來做為區分，其中數字越小則代表使用者資料越不活躍。對應 Goodman 學者所提的顧客價值趨勢及平均判定值分析[12]，即為「流失的顧客」，亦是本研究探討的「藍海」資料。

本研究將以 MRFM 資料分析技術進行巨集資料的活躍度區隔，藉以找出不活躍的資料。接著，再以改良後的布隆過濾器進行不活躍資料的儲存，以提供後續不活躍使用者資料的維護。

3.2 鏈結式布隆過濾器(Linked based Bloom Filter；LBF)

布隆過濾器(Bloom Filter)是由 Burton Bloom 學者[9]在 1970 年時所提出的，它最大的特色就是能夠快速的辨識某個資料是否存在於資料集中。一般布隆過濾器會有一個 m 大小的陣列、及 k 個雜錯函數(Hash Function)，其陣列的初始皆為 0，當有檔案進行存入時，會透過 k 個雜錯函數進行運算，雜錯運算結果為指向陣列之索引位置，被指到的陣列位置其值將轉換成 1，若重複映射至同一位置，則數值保持為 1 即可。進行資料判斷是否在該資料集中，亦是透過 k 個雜錯函數進行計算，若 k 個雜錯映射到的位置皆為 1，即表示該資料存在於資料集中，若是有一個位置為 0，即表示

該資料不存在於資料集中，其概念圖如圖 4 所示，其中在進行查詢時檔案 C 映射位置有一值為 0，透過快速過濾後可判定檔案 C 不存在於資料集中。

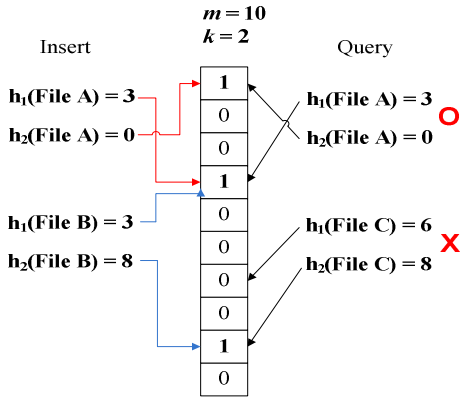

圖 4、布隆過濾器概念圖

布隆過濾器在極大的效率中有著“存在誤判”的可能性(Positive False)，即為過濾器雖然判定該資料存在於資料集中，但事實上卻不存在。但布隆過濾器可以保證不會發生“不存在誤判”的情形(Negative False)，即當布隆過濾器判定該資料不存在於資料集中，就表示該資料絕對不存在於資料集中。“存在誤判”的機率如公式(3)所示[9,11]，其中 m 為陣列大小、k 為雜錯函數數量、n 為存在資料集中的總檔案數量。

$$f' = (1-(1-1/m)^{kn})^k \quad \text{.....(3)}$$

“存在誤判”的機率大小受到以上三個因子影響，表 3 為分析各因子與錯誤率之關係表。

表 3、各因子影響錯誤率

名稱	變數	減少	增加
陣列大小	m	增加 錯誤率	減少 錯誤率
雜錯函數 數量	k	增加 錯誤率	減少 錯誤率
元素數	n	降低錯 誤率	增加 錯誤率

當 m 與 k 較大時，其錯誤率會較小。但在實際應用上，建置布隆過濾器時會採用固定的陣列大小(m)。當採用固定大小的陣列後，k 值會有一門檻值，當 k 超過門檻值時，其錯誤率反而會上升。因此，在設定 k 個雜錯函數時，若等於門檻值，將可得到最小的錯誤率，k 值門檻值的公式如(4)所示[9]。

$$k = (m/n)\ln 2 \approx m/n * 0.69314 \quad \text{.....(4)}$$

布隆過濾能夠透過極小的錯誤率換取極大的效率，因此在能夠接受些許“存在誤判”的應用下，布隆過濾器能夠完全發揮其優勢。

然而，傳統布隆過濾器仍有許多缺點，如：

(a) 當資料刪除時，陣列相關位置上的 1 無法變回 0，因為這個位置可能同時有好幾筆資料指向這個位置；

(b) 當檔案系統容量增加時，此時 n 的數量也會增加，此時若 m 的大小沒有增加，其錯誤率會上升，但若改變 m 的大小，勢必要將全部的檔案重新進行雜錯映射，其成本非常龐大；

