國立交通大學

多媒體工程研究所

碩士論文

藉頭字語與引用文分析追蹤技術發展

Tracing Technology Development by Acronyms and Citation Analysis

研究生：林俞邦
指導教授：王才沛　教授

中華民國　九十九　年七月

藉頭字語與引用文分析追蹤技術發展

Tracing Technology Development by Acronyms and Citation Analysis

研 究 生：林俞邦　　Student：Yu-Pang Lin

指導教授：王才沛　　Advisor：Tsai-Pei Wang

國 立 交 通 大 學

多 媒 體 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Multimedia Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

July 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年七月

藉頭字語與引用文分析追蹤技術發展

學生：林俞邦　　指導教授：王才沛

國立交通大學多媒體工程研究所 碩士班

摘要

學術發展中的每樣技術，皆有不斷進化或衍生的可能性。當研究者因實驗需求而去探討某項技術時，往往必須耗費時間心力在閱覽大量學術文獻，以了解該項技術的發展境況。本論文的研究目的在於探討如何讓電腦利用頭字語和引用文(citation)提供的資訊，來對文件進行篩選並提取出有用的訊息，以幫助研究者有效率地掌握需求技術的後續發展脈絡。我們利用學術搜尋引擎Google Scholar的“被引用次數”搜尋功能，來找出所有與需求技術文件相關的學術文件，建立了三組測試資料組，並分別以人工閱讀方式找出符合我們實驗目的的文件集合，作為衡量實驗結果用的比對資料。在這份論文中，文件分析過程主要可分為三個主要部份：1.從文件中擷取並篩選頭字語、2.利用頭字語和需求技術關鍵字來對蒐集的文件進行篩選、3.對篩選後的文件進行引用文及共引文(co-citation)分析。我們將在論文後段從相關技術資訊取得和有用文件取得兩方面，來探討論文的實驗結果。

Tracing Technology Development by Acronyms and Citation Analysis

Student：Yu-pang Lin　　Advisor：Tsai-pei Wang

Institute of Multimedia Engineering

College of Computer Science

National Chiao Tung University

Abstract

Each technology developed in academic research has the possibility for further improvement or extension in the future. When a researcher investigates a certain technology due to the requirement of his/her own work, it usually takes a lot of time and effort for that researcher in surveying the literature to know the current status of that technology. The purpose of this thesis is to investigate how the information provided by acronyms and citations in the documents can be utilized for the automatic screening of documents and the extraction of helpful information. The goal is to help researchers more efficiently grasp the current development of the technology they need. We use the “Cited by” function of the academic search engine Google Scholar to find all the scholarly literature related to some given technical documents and established three sets of test data. To be able to evaluate our method, we manually read through all the citing documents to find the ones that are consistent with our purpose. In this paper, document analysis consists of three main parts: (1) The extraction and selection of acronyms from the documents; (2) the use the acronyms and keywords of related to the target technology to automatically screen the collection of citing documents; (3) citation analysis and co-citation analysis for the selected citing documents. We will discuss our experimental results on two aspects: to retrieve useful information about the

target technology, and the selection of citing documents that fit our purpose.

誌謝

這篇論文能夠順利完成，首先要感謝的，是我的指導教授王才沛老師，在我遇到研究上的問題或者是困難時，老師都會耐心的指導以及教誨，並且給予我適當的建議或是正確的解決方向，讓我對於論文的研究更加的順利，非常感謝老師的栽培。感謝陳祝嵩老師以及莊仁輝老師擔任我的口試委員，提供許多寶貴的建議，使得本論文更加完備，特此致謝。接下來要感謝的是實驗室一起研究與努力的夥伴們。和我一起口試的偉誌和崇桂，有了你們的陪伴，讓我寫論文的過程不會感到孤獨，在心情低落或是緊張時，也能互相打氣。也要感謝良佑、裕傑、俞丞和俊予，有了你們，讓我的碩士生活更加充實和精彩，最後我要感謝我的家人，謝謝你們能夠支持我完成碩士學業，沒有你們的支持與鼓勵，就不會有今日的我。

目錄

圖目錄

表目錄

第一章　背景與動機

了解研究中採納或開發技術理論的相關發展境況，是學術研究的必經過程。如何更有效率地從大量資料中整理出脈絡，這樣的研究議題便應運而生。各個發展中的技術理論，皆有可能在將來成為其他人的研究基礎之一，或被改良，或被應用於某項環節，然後這些新的成果，同樣也可能被他人用於研究，整個歷程便宛若一幅樹狀圖不停分支、衍伸下去。尤其現今科技日新月異，學術文件正快速地增加著。因此，當我們需要某項技術而對之進行探討的過程中，往往需要閱讀大量文獻來獲取關於該技術更進一步的資訊，甚至必須接觸與我們的議題領域大相逕庭的區塊，而耗費許多無謂的心力與時間。

如果想了解一篇文獻內容的後續發展，從引用它的文件中去發掘是相當好的選擇，因為引用行為正代表研究上的關聯性甚至承繼性。現今網路上有許多如 Google Scholar 和 CiteSeer[1]般的學術搜尋引擎，可以透過關鍵字搜索網上的論文資源，若檢索結果的文獻有被引用，還能查出是哪些文件所引用。然而在這資訊爆炸的數位環境，學術搜尋引擎提供的結果資訊量仍時常太過龐大。為了更有效率地得知某項技術的後續發展境況，若能讓電腦幫助我們將搜尋得到的資訊再進行一次篩選，僅留下具有高度關聯性的文件，尤其是以該項理論或技術為主要發展基礎的研究著作，將可以幫助研究者們在文獻探討這一步上，最大可能地免去人工負擔，更有效率且輕鬆地掌握所求技術的發展脈絡，事半功倍。而這同時也是為了確保能使用最先進的技術以提升效能、找到最適合使用者研究的方法，或避免重複的研究。這類目標在文件分類(document classification)或趨勢分析(trend analysis)等等議題中也有部分相關的研究。

在本篇研究中，我們將最初想進行後續探討的理論或技術論文，稱為“根文獻”；而自網路上蒐集、引用這一份根文獻的文件，稱之為“子文件”。我們希望能從子文件集合中，尋找關於該篇根文獻後續發展的線索。學術著作者會因為各種理由而引用一篇文獻，同一份被引用文獻對於不同的引用文件(citing document)貢獻也不盡相同。Moravcsik 等人在[2]的研究中，從 *Physical Review* 隨機抽取三十篇文章作樣本，對裡面所有參考文獻的引用方式進行分析：是觀念層面(conceptual)或操作性層面(operational)的引用，是必須

的(organic)或應付了事的(perfunctory)引用，是為引用文件的論點所認可(confirmative)或質疑(negational)…等等。其中“對引用文件來說是必須性的或可有可無的”這項分析的結果分別是 69%和 31%，也就是說冗餘的文獻引用佔了將近三分之一。另外還有“提供引用文件發展的基礎(evolutionary)，或只是選擇之一、並列的(juxtapositional)”，比例分別約為 59%和%40。回來談到本篇的實驗，我們希望在經過篩選後，存留下的子文件與根文獻間的關聯是“根文獻提供子文獻發展的基礎，而子文件是對根文獻所述技術進行改進或衍生”。從 Moravcsik 等人的研究不難猜想出，單純藉由學術搜尋引擎蒐集的引用文件，至少會有將近一半不符合我們需求。我們希望在本篇實驗中盡可能地提升篩選文件的準確度，將符合我們需求的文件留下最多，參雜的不符合目的文件數量減到最少。

目前尋找與根文獻具有特定相關性的子文件的作法，主要是利用引文功能(citation function)的角度去分析。在論文中，凡引用到另一份文獻內容的文句，皆會附上引用標籤或作者年份資訊，使閱讀者能在參考文件列表中對照尋找。因此若去分析子文件中根文獻引用標籤前後的字詞文句，或許能找出關於兩者間的蛛絲馬跡，如[3]，只是此種方法必需依賴大量的提示字詞(cue phrase)資料庫。

而我們所想嘗試的，則是利用頭字語來尋找子文件中主軸的理論或技術名稱，並觀察與根文獻技術的關聯性。在一般學術文件中，頭字語常作為原詞彙的同義詞使用，使文件更容易閱讀[4]，比如下例：

> *The greatest difference between the **FCM**-based and **PCM**-based algorithms is for the case where there is but one cluster in the data set. In this case there is essentially no difference between the **FCM**-based methods and hard methods. Figs. 8 and 9 illustrate this idea. Fig. 8 shows the estimates of the prototype parameters for a noisy line when the **FCV** and **PCV** algorithms are used. The estimates of the **FGK** are severely affected by noise. ……(節錄自 A possibilistic approach to clustering; 1993)*

由於我們在實驗中著重“技術發展”這部份，若能將文件中的頭字語取出並分析其組成，不失為一種能達成我們實驗目的之方法，比如下例：

> ***This paper focusses on a new clustering method called evidence accumulation clustering with dual rooted prim treecuts (EAC-DC)**, based on the principle of cluster ensemble also known as “combining multiple clustering methods”. ……(節錄自 Combining multiple partitions created with a graph-based construction for data clustering; 2009)*

