國立交通大學

資訊管理研究所

碩士論文

運用時間序列分群於社會性標籤之研究

A Study of Applying Time Series Clustering to Social Tagging

研究生：曾姿婷

指導教授：柯皓仁　博士

中華民國　九十八　年　六　月

運用時間序列分群於社會性標籤之研究

A Study of Applying Time Series Clustering to Social Tagging

研 究 生：曾 姿 婷　　Student: Tzu-Ting Tseng

指導教授：柯 皓 仁　　Advisor: Dr. Hao-Ren Ke

國 立 交 通 大 學

資 訊 管 理 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

In

Information Management

June 2009

Hsinchu, Taiwan, the Republic of China

中華民國　九十八 年 六 月

運用時間序列分群於社會性標籤之研究

指導教授：柯皓仁　　研究生：曾姿婷

國立交通大學資訊管理研究所

摘要

由於網際網路的普及化，使得網路的服務與經營型態層出不窮，在Web 2.0興起後，分享與參與式的架構更成為網路服務的主流。社會性標籤是Web 2.0的一項重要服務，社會性標籤讓網路使用者對網路資源進行標記，實現了運用集體力量收藏及分享網路資源的機制。社會性標籤之所以能夠蔚為風行，原因是其背後社交的性質，讓社會大眾透過標籤產生對話與互動。

本研究以社會性標籤的社會層面為出發點，檢視社會性標籤隨時間的變化趨勢，以了解社會脈動。本研究利用時間序列分群演算法，首先收集黑米共享書籤網站裡的標籤根據其所標記之網頁內容轉換為時間序列的形式，找出在同一時間區間擁有相似走勢的標籤群聚，形成主題概念；接著計算不同時間區間所形成之群聚的相似度，以萃取出所有時間區間中擁有相似主題概念的群聚和包含於內的標籤；此外，對於同一標籤，分析其在各時間區間的變化趨勢，以及相關連的標籤和網頁。最後，透過本研究所開發之雛型介面將前述研究成果整合。

關鍵詞：Web 2.0、社會性標籤、時間序列分群法、黑米共享書籤

A Study of Applying Time Series Clustering to Social Tagging

Advisor: Dr. Hao-Ren Ke Student: Tzu-Ting Tseng

Institute of Information Management

National Chiao Tung University

Abstract

Due to the widespread adoption of Internet, a variety of Internet applications emerges. With the booming of Web 2.0, participation and sharing becomes the central concept of Web services. Social tagging is one essential service in the Web 2.0 era, which allows users to label Web resources with keywords thought of by themselves. Social tagging also enables the cooperative acquisition and sharing of Web resources. The key factor that social tagging can catch people's attention is the "social" character, which facilitates the communication and interaction of people.

This study, based on the social character of social tagging, detects the chronological variation of social tagging for understanding social trends. Time series clustering for tags is employed in this study. First, tags collected from HEMiDEMi are transformed into the form of time series according to the Web pages labeled by these tags, and then time series clustering is used to identify the tag clusters with similar chronological patterns and trends in the same time period. Next, the similarity between clusters in all time period is calculated to extract the clusters and associated tags with similar concepts in different time periods. Furthermore, the trend variation of an identical tag in different time period is analyzed, and the related tags and Web pages are discovered. Finally, the above research outcomes are integrated into a prototype system.

Keyword：Web 2.0, Social Tagging, Time Series Clustering, HEMiDEMi

誌 謝

隨著口試的結束，研究所生活也要在此畫下句點。回想這兩年來的生活，雖然不是什麼卡司堅強的動作片，但仔細閱讀卻發現是一部令人動容的精緻小品。除了滿滿的感謝還有不捨。

我的論文得以產出，我要先感謝自己也忙得焦頭爛額的鈞智，謝謝鈞智和我一起討論思考出論文研究方向；再來我要大大地感謝當時遠在美國的信源學長，謝謝學長願意花時間一步一步的和我用Skype討論論文、幫我解決問題，讓我每天打開Gmail是既期待又怕受傷害，深怕哪裡又做錯了，哪裡又要多做些什麼，但因為這樣的督促下，讓我深刻體會到研究真不是件馬虎的事，自己不足的地方還很多很多。接著是感謝台積電，感謝台積電在今年年初的時候放了幾個月的無薪假，讓剛進台積電放了無薪假卻不知道要去哪的怡祥學長，可以每週回到研究室，緊盯我的進度，每個星期問的就是：程式寫了沒？資料跑了沒？介面做了沒？論文寫了沒？這口語式的轟炸，無形之中成了我向前跑的動力。最後要由衷感謝的是指導教授柯皓仁老師，謝謝老師兩年來的教導，和老師相處就像如沐春風。老師其實是個大忙人，但每次個別開會總是會花上一個小時和我討論論文細節，讓我感到很不好意思。到最後修改論文階段，還不厭其煩的一直修改我的論文和排版，再加上精益求精的精神，使得我的論文得以輸出。