(c) 傳統布隆過濾器只能做到過濾的效果，也就是說它只能夠初步過濾出此檔案是否存在於本系統中，若判斷為是，將無法繼續判定該檔案儲存的位置，缺少了索引性的特性，在實用上還是有些不足。

本研究所提出鏈結式布隆過濾器(Linked based Bloom Filter；LBF)，能夠直接結合索引值進行快速查詢。其方式是採用鏈結串列進行存放資料，當進行新增資料時，將資料之唯一識別碼透過 k 個雜湊函數運算並映射至布隆過濾器上，接著使用 k 個運算完的雜湊值進行運算，決定鏈結串列之起始位置。k 個雜湊值轉換成鏈結串列之起始位置如公式(5)所示，其中 $val(h_i)$為第 i 個雜湊值、m 為布隆過濾器之長度、%符號表示餘數函式(mod)。

$$\text{Link}(position) = (\sum_{i=1}^{k} val(h_i)) \% m \ldots (5)$$

透過上述公式計算後，將得到布隆過濾器上的某個位置，接著該位置將產生鏈結串列指標，資料之索引值將會儲存在該鏈結串列上。舉例來說，假設有 4 個雜湊函數 $val(h_1)$=6069、$val(h_2)$=36、$val(h_3)$=5、$val(h_4)$=642、m=6463，透過(4)計算得出鏈結串列起始位置為 289 [(6069+36+5+645) %6463]，因此會在布隆過濾器之陣列索引值 289 產生鏈結串列，存放該資料之索引值。

若不同資料經 LBF 計算完畢後之位置相同(如資料 X 計算完畢後亦為 289)，則會由最

後一個鏈結串列產生新的鏈結，根據需求動態產生鏈結空間進行存取，可有效將空間使用率提升，LBF 的整體概念圖如圖 5 所示。

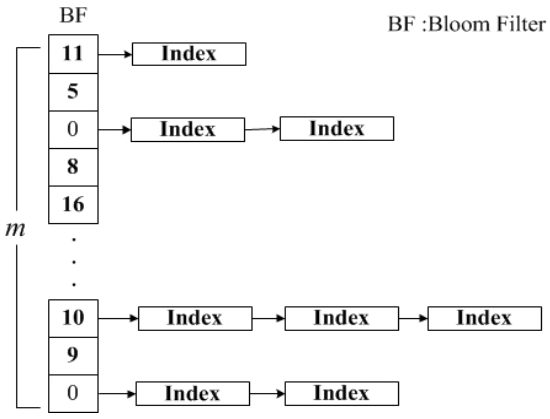

圖 5、鏈結式布隆過濾器的概念

資料比對動作和存入之動作相同，當有比對需求進入後，首先將資料之唯一識別碼透過布隆過濾器判斷檔案是否存在，若布隆過濾器判斷不存在，則可直接上傳資料進行儲存；若布隆過濾器判斷存在，則透過公式(5)進行計算，並至相關位置尋找鏈結串列，透過本研究之方法能夠快速進行比對動作。若是搜尋完所有鏈結串列仍未發現相同資料，表示發生“存在誤判”情形，而 LBF 透過快速的判斷，亦能將“存在誤判”所付出的額外成本降低。

4.研究步驟

在本節中，將說明本研究所提出的在大數據相關應用中以「兩階段資料處理法(Two Phase Data Processing method；TPDP)」，針對在大數據中資料不可用問題中不活躍使用者資料的儲存和管理進行處理。TPDP 的第一階段是以本研究所提出的 MRFM 找出不活躍使用者的資料；接著，在第二階段再以改良後的布隆過濾器(LBF)進行不活躍資料的儲存，以提供後續不活躍使用者資料的維護。