假設實驗所要探討的是 Evidence Accumulation Clustering(EAC)這項技術，在取得例子中

的頭字語 EAC-DC 後，發現該頭字語包含“EAC”這三個字，我們便可推測包含該頭字語的文件或許是我們所想要的結果之一。另外，研究者在摘要或其他方第一次對研究目標進行介紹時，若成果是能以一具體名稱喚之，多會如同上例中順勢定義出頭字語。但因為文章中其他輔助的技術或理論也是相同情況，所以如何在這麼多頭字語中找出前者，便是我們研究中的另一項工作。在後面的章節我們將實驗並驗證這個想法的可行性。

第二章 文獻縱覽

2.1 引用文分析 (Citation Analysis)

一般學術論文必由正文及最末端的參考書目列表組成，這些參考書目即為這篇學術論文引用的文獻，而這些被引用文獻必也包含正文與參考書目列表。引文分析的理論即是建立於這種文件相互引用的關係上，並利用數學與統計學的方式，探討著正文和引用文或參考資料間的引證關係，進而討論相關性或引用特性，譬如引用或被引用的原因、動機，或關聯性程度多寡。其主要目的在於方便後人作研究時了解該份文獻所述內容的相關發展走向以及知識架構[5]。Liu 在 [6]中，將引文分析相關研究的目的歸納為五大類：強化引文索引、引用功能的描述、評估引文的價值、定義對引用文件的貢獻以及判斷引用的動機。以下僅討論與本篇研究較具相關性的引文索引與引用文功能兩項：

2.1.1 引用文功能分析 (Citation Function Analysis)

這項議題探討的是作者基於什麼樣的理由而引用某篇論文[3]。最早在電腦科技尚未發達時，研究者們以人工閱讀、郵件詢問或面對面採訪作者的方式，來取得引用文功能的相關資料，並進行分析[2]。Chubin 等人從物理期刊中挑選 33 份快報和 10 份論文，分別分析它們與引用文間的關係，並以基本需求、間接需求、額外補充、冗餘補充、部分反對、完全反對這六個層面來分類[7]。Oppenheim 等人挑選 23 篇高引用率的早期文件，並從 1974 到 1975 年間引用它們的文獻中隨機挑出 987 份進行分析，結果發現僅有 16% 是基於理論或技術上的用途而引用[8]。

為了能不依靠人工的手段來辨識引用文功能，研究者們嘗試將文件中一些包含明確意向的文詞提示與引用文功能作聯結，尋找是否有對應的關連性，例如引用文是使用在引用文件的哪個章節或被提及次數的多寡。Maricic 等人的研究結果顯示，在文獻介紹(Introduction)章節中出現的引用文，通常是粗略提及、不是那麼重要的；較有意義或較具關聯性的引用文主要是在方法、結果、和討論等章節出現[9]。Hooten 的研究則指出，

若某份引用文在同一項研究中被提及越多次，便可依此推斷該份引用文和引用它的文件具有密切關係[10]。不過 Hanney 等人則對此抱持有質疑的看法[11]。另外 Simone Teufel 等人在[3]中，將引用文功能歸納成數類，並以語言學中“cue phrase”概念，藉由一些對於後續文句內容有提示功能的字詞(例如：adopt, agree with, base, be based on, be derived from, be originated in, be inspired by, build on,…)，來對照所屬引用文功能類型。不過直至今日，關於這些文詞提示是否真的能正確地判斷出引用文功能，還仍沒辦法明確地證明[12]；儘管如此，仍有不少研究者投入這項研究，尋找更具說服力的方法與可能性。

2.1.2 引用文索引(Citation Index)

引用文在學術文件間可視作一種關聯性上的連結，為了利用這種連結架構來搜尋相關的文件群，早期便有了引用文索引這類工具的研究[13]。到了近期，藉由全自動引用文索引線上搜尋工具，例如 Citeseer[1]和 Google Scholar，研究者可以輕易地蒐集具有某種共通關聯的學術文件。另外，為了能尋找引用同一篇引用文的文件，便產生了共引用分析(co-citation analysis)，藉此我們可以從兩篇文獻的共同引用數目來衡量關聯強度，例如在[14] 中，Strohman 等人將文件間共引用的資訊視作一種特徵，讓搜尋系統在使用者輸入一份文件資料後，可以獲知這份文件中哪些引用文具有較高閱讀價值。Meij 與 de Rijke 則在語言模組中，利用引用次數(citation counts)來衡量相關文件的事前機率(prior probability) [15]。

2.2 頭字語分析 (Acronyms)

頭字語(acronyms)，也稱首字母縮略字，是由其他一連串單字的第一個字母所組成的字 [16]，用以代表原本冗長的實體名稱，本篇論文中我們以“全名”來稱呼這個實體名稱，例如“CPU”即是代表全名為“Central Processing Unit”的頭字語。目前學界對頭字語的主要研究工作可分為兩類：第一類是探討頭字語的組成特性，以辨認並取出文本中頭字語為主。由於目前頭字語的組成方式複雜化，已不再單純由第一個字母組成，而可能從

全名中其他位置擷取有意義的片段字母加以組成，於[17]中 Larkey 等學者提出 contextual、 canonical、 canonical/contextual 以及 simple canonical 這四類演算法，來針對不同特性(比如是否包含小寫字母)的頭字語進行辨認。第二類是頭字語辨識(acronym identification)，目的是取得文本中包含的頭字語與其全名[18]，其中也包含將頭字語還原成全文型態的擴展(acronym expansion)這項重要議題。另外也因為頭字語多樣的組成範疇，導致同一全名可能擁有不同長度的頭字語、或不同全名擁有相同的頭字語的狀況發生，是這議題最需要克服的難題[19]。 [20]中利用頭字語-全名對照表，來實現一種利用各式預先訂好的限制條件來尋找頭字語並配上適合的全名，這方法在辨識那些沒有成對出現的頭字語或全名時，特別有效。

文獻中重要的技術或理論，其全名在第一次出現時，通常會搭配一個括號附註其頭字語寫法，以方便後續內容的撰寫。本篇實驗將利用這個特性來尋找文本中與內容具高度相關性的頭字語。

2.3 文件分類 (Document Classification)

由於現代資訊數位化的影響，文件數量增長過大而不再適合以人工方式進行分類，因此文件分類此一議題更顯重要。文件分類，顧名思義，指的是依照文件內容來歸類成數個類別。目前除了字面上的用途，也普遍應用於電子郵件或全球資訊網頁的管理上。在學術界中，則提供了提供一種篩選及統整文件的方式，讓研究者們能更有效率地找到相關學術文件群。文件分類在方法主要可以歸類為兩大類：第一種是以關鍵字為基礎(keyword-based)，以事先定義好的關鍵字彙資料庫來對文件進行比對並分類，為較傳統的方式。不過現今學界以第二種、也就是以內容為基礎(content-based)或以特徵為基礎(feature-based)的文件分類[21]研究居多。這類技術使用特殊的演算法，從文件中分析並擷取出足以代表這份文件的特徵集合(例如文字出現頻率、文本架構乃至於圖片)，並依此作分類。最普遍的為 Salton 等人在[22]中提出的 TF*IDF，藉由計算字彙在個別文件與所有文件中出現頻率這兩種值，來判斷這個字彙對這份文件是否具代表性。另外 Bader

Aljaber 等人在[23]中，提到藉由“bag-of-words model”這種標籤分類機制與分群法的組合，有助於將含有特殊關鍵字且與研究之主題高度相關的文件篩選出來。另外也有 Small 等人結合前幾節所述引文架構，應用到開發對資料庫文件進行分類的系統中[24]。

第三章　實驗方法

3.1　建立測試資料組

3.1.1　文件蒐集

為了驗證我們在第一章提出的想法，首先必須建立作為測試用的資料。為了有助於之後的人工篩選與分析探討，我們從自身熟悉的領域中，挑選具有適量被引用次數(至少 150 次)，且技術被廣泛使用的學術文獻作為根文獻。我們將根文獻標題輸入 Google Scholar，找到目標後點擊下方“被引用次數”連結，顯示引用文件列表，即是我們所要蒐集的子文件。我們希望將每組文件數量控制在 200 份內，若 Google Scholar 顯示的引用文件超過這數量，便依循下列原則來挑選：

i. 略過書本和非英語系文件，

ii. 選擇發表於期刊上的文獻，

iii. 會議文獻的話，引用次數需高於預設閾值(實驗中設定為 25)才選入。

我們以人工方式盡可能自 Google Scholar 提供的連結或其他學術相關網站取得這些文件的電子檔，並利用提供線上轉檔服務的網頁 *PDFTextOnline*[25]來進行 PDF 轉換成純文字格式檔案的工作，這是為了方便之後程式讀取以及分析。由於部分早期文件的電子檔僅僅是文本的掃描圖檔，我們也將剔除這類文件，而轉檔發生錯誤的文件也會予以刪除。