4.1 以 MRFM 找出不活躍使用者的資料

在本研究中，以王順生等學者及陳聖中學者在研究中所使用的雲端資料進行說明**[1,5]**。在實作部分，本研究所進行的 MRFM 資料分析技術是使用 SPSS 統計軟體進行資料分析。另外，在 MRFM 資料活躍度分析中的 MLE 和 WMLE 分析，則是運用 Microsoft Excel 來進行相關的資料分析與計算。

在研究中，TPDP 的第一階段是以 MRFM 找出不活躍使用者的資料。本研究透過 MRFM 資料分析技術進行使用者資料的活躍度分析，分別將資料最近一次被存取的時間(Recently)、資料被存取的頻率 (Frequency)與資料存取所需的時間長度(aMounts)等三項進行五等分後，即各等分均為 20%，分別給定所屬於的 R、F、及 M 值，計算出個別的 RFM 總分。如圖 6 所示，可以看出每個資料的 RFM 得分分佈狀態，用以找出其資料的活躍度分析，其分數越高代表資料的活躍度越高。

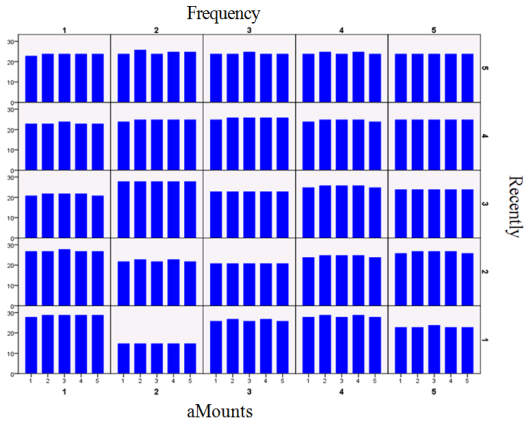

圖 6、每個資料的 MRFM 得分分佈狀態

其次，本研究運用Goodman學者所提出MLE(Maximum Likelihood Estimation) 及WMLE(Weighted Maximum Likelihood Estimation)進行資料活躍度趨勢分析，分析結果如圖7所示。其中資料活躍度趨勢大於0為89.45%，即具有較高活躍度的資料；資料活躍度趨勢等於0為0.08%，表示資料活躍度趨勢趨向穩定；資料活躍度小於0為10.44%，為較低活躍度的資料。在本研究中，資料活躍度小於0的10.44%資料將被篩選出來，並代入，TPDP的第二階段以LBF進行不活躍資料的儲存與維護。

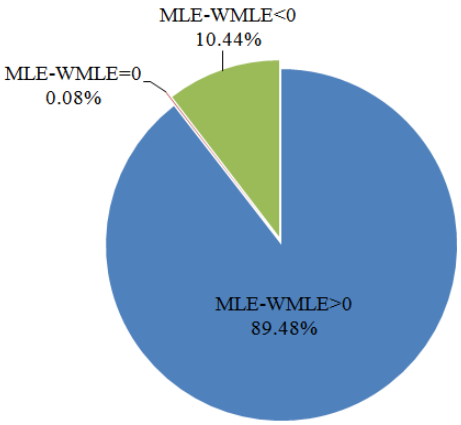

圖 7、資料活躍度趨勢分析分佈圖

4.2 以 LBF 進行不活躍資料的儲存

在研究中，TPDP 的第二階段是以 LBF 進行不活躍資料的儲存。由於傳統布隆過濾器最主要的功能就是快速的判斷出資料是否已儲存在資料集中[7,9]，然而若是將其使用在分散式檔案系統中的查詢，仍有其不足的地方，例如(1)若布隆過濾器判斷該資料是存在的，卻無法繼續判定該資料儲存在那裡，缺少了索引性的特性、(2)在資料系統中，使用者會經常進行增加與刪除資料的動作，然而布隆過濾器卻無法任意進行刪除，使其使用上的彈性不足。因此，在 TPDP 第二階段進行資料比對作業時，採用 LBF 進行不活躍資料的比對與儲存。LBF 初始化之流程如圖 8 所示。