3.1.2　建立比對資料

為了確認程式分析結果的準確度，我們仍需先依靠人工閱覽來判斷子文件是否滿足實驗目的：以根文獻為基礎發展改進或衍生的技術。之後的程式參數調整和結果的比對分析也會使用到這項結果。人工判斷的依據主要可歸類為下列幾項：

i. 觀察子文件研究之主體(技術)名稱：若包含有根文獻技術名稱或關鍵單字，有很高機率滿足實驗目的。如 Competition Algorithm 為根文獻技術名稱，子文件研究主體為 Robust Competition Algorithm，則可判定該子文件滿足實驗目的；另外在這個例

子中，關鍵單字為“Competition”，是為根文獻所述技術之特徵。

ii. 觀察摘要、方法與結論這三個章節：如同在 2.1.1 節所提到，重要的引用文通常會於這些章節被頻繁提及。根文獻技術若為子文件研究的必要基礎或主軸，子文件著作者多會於摘要或結論的前幾段提及，或是在方法章節經常性的出現根文獻技術名稱和引用標籤(reference tag)，甚至以根文獻技術名稱為章節標題。

iii. 某些特定的提示字詞：一段話中若有提及引用文，則在提及之前的某些字詞可能暗示著子文件研究內容和根文獻技術的關係。常見如 “extend the”、“enhanced”、“improved”、“based on”…等等。

分類完後記錄下滿足實驗目的的文件標題，即完成比對用資料組的建立工作。在之後的文章中，我們皆以“比對組”來稱呼這些通過人工方式篩選、符合我們實驗目的的文件。

3.2 頭字語搜尋與處理

3.2.1 搜尋頭字語

頭字語在文件中第一次提及時，通常以 *Full-name (acronyms)*的型式被介紹[26]，例如：“*We present a new clustering algorithm called Robust Competitive Agglomeration (RCA).*”。我們希望僅取出文件中的主體技術名稱，除此外的不必要資訊越少越好；基於這個緣故，我們只對括號內的字串進行是否為頭字語的判斷：

> 於句子中找到左括號後，在遇到右括號前：
>
> - 以空格為區隔，找到包含連續三個以上大寫字母的連續字元
> - 搜尋範圍不超過左括號的六個字元，但遇到空格重新計算
>
> 當滿足上述兩項條件時，直接中斷搜尋並將該連續字元視為頭字語。

為了避免搜尋到過多不需要的頭字語，導致實驗結果產生偏差，我們將搜尋範圍限制在標題和摘要部份。除了這項根本因素外，根據人工分類時的觀察，我們制定這項限制的理由尚有：

i. 文件本身主體技術和發展基礎技術多會在摘要部份便提及。

ii. Introduction 章節雖然也會提及，但同時更包含大量無助於實驗結果的資訊。對頭字語占有重要地位的本實驗來說，是最需要迴避的區塊。

iii. 只有極其少數的文件會於方法等章節才首次介紹文件主體技術的頭字語(約十幾分之一的機率)。由於這類章節多有一定份量的內容，與根本因素相衡量下，我們也迴避這個區塊。

iv. 實驗結果與分析比較這類章節的內容，明顯地和我們的實驗目的較無關聯。

v. 文件中最後對整篇研究作總結的章節，除了章節標題類型繁多、導致不易判定起始行數這項變因外，文件主體技術頭字語在這章節宣告卻未於摘要提起的情況，也是極其少數。另外由於這章節有回顧論文所有工作的特性，文件中許多不符合實驗需求的頭字語會在此重新提起，而這也違反了前述根本因素考量。

在後面的 4-1 章節中，我們也會驗證這些想法，並証明將搜尋範圍限制在摘要是最適合本實驗的作法。

3.2.2 頭字語篩選

舉凡各式研究中的新技術，在吸收基礎技術優點同時，也會繼承其演算法特性。尤其是以改良或延伸某項固有技術為目的者，在名稱上更容易出現與前項技術相似之處，比如一些暗示著前項技術特性的關鍵單字。舉例來說：建立在 *Competitive Agglomeration (CA)*之上的 *Robust Competition algorithm(RCA)*、*Competitive Agglomeration for Relation DATA (CARD)*，皆能明顯看出同原技術演算法中“competition”的特性。在本篇實驗，將利用這個現象來篩選於 3.2.1 所記錄下的頭字語集合，希望藉此留下較可能與根文獻技術有關聯的頭字語。篩選主要有兩個階段，各階段規則如下：

篩選所需資訊：根文獻技術名稱 D_{root} 與頭字語 A_{root}，由使用者提供

例：“Competitive Agglomeration”和“CA”

第一階段：去除掉與 root_arc 沒有相同字母的頭字語

第二階段：對所有頭字語，檢察其全名是否包含與 D_{root} 或 A_{root} 相同或類似的單字，若有兩字以上相同便予以保留

例：(CA) competitive agglomeration

(RCA) robust competitive agglomeration

例：(PCM) possibilistic c-means

(RPCM) robust PCM

附註：兩個單字相似度計算方式

例：competition 與 competitive 作比對

相似度 = (相同部分長度/比對目標字串長度) = 9/11 ≒ 0.82

於本實驗中若相似度大於 65%，我們視為相同的字。

3.2.3 以頭字語和根文獻技術關鍵字作文件篩選

若文件不包含任一通過篩選的頭字語，將被從子文件集合中剔除。

例：於子文件 *Fully automatic clustering system* 中僅找到頭字語集合{FACS}，但 FACS 不包含於通過篩選的頭字語集合中，因此該篇文件將被過濾掉。

例：子文件 *Coherence criterion for region labelling and description* 中包含頭字語集合{ARC}，且 ARC 屬於通過篩選的頭字語，因此該文件將會存留下來。

並非所有作者皆習慣使用頭字語，或特意在著作中自行創造新的頭字語使用。但如同 3.2.1 開頭所述，不只技術名稱，標題中有時也會隱藏著線索。因此在利用頭字語過濾文件的同時，我們以 3.2.2 第二階段的規則檢查文件的標題，希望藉此留下更多有用

的文件。

例：在 *Improved possibilistic c-means clustering algorithms* 中僅以“Improved PCM”來稱呼內容所述技術，而非其他文件所提的 IPCM。但我們可於標題中的 *possibilistic c-means* 發覺文件與根文獻技術 PCM 的關聯性。

3.3 引文與共同引文分析

本章節所處理的文件，為經過 3.2 章節刪減後的子文件集合。我們希望從這些具有滿足我們實驗目的可能性的文件中，藉由引用文分析找出他們彼此關聯性以及有幫助的引用文。我們認為在未來，當研究者探討或使用這些子文件時，這些關聯性或引用文，能提供研究者選讀上的依據、或關於子文件理解及實作上的支援，更進一步地提升文獻探討的效率。

3.3.1 各文件參考文件列表

目前學術文件的參考文件列表標籤主要分為三種類型：

[編號] *引用文資訊*

例：[3] Simone Teufel , Advaith Siddharthan and Dan Tidhar. 2006. “Automatic Classification of citation function”. In Proc. of EMNLP 2006, pages 103–110.

編號. *引用文資訊*

例：1. Simone Teufel , Advaith Siddharthan and Dan Tidhar. 2006. “Automatic Classification of citation function”. In Proc. of EMNLP 2006, pages 103–110.

引用文資訊 (沒有任何標籤)

例：Simone Teufel , Advaith Siddharthan and Dan Tidhar. 2006. “Automatic Classification of citation function”. In Proc. of EMNLP 2006, pages 103–110.

在文件讀取至參考文件列表起始位置後，我們逐一地將各個引用文組合並依序儲存起來。前兩種參考文件格式皆能依靠標籤來辨識引用文開頭與結束的地方，第三類型則需依靠該行是否出現作者姓名的方式來判斷。

例：*Andrews, D., P. Bickel, F. Hampel, P. Huber, W. Rogers and J. Tukey (1972)…*

例：*Zhang D, Pal SK (2002)…*

我們在觀察後數十份文件後，統整出如下作者姓名格式，作為檢查字串中是否包含作者資訊的依據：

[*upper_letter*] [,|.] [*not letters*]

即是說，只要有連續字元滿足“大寫字母後接著逗號或點，然後再下一個字元不為英文字母”的格式，便可認定該字串包含作者姓名。將這第三類型的文件參考列表分析結果與原文本對照，我們的方法準確度為平均每份文件漏失一篇以下的引用文。

此外，部分文件在參考文件列表後尚有其他資料，如附錄或圖表，因此有時會被誤判而視作該份文件的參考文件之一。

3.3.2 尋找引用文標題

為了之後方便比對子文件間是否具備引用關聯性或有共同的引用文，必須先將參考文件列表陣列的引用文資料簡化，僅僅留下標題部分作為識別，而之後在比對不同文件間引用文時，也將根據標題是否相同來作判斷。由於參考文件列表中引用文的撰寫格式相當多樣，間隔方式或使用的標點符號也不盡相同，而造成參考文件語法分析器(reference parser)研究者的困擾。但因為本篇實驗中只需切出標題這部分，因此以簡單的方式去分析取得即可。