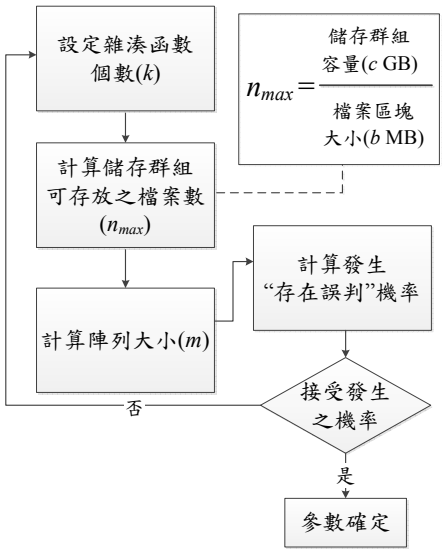

圖 8、IBF 初始化

步驟與實例說明如下：

1. 設定雜湊函數個數(k)，假設共有 7 個雜湊函數。
2. 計算該資料集內共可以存放多少個資料數量(n_{max})，假設該儲存群組織容量為 40 GB，資料檔案區塊大小為 64 MB，因此該資料集可存放 640 個資料檔案(40 GB /64 MB)。
3. 計算布隆過濾器之陣列大小(m)，在本例中假設 m=6463。
4. 計算 "存在誤判"之機率，在本例中"存在誤判"之機率為 0.0078，亦即平均查詢 1000 次會發生 8 次存在誤判之情形。
5. 大數據應用服務供應商選擇是否接受該"存在誤判"機率，若接受則完成布隆過濾器參數設定；若不接受則重新回到第一步，設計較大的 k 值，即可得到更小的"存在誤判"機率。

在本研究中，資料檔案儲存系統裡的實體資料檔案不會重複，因此 n_{max} 可視為該資料集之實體檔案最大數量，而為了能夠動態在布隆過濾器上進行刪除動作，因此本研究採用的陣列為整數陣列。傳統布隆過濾器不管有幾個資料集檔案指到同一位置，該位置皆設定為 1，在進行刪除時很難判別是否還有資料集檔案映射至該位置，因此在 LBF 中採用整數陣列，以表示還有幾個資料集檔案指向該位置，在進行查詢時，k 個映射指向之位置若不為 0，即判斷該檔案存在，如圖 9 所示，資料檔案 A 判定為存在。

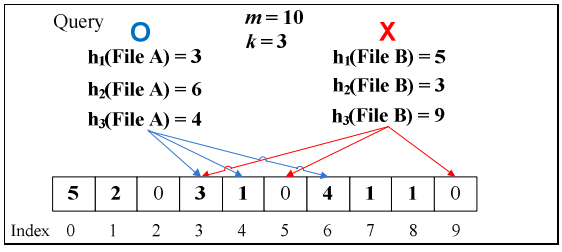

圖 9、以 LBF 進行資料檔案 A 是否存在的判定

圖 9 之布隆過濾器陣列中，陣列索引編號 3 的內容值為整數 3，表示有 3 個資料檔案指向該位置，若使用者後續使用時將資料檔案 A 刪除，則其所映射位置之值將會減 1，如圖 10 所示。

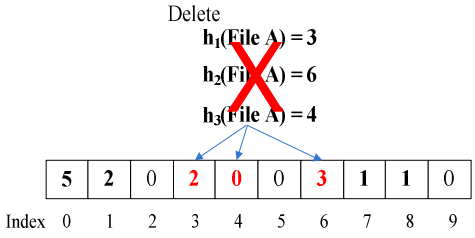

圖 10、刪除資料檔案 A 之結果

本研究所提出 LBF，直接採用鏈結串列進行存放資料檔案，當進行新增資料檔案時，將資料檔案之唯一識別碼透過 k 個雜湊函數運算並映射至布隆過濾器上，接著使用 k 個運算完的雜湊值進行運算，決定鏈結串列之起始位置。