輸入：參考文件列表陣列每項欄位中的字串

初始化：Title_Flag = 0

步驟一、如果字串包含(“ , ”)、(`` , ”)任一對符號，截取夾在符號中間的字串即為標題，Title_Flag = 1。若含兩對以上則截取第一對

步驟二、

1. 找尋字串中首組未有標點符號間隔的連續三個(或以上)單字

例：**H. Frigui** and R. Krishnapuram, Clustering by Competitive Agglomeration,…
└有標點符號間隔

例：H. **Frigui and R.** Krishnapuram, Clustering by Competitive Agglomeration,…
└連續三字且無標點符號間隔

2. 檢查該組連續字串是否：

i. 開頭不為小寫字母，否則該組字串無效，返回 1.繼續尋找(標題部分首位字母必為大寫)

例：Z Zhang, R Deriche, O Faugeras, and QT Luong,…

ii. 末端單字不符合作者姓氏格式，否則該組字串無效，返回 1.繼續尋找

例：Z. Zhang, R. Deriche, O. Faugeras and Q. T. Luong,…

iii. 標點符號除外的末端字元不為數字，否則該組字串無效，返回 1.繼續尋找

iv. 滿足上三項條件且不包含逗號(,)，或包含逗號但連續單字數大於二、且逗號前字母為小寫，則 Title_Flag = 1，表示找到標題字組了

步驟三、

1. 如果 Title_Flag = 1，整理標題字串：

   i. “-”的處理，連接兩關聯單字或文件中行尾單字過長而被斷至次行的情形

   例：Semi-unsupervised，則保留“-”

   例：Clustering by Competitive Agglo-meration，需去除“-”

   ii. 去除頭尾的標點符號

   iii. 去除開頭為年份。處理下述格式的參考文件列表：

   例：H. Frigui and R. Krishnapuram, (1997) Clustering by Competitive Agglomeration, …

   iv. 去除開頭為作者名。處理下述格式的參考文件列表：

   例：H. Frigui and R. Krishnapuram: Clustering by Competitive Agglomeration, …

2. 如果 Title_Flag = 0，表示在步驟二未能找到標題。主要原因有兩個：該欄位非屬於參考文件列表內容(附錄或頁面註解)，或標題僅由兩個單字組成。但在此皆視作後者進行處理：

   i. 以點和逗點兩種符號對字串進行切割

   ii. 檢查切段是否開頭為大寫字母且末端單字不符合作者姓氏格式

   iii. 切段長度是否大於 9。在使用的三組測資中，我們所發現最短的引用文標題為 10 字(“Fuzzy sets”;1965)，作為最低長度閾值。

   iv. 若切段滿足 ii 和 iii 兩項條件，則視為找到標題字組，否則不予理會

步驟四、找到標題字組的話，回存參考文件陣列同欄位

3.3.3 建立引用關係表

符合實驗目的的子文件中，彼此也可能有承先啟後的關聯性；我們以暴力法直接將子文件的標題和所有參考文件列表中的引用文標題(除了所屬自己以外)作比對，並整理成如表 3-1 範例格式。D 為篩選後的子文件集合，上方橫向欄位是被集合中其他文件引用的子文件，左方縱向則為引用方。舉例說，範例中的 D_1 引用了 D_3 和 D_{10} 兩份文件。藉由觀察這些文件間的引用關係，有助於我們衡量文件的重要性。

表 3-1：子文件間引用關係表格範例

	D_1	D_3	D_5	D_{10}
D_1	0	1	0	1
D_5	0	1	0	1
D_9	1	0	1	1

3.3.4建立共引用關係表

符合實驗目的的子文件間可能存在共同的引用文，而這些引用文之中，或許存在一些有用的訊息，在研究者將來實際應用根文獻時能作為觀念和技術上的輔助、或提供相關資訊，因此我們嘗試以共引用分析的方法來找出這類引用文。在這部分我們同樣以暴力法直接檢查，從參考文件陣列的最上方第一個欄位開始，每行的欄位都會與其下方所有欄位作比較；若發現相同標題，則將該欄位標上記號，之後比對時跳過該項。共引用關係表為一布林陣列，行數為被共引用的引用文總數；列數為符合實驗目的的子文件總數。

第四章　實驗結果與分析

我們選了三篇論文作為根文獻：*Clustering By Competitive Agglomeration*(H. Frigui 等人，1997)，*Data Clustering Using Evidence Accumulation*(L.N. Fred 等人，2005)，*A possibilistic approach to clustering*(R. Krishnapuram 等人，1993)，並建立子文件相關資料如表 4-1：

表4-1：測試資料集

TestData.1	TestData.2	TestData.3
Clustering By Competitive Agglomeration	A possibilistic approach to clustering	Data Clustering Using Evidence Accumulation
主體技術(頭字語)		
Competitive Agglomeration(CA)	Possibilistic C-Means(PCM)	Evidence Accumulation Clustering(EAC)
比對組數量 / 資料集總數		
12 / 114	**14 / 171**	**8 / 103**

4.1 藉由頭字語尋找以根文獻為基礎發展的技術

表4-1中顯示各測試資料比對組(人工篩選的結果)的數值。而下面的表4-3、4-4、4-5，則將顯示我們的系統分析測試資料後取得成果。在這裡，我們使用資訊擷取理論中用以評估分類績效的精確率/召回率 (Precision/Recall)[27]：

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

若將上述參數對照我們的實驗，tp即是通過人工與系統兩項篩選的文件集合，fn為通過人工篩選但被系統過濾掉的文件集合，fp則是通過系統篩選、但事實上並不符合實驗目

的之文件集合。

首先，我們將兩種頭字語的搜索範圍：全文(不含參考文獻)，以及僅限摘要部份，兩者的效果進行比較，以確認哪種搜索方式可以使我們的方法或得最佳效益，如表4-2。

表4-2 全文搜索與摘要搜索成果比較

	測試資料組一		測試資料組二		測試資料組三	
系統篩選文件剩餘數	25	16	20	11	11	6
比對組∩系統篩選	13	11	11	10	4	4
精確度	52.0%	68.8%	55.5%	90.9%	36.4%	66.7%
召回率	100%	91.7%	78.6%	71.4%	50.0%	50.0%

*各組下方的欄位，左欄為全文搜索，右欄為摘要搜索結果

從表4-2中我們可以看出，全文搜索在召回率上較僅限摘要部份搜索略佳，但在精確度上卻遠不如摘要搜索。平均每多找出一份符合實驗目的的文件，篩選文件結果會相對增加7.7份不必要文件，因此我們在衡量精確度與召回率間得失後，決定僅僅以摘要部份作為搜索頭字語的範圍。詳細分析結果如下表4-3、4-4、4-5。

表4-3

(a) TestData.1分析結果

頭字語 / 包含該頭字語文件數	▼符合實驗目的	
	RCA (Robust Competition algorithm)	2
	CARD (Competitive Agglomeration for Relation DATA)	3
	PCCA (Pairwise-Constrained Competitive Agglomeration)	4
	ARC (Adaptive Robust Competition algorithm)	5
	▼非實驗所求	
	CAA (Competitive Agglomeration Algorithm)	2
標題比對	符合實驗目的	0
	非實驗所求	0
系統篩選總量 / 總數	16 / 114	
比對組∩系統篩選	11	
精確度 / 召回率	68.8% / 91.7%	

*標題比對數量為扣去與頭字語重複部分結果

(b) TestData.1通過系統篩選的文件列表

[RCA] A Robust Competitive Clustering Algorithm withApplications in Computer Vision

[RCA] Robust fuzzy Gustafson-Kessel clustering for nonlinear system identification

[ARC] Coherence criterion for region labelling and description

[ARC] Image database clustering with SVM-based class personalization

[ARC] Region labelling using a Point-Based Coherence Criterion

[ARC] Unsupervised categorization for image database overview

[ARC] Unsupervised robust clustering for image database categorization

[CARD] Clustering and aggregation of relational data with applications to image database categorization

[CARD] Extracting web user profiles using relational competitive fuzzy clustering

[CARD] Mining Web Access Logs Using Relational Competitive Fuzzy Clustering

[PCCA] Fuzzy Clustering with Pairwise Constraints for Knowledge-Driven Image Categorization

[PCCA] Semi-Supervised Fuzzy Clustering with Pairwise-Constrained Competitive Agglomeration

[PCCA] Semi-supervised image database categorization using pairwise constraints

[PCCA] Some Pairwise Constrained Semi-Supervised Fuzzy c-Means Clustering Algorithms