整體而言，TPDP 的第二階段會將不活躍資料的唯一識別碼與欲存放該資料之資料集，進行資料檔案是否重複的比對。不活躍資料的唯一識別碼將被輸入至 LBF 進行處理，判斷該唯一識別碼是否存在。若是第一次使用，則需要進行初始化動作。以資料集編號 000 為例，假設共有 7 個雜湊函數(k)，該群組可容忍之最大元素數(n_{max})為 2880 (c GB/ b MB = 180 GB / 64 MB)，經計算可得出布隆過濾器之陣列大小(m) = 13974，及算出“存在誤判”機率為 0.151，若大數據應用供應商接受該機率則完成參數設定。而其他儲存群組透過相同方式，亦可計算出該儲存群組之布隆過濾器大小與其最小錯誤率。

接著，不活躍資料的之唯一識別碼透過 k 個雜湊函數的運算，若運算完的結果皆映射至“非 0”的位置時，表示儲存位置擁有該唯一識別碼，並快速找出實體檔案存放之位置，若發現相同唯一識別碼，則不須上傳完整檔案，只需傳輸相關資料製作索引檔即可。只要有一個映射之位置為“0”，則表示沒有該實體資料，則需上傳完整檔案。而在此時有可能發生“存在誤判”的情形，但透過查詢後，可快速確定是否發生存在誤判的情形，將所造成的損失降到最低。建置完成的鏈結式布隆過濾器與相對應的儲存位置之關係如圖 11 所示。

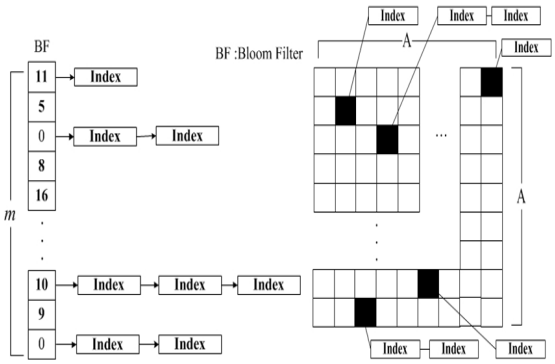

圖 11、LBF 與相對應的儲存位置之關係

5. 結論與未來研究

大數據具有三項主要特性，包括大量、快速、及多樣性。由於傳統的關聯式資料庫系統及在此系統上運作的資訊探勘、統計分析等各種技術，無法直接因應上述幾項特性，是以需要設計能夠大量儲存並實行平行運算的各種新方法來處理、利用大數據，以期找出大數據中蘊含的巨大價值(Value)。

依據朱揚勇與熊贇的研究，一個大數據應用需求，經常會遇到“6用問題” [2]。其中，資料不可用中的不活躍使用者之資料具有藍海競爭的潛力。因此，本研究所提出的在大數據相關應用中以「兩階段資料處理法(Two Phase Data Processing method；TPDP)」，針對在大數據中資料不可用問題中不活躍使用者資料的儲存和管理進行處理。TPDP的第一階段是以本研究所提出的MRFM找出不活躍使用者的資料；接著，在第二階段再以改良後的布隆過濾器(LBF)進行不活躍資料的儲存，以提供後續不活躍使用者資料的維護。

由於目前本研究主要是運用非同步的方式來達到不活躍資料的維護，在處理不活躍資料的儲存或查詢時，則會使用資料的活躍度以進行更新，因此可以更加符合在網路中資料檔案的快速傳播之性質。但由於大數據應用服務商提供多元化的服務，而不同的應用服務對資料活躍度的要求不盡相同，因此在未來的研究中將加入時間戳記(Timestamp)的機制，以時序的方式對於版本進行控管，並且在讀寫控制部分增加多人同時修改的功能，使得大數據應用服務能更加人性化。

參考文獻

[1] 王順生、王淑卿、陳聖中、楊欲富，“雲端運算環境中以檔案熱門程度的浮動門檻進行檔案副本儲存一致性的快速調節策略，”***第八屆智慧生活科技研討會(ILT2013)論文集***，pp. 1341-1348，2013.