[CAA] Lossless and Gradual Coding of Hyperspectral Images by Lifting Scheme

[CAA] Multiscale Retrieval with partial query of multispectral satellite images

附註：底線標示者為不符合實驗目的之文件

在第一組測試資料結果4-3a中，系統取得了五項頭字語，前四項正為我們所想要的，且囊括了比對組文件內的所有主體技術。而比對組的文件在系統篩選過程中，擁有九成以上的召回率，以相當高的比例存留下來；僅有一份因為主體技術沒有明確稱呼而未能通過系統篩選，該份文件必須依靠觀察文句敘述才能判斷與根文獻的關係。在通過系統篩選的文件中，除了標籤為CAA的兩份文件外，另外三份屬於ARC或CARD標籤的文件，僅是純粹使用ARC或CARD技術而沒有進一步地改良或發展，因此未在人工篩選時納入。第五項CAA實為根文獻技術CA的另一種寫法，系統無法分辨出而錯誤納入。在本份資料中，精確度雖只有約七成，但系統篩選出的十六份文件中，扣去所求的十一篇外，僅有兩份是完全偏離目的，另外三篇仍與我們想要的技術有所關聯。

表4-4

(a) TestData.2分析結果

頭字語 / 包含該頭字語文件數	▼符合實驗目的	
	PNFCM (Possibilistic Neuro Fuzzy C-Means algorithm)	1
	FPCM (Fuzzy-Possihilistic C-Means)	7/3
	SPCM (Similarity-based PCM)	1
	PFCM (Possibilistic-Fuzzy C-Means)	1/2
	EPCM (Enhanced PCM version)	1
	FPHN (Fuzzy Possibilistic Hopfield Network)	1/2
	KPCM (Kernel Possibilistic C-Means)	1
	RPCM (Relational PCM)	1
	PFPCM (Penalized Fuzzy Possibilistic C-Means)	1/3
	CFPCM (Compensated Fuzzy Possibilistic C-Means)	1/3
	▼非實驗所求	
	None	0
標題比對	符合實驗目的	1
	非實驗所求	1
系統篩選總量 / 總數	11 / 171	
比對組∩系統篩選	10	
精確度 / 召回率	90.9% / 71.4%	

*分數代表該頭字語的所屬文件包含複數個頭字語，為文件數平均分配後的結果

*標題比對數量為扣去與頭字語重複部分結果

(b) TestData.2通過系統篩選的文件列表

[EPCM] An enhanced possibilistic C-Means clustering algorithm EPCM

[FPCM] A mixed c-means clustering model

[FPCM][FPHN]Fuzzy possibilistic neural network to vector quantizer in frequency domains

[FPCM] [PFCM] A possibilistic fuzzy c-means clustering algorithm

[FPCM][PFPCM][CFPCM] Vector quantization in DCT domain using fuzzy possibilistic c-means based on penalized and compensated constraints

[KPCM] Kernel based fuzzy and possibilistic c-means clustering

[PNFCM] A fuzzy clustering based segmentation system as support to diagnosis in medical imaging

[RPCM] On relational possibilistic clustering

[SPCM] A novel similarity-based fuzzy clustering algorithm by integrating PCM and mountain method

[x]Improved possibilistic c-means clustering algorithms

[x]The possibilistic c-means algorithm: insights and recommendations

附註：底線標示者為不符合實驗目的之文件

在第二組測試資料結果4-4a中，沒有多餘頭字語的問題；但與比對組內文件對照後，發現有幾項頭字語因作者未明確宣告、或在摘要之後的部分才提到這等等因素，沒能被系統所發現，也導致出處文件未能通過篩選；雖然召回率因這個關係而僅有七成，但於本組資料中，卻擁有九成的精確度。本組在篩選後冗餘的文件僅有一份，其內容純粹為PCM的概觀論述，而非根文獻技術上的延伸或改進。

4-4b 最後兩項，因系統判斷標題與根文獻技術具有強烈關聯性而存留下來，其中 *Improved possibilistic c-means clustering algorithms* 主體技術為 Improved PCM(IPCM)，是屬於符合我們實驗目的的技術。

表4-5

(a) TestData.3分析結果

頭字語 / 包含該頭字語文件數	▼符合實驗目的	
	EAC-DC (Evidence Accumulation Clustering with Dual rooted prim tree Cuts)	1
	▼非實驗所求	
	None	0
標題比對	符合實驗目的	3
	非實驗所求	2
系統篩選總量 / 總數	6 / 114	
比對組∩系統篩選	4	
精確度 / 召回率	66.7% / 50.0%	

*標題比對數量為扣去與頭字語重複部分結果

(b) TestData.3通過系統篩選的文件列表

[EAC-DC] Combining multiple partitions created with a graph-based construction for data clustering

[x] An Evidence Accumulation Approach to Constrained Clustering Combination

[x] Definition of MV load diagrams via weighted evidence accumulation clustering using subsampling

[x] Sub-Space Clustering and Evidence Accumulation for Unsupervised Anomaly Detection in IP Networks

[x] Classification Rules Obtained from Evidence Accumulation

[x] Computation of initial modes for K-modes clustering algorithm using evidence accumulation

附註：底線標示者為不符合實驗目的之文件

最後討論到第三組測試資料 4-5a，與前面兩組比較起來，第三組的結果明顯較不理想。精確度與召回率雖皆有五成以上，但比之前兩組測試資料仍遜色許多。再討論原因前，我們先看到表 4-6：

表4-6　各組測試資料頭字語的篩選前數值

	TestData.1(CA)	TestData.2(PCM)	TestData.3(EAC)
頭字語個數	97	121	56
使用總次數	183	219	71

從上表可以明顯看出，第三組無論是數量或是使用次數，皆遠低於前面兩組，而且這也正是造成本組實驗結果較不理想的原因。於表 4-5a 中可以看到，僅有一項技術被發現，而其他文件皆是因標題具關聯性而通過篩選；我們分析這些文件，未發現頭字語的原因分別是：(1)於方法章節中提及而並非摘要，(2)技術全名為標題中的 *weighted evidence accumulation clustering using subsampling*，但於文件中卻直接使用頭字語 WEACS，並未再特別宣告兩者關係，(3)技術未有具體頭字語，最簡單的描述為 Evidence Accumulation for UAD。

4.2 子文件間的引用關係

通過篩選的文件，彼此間可能也存在著引證關係。透過建立他們間的引用關係架構，或許會有利於了解技術發展過程，或提供研究者一些想法。圖 4-1 為利用系統建立的引用關係表，所完成的關係圖，未出現的文件則代表沒引用他人也沒被他人所引用。

(Cited Doc. → Citing Doc.)

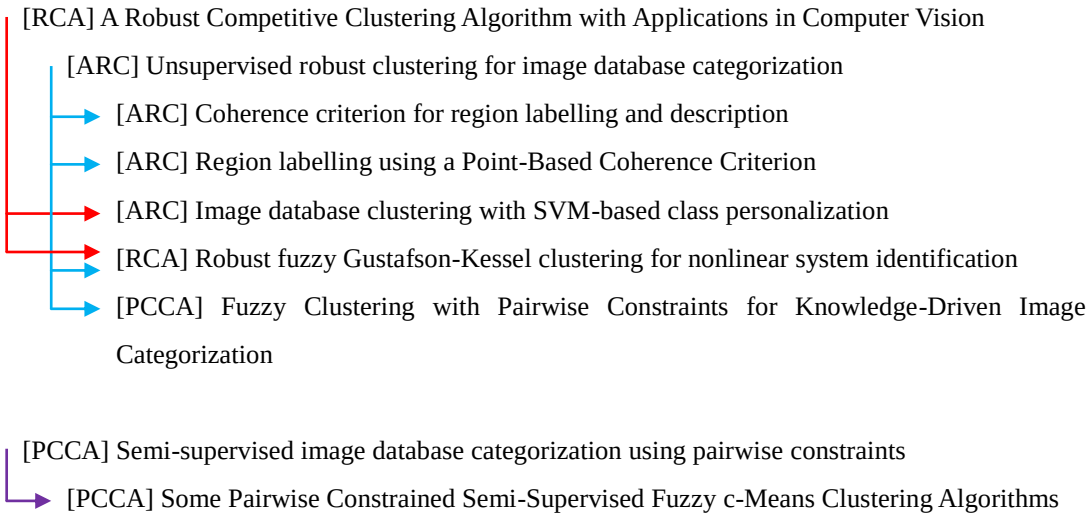

圖4-1 TestData.1子文件間引用關係圖

圖 4-1 為第一組資料產生結果。從圖表中我們可以看出位於箭頭根部的三篇各是該系列衍生技術中，較基層的文件。當使用者意欲探討其中某項技術時，該三篇文件便適合做為使用者入門或閱讀輔助的選擇。反過來說，也可以知道包含某項技術最新發展資訊的文件是哪些。

(Cited Doc. → Citing Doc.)

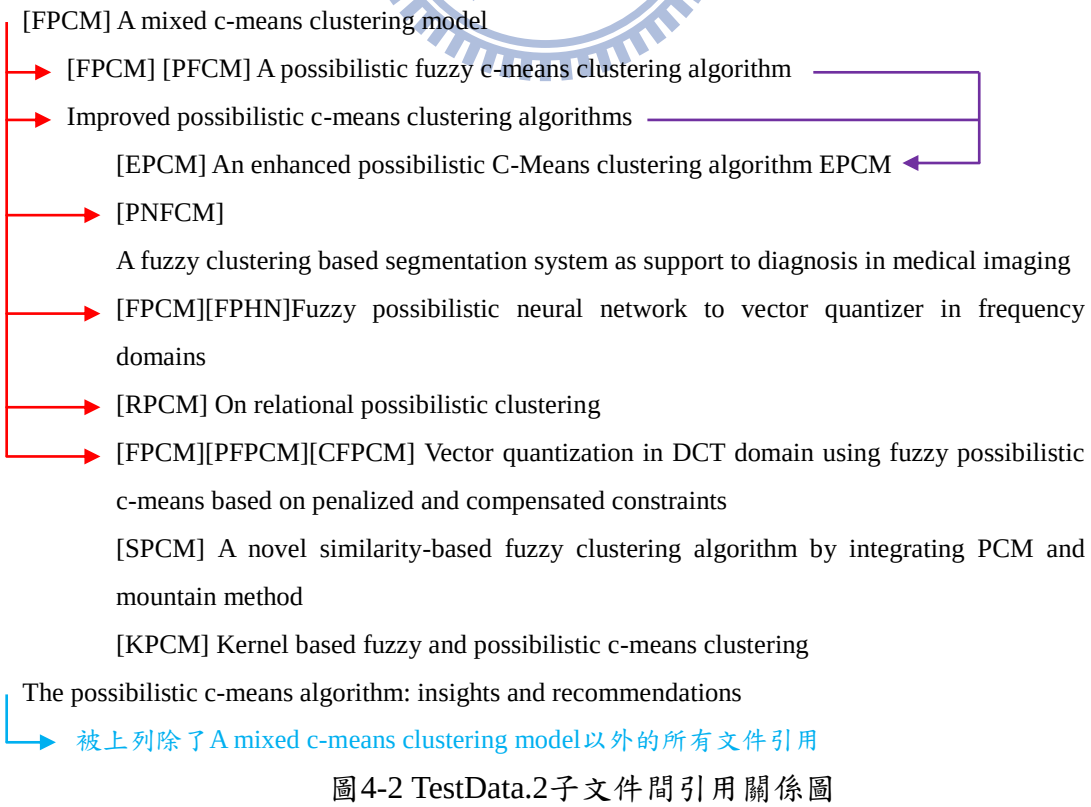

圖4-2 TestData.2子文件間引用關係圖

圖4-2為第二組資料產生結果。從圖表中可以看出，具有FPCM標籤的文件皆引用了*A mixed c-means clustering model*，可見是相當熱門的FPCM研究文獻，是研究者探討FPCM這項技術時必讀的文獻。再觀察那些引用他的文件，可以發現其中存在著各式各樣的FPCM衍生技術，可見得在PCM的發展樹中，FPCM是非常重要的一個節點，其底下可能擁有龐大的分支數量。*The possibilistic c-means algorithm: insights and recommendations*擁有極高被引用率，從標題可猜想，應是理論概念方面的引用功能居多。雖然這篇論文在本實驗中為錯誤判斷的文件，但若使用者有了解技術背景這方面需求時，也許將是個好選擇。

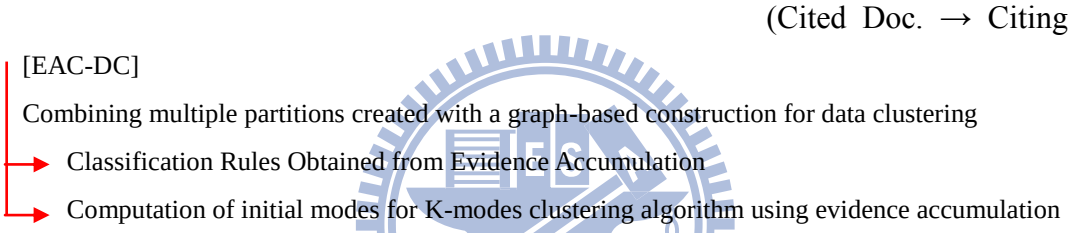

圖4-3 TestData.3子文件間引用關係圖

圖 4-3 為第三組資料產生結果，由於最初在文件篩選時成效便不佳，從圖中無可供參考的資訊，且下方兩篇皆是不符合實驗目的的誤判文獻。

4.3 子文件間的共引用關係

在前面章節曾提到，相關議題中越多論文共同引用的文獻，代表在這塊領域越有一定地位。本章節將分析 4.1 篩選出來的子文件參考文件列表，並找出可能對使用者有幫助的共引用文獻。在本節的三個表格中，橫向欄位皆是保留共引用次數超過兩篇的引用文，也就是在篩選出來的子文件中、被三份以上引用；注意的是根文獻(全體皆有引用)不算在內。

表4-7 TestData.1子文件間的共引用關係表

		B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
A1	*ARC*		1			1			1	1			
A2	*ARC*	1				1		1		1			1
A3	*ARC*		1			1			1	1			
A4	*ARC*	1					1					1	1
A5	*ARC*	1					1					1	1
A6	*RCA*	1											
A7	*RCA*												
A8	*CARD*		1	1	1	1	1	1					
A9	*CARD*			1	1								
A10	*CARD*			1	1								
A11	*PCCA*	1						1		1	1	1	
A12	*PCCA*	1									1	1	
A13	*PCCA*	1									1	1	
A14	*PCCA*										1		
A15	*CAA*								1				
A16	*CAA*												

*文件標題詳見附錄A索引編號; 引用文標題詳見附錄B索引編號

從表 4-7 中，我們可以看出有幾份引用文很明顯的，在某些技術探討的文件中常被引用。例如 B5、B9、B12 在與 ARC 有所關聯的文件中，皆有三份以上文件引用。接著看到探討 CARD 的文件中，B3、B4 兩份引用文在這類群被廣泛引用，但在其他文件中沒有引用紀錄；可見得是關於 CARD 技術中較特異、不同於其他的部分，這點從標題中也可推測出。具同樣特性的引用文還有 B10，在 PCCA 這類群中皆有引用記錄。上述的六篇引用文，由於皆具有高被引用率、以及只被特定技術引用的專門性，因此在研究者探討那幾項技術時，會相對的有利用價值，不管是從技術或觀念層面的輔助需求。引用文 B1 所探討的是分群時處理雜訊的問題，在 ARC、PCCA 皆有所應用，是為了提升效能而使用的輔助技術。A4 與 A5 兩份文件共引用了相同文件，從共引用分析的理論來看，這兩篇具有高度關聯性；因此如果研究者僅是需要大概了解技術，或許擇一閱讀即可。B8 和 B11 實質上為同一篇論文，但因標題單字使用不同而各自獨立；在 ARC 與 PCCA 兩項技術中被廣泛引用，也是篇相當具可讀性的技術文件。CCA 兩份文件原本便

屬於誤判，在關聯性上也明顯較少。

表4-8 TestData.2子文件間的共引用關係表

		B13	B14	B15	B16	B17	
A17	*EPCM*		1	1	1	1	
A18	*FPCM*	1	1				
A19	*FPCM*	1	1		1		*PFCM
A20	*FPCM*						*FPHN
A21	*FPCM*						*PFPCM *CFPCM
A22	*PNFCM*						
A23	*RPCM*				1	1	
A24	*RPCM*			1			
A25	*SPCM*			1			
A26	*IPCM*			1		1	
A27	*none*	1	1			1	

*文件標題詳見附錄A索引編號; 引用文標題詳見附錄B索引編號

表 4-8 的頭字語類別較多元且多是自成一派，而不像表 4-7 中各類技術皆有一定數量的相關文獻。因此在這裡，我們僅以“對於 PCM 的後續發展中，有幫助的引用文”這角度來分析。B14、B15、B17 皆為大宗的引用文，前兩項皆是技術開發性質的文獻，唯 B17 從標題來看，引用上會比較屬於概念層面。由於這些衍生技術皆是具有相同基礎與目的，因此這類熱門的引用文有時或許能對研究提供相關的輔助。另外在表中有個值得注意的地方是：A19、A20 和 A21 這三篇皆是以 PCM 的衍生技術 FPCM 為基礎，而再發展出更進一步的技術(後方附註)，可見得 FPCM 在 PCM 的發展中，占有一定地位。

表4-9 TestData.3子文件間的共引用關係表

		B18	B19	B20
A28	*EAC-DC*	1		1
A29	*none*	1		
A30	*none*			
A31	*none*		1	1
A32	*none*	1	1	
A33	*none*		1	1