[2] 朱揚勇、熊贇，“大數據是數據，技術，還是應用”，***大數據***，第 1 期，pp. 2015007-1~2015007-11，2015。

[3] 林陽助，***服務行銷***，鼎茂出版社，2003。

[4] 陳彤生，“運用改良 RFM 提升行銷效益的實證研究”，***第七屆人工智慧與應用研討會論文集***，2002。

[5] 陳聖中，***以熱門程度為基礎在雲端運算環境中建構可自適應檔案複製機制***，朝陽科技大學資訊管理系碩士論文，2013。

[6] 連惟謙，***應用資料分析技術進行顧客流失與顧客價值之研究***，碩士論文，中原大學資訊管理研究所，2003。

[7] Antichi, G., Pietro, A.D., Ficara, D., Giordano, S., Russo, F. and Vitucci, F., “Achieving Perfect Hashing through an Improved Construction of Bloom Filters,” ***Proc. of IEEE International Conference on Communications***, pp. 1-5, 2010.

[8] Baldor, S.A., “Applying a Cloud Computing Approach to Storage Architectures for Spacecraft,” ***IEEE Aerospace Conference***, pp. 1-6, 2013.

[9] Bloom, B.H., “Space/time Trade-offs in Hash Coding with Allowable Errors,” ***Commutations of the ACM***, Vol. 13, No. 7, pp. 422-426, 1970.

[10] Chan, W.K. and Mauborgne, R., ***Blue Ocean Strategy,*** INSEAD Blue Ocean Strategy Institute, 2005.

[11] Chen, T., Liu, F. and Xiao, N., “RADPA Reliability-aware Data Placement Algorithm for Large-scale Network Storage Systems,” ***Proc. of the IEEE International Conference on High Performance Computing and Communications***, pp. 648-653, 2009.

[12] Goodman, J., “Leveraging the Customer Database to your Competitive Advantage,” ***Direct Marketing***, Vol. 55, No. 8, pp. 26-27, 1992.

[13] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., Pierre, S.S., Twigger, S., White, O. and Rhee, S.Y.,“Big Data: The Future of Biocuration,” ***Nature***, Vol. 455, pp. 47-50, 2008.

[14] Hughes, A.M., “Boosting Response with RFM” ***Marketing Tools 5***, pp. 4-10, 1996.

[15] Jacobs, A., “The Pathologies of Big Data,” ***Communications of the ACM***, Vol. 52, Issue 8, pp. 36-44, 2009.

[16] Kahan, R., “Using Database Marketing Techniques to Enhance Your One-to-One Marketing Initiatives,” ***Journal of Consumer Marketing***, Vol. 15, pp. 491-493, 1998.

[17] Luigi, A., Antonio, I. and Giacomo, M., “The Internet of Things: A Survey,” ***Computer Networks***, Vol. 54, No. 15, pp. 2787-2805, 2010.

[18] Luo, T., Chen, G. and Zhang, Y., “H-DB: Yet Another Big Data Hybrid System of Hadoop and DBMS,” ***Lecture Notes in Computer Science***, Vol. 8285, pp. 324-335, 2013.

[19] Papadopoulos, A. and Katsaros, D., “A-Tree: Distributed Indexing of Multidimensional Data for Cloud Computing Environments,” ***Proc. of Third International Conference on Cloud Computing Technology and Science***, pp. 407-414, 2011.

[20] Paul, B., Lee, G., Alon, H., Jiawei, H., Marti, H. and Jure, L., “Channeling the Deluge: Research Challenges for Big Data and Information Systems,” ***Proc. of the 22nd ACM International Conference on Information & Knowledge Management***, pp. 2537-2538, 2013.

[21] Wei, J., Jiang, H., Zhou, K. and Feng, D., “DBA: A Dynamic Bloom Filter Array for Scalable Membership Representation of Variable Large Data Sets,” ***Proc. of 19th International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems***, pp. 466-468, 2011.

[22] Zikopoulos, P. and Eaton, C., ***Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data***, McGraw-Hill Osborne Media, 2011.

[23] 大數據簡介，www.nacs.gov.tw/.../3c2136cf0801ea10c866aaa770aa3e94.pptx, 2014/12/8。

[24] The Hadoop Archive, http://hadoop.apache.org/core/docs/r0.20.0/hadoop_archives.html。