*文件標題詳見附錄 A 索引編號

*引用文標題詳見附錄 B 索引編號

最後是表 4-9，前四項為符合實驗目的的子文件，後兩項則為誤判。這組資料量過於稀少，且前四份並沒有特別突出的共引用文獻，因此無法看出端倪、提供有用資訊。

第五章　結論與未來展望

為了更有效率地掌握技術的後續發展脈絡，若能讓電腦幫助我們將搜尋得到的資訊再進行一次篩選，僅留下以該項理論或技術為主要發展基礎的研究著作，將可以幫助研究者們在文獻探討這一步上事半功倍。學術論文在撰寫時，常使用頭字語來替代技術全名，以方便重複提及或易於讀者閱覽；我們注意到了這項特性，並在本篇實驗中加以利用：藉由分析學術文件內的頭字語的方式，來尋找以根文獻技術為基礎改進或延伸發展的技術和其名稱，並篩選出包含我們所需資訊的文件。我們的系統在前兩組測試資料中，能取得與人工篩選方式相當接近的成果，唯第三組因文件群使用頭字語不普遍而導致成效不彰，如何降低甚至避免這種現象帶來的影響，也是我們往後改進的方向之一。另外我們藉由分析參考文件列表，來觀察通過篩選的文件間彼此引證情形，不乏為一樣可以評估這些文件重要性的方式。另外也進行共引文分析，試圖找出與根文獻技術、或是各頭字語代表的技術，有所助益的相關引用文；雖然從結果中能獲取的資訊不如預期的多，但仍足供參考。從三組測試資料的分析過程，我們也發現到有喜好使用頭字語的學術著作者，但也有只是遵循參考文獻而使用、但本身不會特意創造新頭字語的作者，甚至偏好直接使用全名。頭字語雖然普遍的被使用，但僅是習慣問題而非規定，這現象直接影響到了我們的結果。雖然如此，對於這類憑著共同引用一篇文獻這項關聯性，所聚集起來、蘊藏著各式具有相同基礎與目的的技術之大量學術文件，除卻第三資料組那般的特例外，我們的方法在大多部份仍具有相當的效果。

因為文件中各章節的冗餘資訊繁多，因而我們在搜索頭字語時僅限摘要部分，若能找到方法解決這問題、來對全文進行搜索，相信將可大大地提升準確率。另外由於目前測試資料皆以人工建立，若能克服技術問題、讓系統能直接從網路上下載文件並自動轉為純文字格式檔，來提供系統分析，才是最人性化的方式。

參考文獻

[1] S. Lawrence, K. Bollacker and C. L. Giles, “Indexing and retrieval of scientific literature,” Proc. of the 8th Conference on Information and Knowledge Management, ACM Press, pp. 139-146, 1999.

[2] L. Bornmann and H. D. Daniel, “What do citation counts measure:A review of studies on citing behavior,” Journal of Documentation, Vol.64, No.1, pp.45-80, 2008.

[3] S. Teufel , A. Siddharthan and D. Tidhar, “Automatic classification of citation function,” Proc. of EMNLP 2006, pp.103-110, 2006.

[4] G. Nenadic, I. Spasic and S. Ananiadou, “Automatic acronym acquisition and term variation management within domain-specific texts,” Proc. of LREC-3. Las Palmas, Spain, pp. 2155-2162, 2002.

[5] Wiki in library and information Science (http://morris.lis.ntu.edu.tw/wikimedia/index.php/)

[6] M. Liu, “The complexities of citation practice: a review of citation studies,” Journal of Documentation, Vol.49, No.4, pp.370-408, 1993.

[7] D. E. Chubin and S. D. Moitra, “Content analysis of references: adjunct or alternative to citation counting?,” Social Studies of Science, Vol.5, pp.423-441, 1975.

[8] C. Oppenheim and S. P. Renn, “Highly cited old papers and reasons why they continue to be cited,” Journal of the American Society for Information Science, Vol.29, pp.225-231, 1978.

[9] S. Maricic, J. Spaventi, L. Pavicic and G. Pifat-mrzljak, “Citation context versus the frequency counts of citation histories,” Journal of the American Society for Information Science, Vol.49, pp.530-540, 1998.

[10] P. A. Hooten, “Frequency and functional use of cited documents in information science,” Journal of the American Society for Information Science, Vol.42, pp.397-404, 1991.

[11] S. Hanney, I. Frame, J. Grant, P. Green and M. Buxton, “From bench to bedside: tracing the payback forwards from basic or early clinical research – a preliminary exercise and proposals for a future study,” HERG Research Report No. 31. Uxbridge, UK, Health Economics Research Group, Brunel University, 2003.

[12] A. Ritchie. “Citation context analysis for information retrieval,” PhD thesis, University of Cambridge, 2008.

[13] E. Garfield, “Citation indexing: its theory and application in science,” Technology and Humanities, Wiley, 1979.

[14] T. Strohman, W. B. Croft and D. Jensen, “Recommending citations for academic papers,” Proc. of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp.705-706, 2007.

[15] E. Meij and M. D. Rijke, “Using prior information derived from citations in literature search,” Proc. of the International Conference on Recherched Information Assistée par Ordinateur (RIAO), 2007.

[16] K. Taghva and J. Gilbreth, “Recognizing acronyms and their definitions,” Technical report 95-03, ISRI, UNLV, June 1995.

[17] L.S. Larkey, P. Ogilvie, M.A. Price, and B. Tamilio, “Acrophile: an automated acronym extractor and server,” Proc. of 5th ACM Conference on Digital Libraries, San Antonio, TX, pp205–214, June 2000.

[18] D. Nadeau and P. D. Turney, “A Supervised Learning Approach to Acronym Identification,” in 18th Conference of the Canadian Society , for Computational Studies of Intelligence, Victoria, BC, Canada, pp319–329, 2005.

[19] D. Dann´ells, “ Acronym classification using feature combinations,“ 2007.

[20] D. Dann´ells, “Automatic acronym recognition,” Proc. of the 11th conference on European chapter of the Association for Computational Linguistics, pp.167–170, 2006.

[21] S. N. Sanchez, E. Triantaphyllou, and D. Kraft, “A featuremining based approach for the classification of text documents into disjoint classes,” Inf. Process. Manage. Vol.38, No.4, pp.583–604, 2002.

[22] G. Salton, A. Wong, C. S. Yang. “A Vector Space Model for Automatic Indexing,” Communications of the ACM, Vol.18, No.11, pp.913–620, 1975.

[23] B. Aljaber, N. Stokes, J. Bailey and J. Pei, “Document Clustering of Scientific Texts Using Citation Contexts,” Information Retrieval, Vol.13, No.2, pp101–131,2009.

[24] H. Small and E. Sweeney, “Clustering The Science Citation Index @ Using Co-Citations,” Scienrometrics, Vol.7, No.3-6, pp.391–409, June 1984.

[25] PDFTextOnline. (http://pdftextonline.com/q/)

[26] S. Yeates, “Automatic extraction of acronyms from text,” Proc. Of the 4th New Zealand Computer Science Research Students’ Conference. , pp.117–24, 1999.

[27] D. L. Olson and D. Delen, ”Advanced Data Mining Techniques,” Berlin Heidelberg : Springer-Verlag, pp.138, 2008.

[28] D. Kaplan and T. Tokunaga, ”A Citation-based Approach to Automatic Paper Summarization,” 言語処理学会第15回年次大会発表論文集, pp.128–131, 2009.

[29] A. Ritchie, S. Teufel and S. Robertson, “How to Find Better Index Terms Through Citations,” Proc. of the Workshop on How Can Computational Linguistics Improve Information Retrieval, pp.25–32, July 2006.

[30] K. Lang, “Newsweeder: learning to filter net news,” Proc. of ICML-95, 12th International Conference on Machine Learning(Lake Tahoe,CA,1994),pp.331–339, 1995.

[31] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” Proc. of 5th Berkeley Symposium on Mathematical Statistic s and Probability, 1967.

[32] C. D. Manning, P. Raghavan and H. Sch¨utze. “Introduction to Information Retrieval,” Cambridge, England: Cambridge University Press, July 2008.

附錄A　各測試資料組通過系統篩選的文件列表

TestData.1

1. H. Houissa and N. Boujemaa, “Coherence criterion for region labelling and description,” 2005.
2. B. L. Saux and N. Boujemaa, “Image Database Clustering with SVM-based Class Personalization,” SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia, 2004.
3. H. Houissa and N. Boujemaa, “Region labelling using a Point-Based Coherence Criterion,” Proc. of SPIE, 2006.
4. B. Saux and N. Boujemaa, “Unsupervised categorization for image database overview,” Proc. Recent Advances in Visual Information Systems,2002. Proc. Recent Advances in Visual Information Systems / International Conference on Visual Information System, Vol. 23,14 of Lecture Notes in Computer Science, pp.163-174, March 2002.
5. B. L. Saux and N. Boujemaa, “Unsupervised robust clustering for image database categorization,” Proc. of the International IEEE Conference on Pattern Recognition (ICPR), 2002.
6. H. Frigui and R. Krishnapuram, “A Robust Competitive Clustering Algorithm withApplications in Computer Vision,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.21, No.1, pp.450-465, January 1999.
7. L. Ren, G. W. Irwin, “Robust fuzzy Gustafson-Kessel clustering for nonlinear system identification,” International Journal of Systems Science, Vol.34, No.1, pp.787-803, 2003.
8. H. Frigui, C. Hwang and F.C.-H. Rhee, “Clustering and aggregation of relational data with application to image database categorization,” Pattern Recognition, Vol.40, pp.3053-3068, 2007.
9. O. Nasraoui, H. Frigui, R. Krishnapuram and A. Joshi, “Extracting web user profiles using relational competitive fuzzy clustering,” International Journal on Artificial Intelligence Tools, Vol.9, pp.509-526, 2000.
10. O. Nasraoui, H. Frigui, A. Joshi and R. Krishnapuram, “Mining Web Access Logs Using Relational Competitive Fuzzy Clustering,” Proc. of the Eight International Fuzzy Systems Association World Congress, August 1999.
11. N. Grira, M. Crucianu, N. Boujemaa, “Fuzzy Clustering with Pairwise Constraints for Knowledge-Driven Image Categorization,” IEEE Proc. on Vision, Image and Signal Processing, 2006.
12. N. Grira, M. Crucianu and N. Boujemaa, “Semi-Supervised Fuzzy Clustering with Pairwise-Constrained Competitive Agglomeration,” IEEE International Conference on Fuzzy Systems (Fuzz'IEEE 2005), May 2005.
13. N. Grira, M. Crucianu and N. Boujemaa, “Semi-supervised image database categorization using pairwise constraints,” IEEE International Conference on Image Processing. Genova, Italy, 2005.
14. Y. Kanzawa, Y. Endo and S. Miyamoto, “Some Pairwise Constrained Semi-Supervised Fuzzy c-Means Clustering Algorithms,” MDAI 2009, LNAI 5861, pp.268-281, 2009.

15. A. Chourou and A. Benazza-Benyahia, “Lossless and Gradual Coding of Hyperspectral Images by Lifting Scheme,” Proc. of 15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 2007.
16. R. Tebourbi, A. Benazza-Benyahia and Z. Belhadj, “Multiscale Retrieval with partial query of multispectral satellite images”.

TestData.2

17. Z. Xie, A. Wang, F. L. Chung, “An enhanced possibilistic C-Means clustering algorithm EPCM,” Soft Computing, Vol.12, pp.593-611 ,2008.
18. N. R. Pal, K. Pal, J. C. Bezdek, “A mixed c-means clustering model,” IEEE Proc. of the International Conference on Fuzzy Systems, pp.11-20, 1997.
19. N. R. Pal, K. Pal, J. M. Keller and J. C. Bezdek, “A possibilistic fuzzy c-means clustering algorithm,” IEEE Trans. Fuzzy Systems, Vol.13, No.4, pp.517-530, August 2005.
20. J. S. Lin, “Fuzzy possibilistic neural network to vector quantizer in frequency domains,” Optical Engineering, pp.839-847, April 2002.
21. S. H. Liu and J. S. Lin, ”Vector quantization in DCT domain using fuzzy possibilistic c-means based on penalized and compensated constraints,” Pattern Recognition, Vol.35, No.10, pp.2201-2211, 2002.
22. F. Masulli, A. Schenone, “A fuzzy clustering based segmentation system as support to diagnosis in medical imaging,” Artificial Intelligence in Medicine, Vol.16, No.2, pp.129-147, 1999.
23. D. Q. Zhang and S. C. Chen, “Kernel based fuzzy and possibilistic c-means clustering,” Proc. International Conference on Artificial Neural Network (ICANN '03), Istanbul, Turkey, pp.122-125, June 2003.
24. M. De C´aceres, F. Oliva and X. Font, “On relational possibilistic clustering,” Pattern Recognition, Vol.39, No.11, pp.2010-2024, 2006.
25. V. S.Tseng and C. Kao, “A novel similarity-based fuzzy clustering algorithm by integrating PCM and mountain method,” IEEE Trans. on Fuzzy Systems, Vol.15, pp.1188-1196. December 2007.
26. J. S. Zhang , Y. W. Leung, “Improved possibilistic c-means clustering algorithms,” IEEE Trans. on Fuzzy Systems, Vol.12, pp.209-217, 2004.
27. R. Krishnapuram and J. M. Keller, “The possibilistic c-means algorithm: insights and recommendations,” IEEE Trans. Fuzzy Systems, Vol.4, pp.385-393, 1996.

TestData.3

28. L. Galluccio, O. Michel, P. Comon, A. O. Hero and M. Kliger, “Combining multiple partitions created with a graph-based construction for data clustering,” IEEE International Workshop on Machine Learning for Signal Processing, September 2009.
29. D. D. Abdala and X. Jiang, ”An Evidence Accumulation Approach to Constrained Clustering Combination,” MLDM 2009, LNAI 5632, pp.361-371, 2009.
30. J Duarte, A Fred, F Rodrigues, J Duarte, S Ramos and Z. Vale, “Definition of MV load diagrams via weighted evidence accumulation clustering using subsampling,” Proc. of the 6th WSEAS International

Conference on Signal Processing, Robotics and Automation, Corfu Island, Greece, February 2007.

31. P. Casas, J. Mazel, P. Owezarski and Y. Labit, “Sub-Space Clustering and Evidence Accumulation for Unsupervised Anomaly Detection in IP Networks,” Hal-00485427, Version 1-20, May 2010.
32. W. Hasperue, L. Lanzarini , “Classification Rules Obtained from Evidence Accumulation,” On Information Technology Interfaces, ITI 2007, 2007.
33. S. S. Khan, Dr. S. Kant, “Computation of initial modes for K-modes clustering algorithm using evidence accumulation,” Proc. of the 20th International Joint Conference on Artificial Intelligence, Vol.7, pp.2784-2789, 2007.

附錄 B　各測試資料組共引用文件標題列表

TestData.1

1. R. N. Dav´e, “Characterization and detection of noise in clustering,” Pattern Recognition Letters, Vol.12, No.11, pp. 657–664, 1991.
2. J. Z. Wang, J. Li and G. Wiederhold. “SIMPLIcity: Semantics-sensitive integrated matching for picture Libraries,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.23, pp.947-963, 2001.
3. R. J. Hathaway, J. W. Davenport and J. C. Bezdek, “Relational duals of the c-means algorithms,” Pattern Recognition, Vol.22, pp.205-212, 1989.
4. R. J. Hathaway and J. C. Bezdek, “NERF c-Means: Non-Euclidean relational fuzzy clustering,” Pattern Recognition, Vol.27, pp.429-437,1994.
5. H. Frigui and O. Nasraoui, “Unsupervised Learning of Prototypes And Attributes Weights,” Pattern Recognition, 2004.
6. R. R. Yager, “On ordered weighted averaging aggregation operators in multicriteria decision making,” IEEE Trans. on Systems, Man and Cybernetics, Vol.18, pp.183-190, 1988.
7. A. Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of Royal Statistical Society, Vol.39, pp.1-38, 1977.
8. N. Boujemaa, “On competitive unsupervised clustering,” Proc. of International Conference on Pattern Recognition, Spain, 2000.
9. B. Le Saux and N. Boujemaa, “Unsupervised robust clustering for image database categorization,” Proc. International Conference on Pattern Recognition, 2002.
10. S. Basu, A. Banerjee and R. J. Mooney, “Semi-supervised clustering by seeding,” Proc. Of 19th International Conference on Machine Learning (ICML’02), pp.19-26, 2002.
11. N. Boujemaa, “On competitive unsupervized clustering,” Proc. of the International Conference on Pattern Recognition, Vol.1, pp.631-634, Barcelona, Spain, September 2000.
12. S. A. Nene, S. K. Nayar, and H. Murase, “Columbia object image library (coil-100),” tech. rep., Department of Computer Science, Columbia University, http://www.cs.columbia.edu/CAVE/, 1996.

TestData.2

13. O. Nasraoui and R. Krishnapuram, “Crisp Interpretations of Fuzzy and Possibilistic Clustering Algorithm,” Proc. EUFIT, Aachen, Germany, pp.1312-1318, 1995.
14. E. E. Gustafson and W. Kessel, “Fuzzy Clustering with a Fuzzy Covariance Matrix,” Proc. 1978 IEEE CDC, pp.761-766,1979.
15. R. N. Dave and S. Sen, “Possibilistic c-means clustering for relational data,” IEEE Trans Fuzzy System, Vol.10, No.6, pp.713-727, 2002.
16. H. Timm, C. Borgelt, C. Döring, R. Kruse, “An extension to possibilistic fuzzy cluster analysis,” IEEE Trans Fuzzy Sets and Systems, Vol.147, pp.3-16, 2004.
17. R.N Dave, R. Krishnapuram, “Robust clustering methods: A unified view,” IEEE Trans. Fuzzy System Vol.5, pp.270–293, 1997.

TestData.3

18. A. Topchy, A. K. Jain and W. Punch, “Clustering ensembles: Models of consensus and weak partitions,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.27, No.12, pp.1866-1881, December 2005.
19. A. Fred and A. K. Jain, “Data Clustering Using Evidence Accumulation,” Proc. 16th International Conference Pattern Recognition, pp.276-280, 2002.
20. S. Theodoridis and K. Koutroumbas, “Pattern Recognition,” Academic Press, third edition, 2006.