在我的研究生活中要感謝陪伴我兩年的研究所同學：感謝我的飯友，盈羽和力廣，讓每天在研究室的生活，在吃飯時總是能有個聊天對象；感謝另外三位千金，欣穎、盈羽、盈佑，大家總是會彼此鼓勵、彼此說著別人的八卦，排憂解悶促進生活的樂趣。感謝研究室兩位貼心的學妹，筑婷、雅雯，總是會聽我不停的吐苦水，幫我思考論文的盲點在哪，文句哪不通順。感謝淑惠總是那樣的親切對待我們每一個人，讓所辦就像是我們研究做累時串們子的好地方。也謝謝淑惠總是不厭其煩的幫我們張羅每學期所上的活動，讓我們的研究所生活多增添了一些

風采。此外我還要感謝我的大學同學小萩和一哥，謝謝小萩總是帶給我正面的思考價值；謝謝一哥總是在我回家的路上可以撥通電話給你，讓我呼吸不一樣的空氣。

最後的最後，要謝謝我的家人，謝謝爸爸媽媽讓我可以不用擔心任何生活問題繼續的念書；謝謝爸爸媽媽總是讓我做自己想做的事情。謝謝你們的包容，才有今天的我。

曾姿婷 謹誌於交通大學資管所

2009年7月

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究背景與動機

隨著網際網路的普及化，民眾平均上網時間年年升高。根據資策會FIND統計，截至2008/12/31年底，國人經常上網人口為1,046萬人，網際網路連網應用普及率為45% [28]；時報周刊也統計，台灣人一週的平均上網時數為12.6小時[29]，占日常生活各項作息時間之比重愈來愈大。由此可知網路已經變成生活的一部分。

網際網路的使用型態與經營模式，自2000年網路泡沫化後，邁向了另一個新階段，在2004年O'Reilly提出Web 2.0的概念後[15]，更朝著使用者導向的服務模式發展。想要在網路上打響名氣，沒有集合大眾智慧、讓使用者有展現自我的平台，是很難成功的。舉凡自網路泡沫化存活下來的Amazon、eBay，及現在變成人們生活一部分的Wikipedia、flickr、del.icio.us、facebook、無名小站、YouTube、HEMiDEMi…等，這些網站都是以人為基礎，且由大量使用者摒棄利己小我的精神，各自提供少量的資訊，最後形成利他大我龐大的資料庫。這種「分享式」的行為模型與「參與式」的軟體架構已成為Web 2.0的核心價值[15]。

在Web 2.0各種應用中，從del.cio.us的網路書籤共享推出後，再加上YouTube影音分享，還有flickr的圖片共享…等網站的鼓舞下，使用者能夠恣意地將他們喜愛的網站、影音、照片、甚至學術性質的文章…等，用自己的方式進行分類、標記關鍵字，這些標記的關鍵字通常稱為標籤(Tag)，使用者可以分享彼此的標籤，甚至再把標註相同標籤的物件集合起來成為一個新的組織分類。因此，可以說這些社會性標籤(Social Tagging)是Web 2.0網站的重要辨別特徵[24]。

社會性標籤是一種運用集體力量收藏和分享標籤的機制，就像長尾理論

中，尾巴「聚少成多」的力量得以發揮最大的效益，眾多小人物的智慧是不可被低估和忽略的。社會性標籤也是一種不同於傳統由內容作者自行下關鍵字，或是由分類學專家來對內容加以分類的方式。

如此讓社會性標籤蔚為盛行的原因，其實不是標籤本身，而是其背後社交(Social)的特質，是一種促進社會對話的標記行為。具體而言，即是希望一群人，不論主動或被動、刻意或自然，透過標籤來產生對話或互動。而社交特質最重要的目的即是產生有意義的關聯，以便使用者搜尋及發掘資訊，例如找到志同道合的人就是一種頗有價值的關聯[23]。社會性標籤因為將分類的權力下放給每位網友，每位網友可以天馬行空對文章、圖片、影音等進行標記，從這些標籤可以學習到許多趨勢上的變化及社會、文化現象上的脈動，例如在不同的時空環境下，使用者使用「很囧」這標籤時，所表示的事件不一定相同；在同樣的時空環境中，亦能透過標籤找出具有類似概念的網路資源，同時得知來自四面八方的使用者對這些資源的看法描述。網路使用者如何多樣化地描述各式網路內容，就是讓社會性標籤的價值得以彰顯的原因。

本研究以社會性標籤的社會性特質為出發點，藉著時間序列分群(Time Series Clustering)方式，更精細地檢視社會性標籤隨時間變化的趨勢。就以下例子作說明：

假設「奧運」、「中國」、「北京」、「政治」、「台灣」這五個標籤出現在 p_i、p_j、p_k、p_l、p_m 等五個時間點的次數為表 1。若以整個時間區間來看，傳統不考慮時間序列的分群法會將中國和奧運分在同一群，因為其字詞出現總頻率較接近。但換用時間點檢視(如圖 1)，會發現和中國、政治、台灣等三個標籤的時間序列曲線最為相似，因此，中國、政治、台灣所表示的事件其關連度應該比中國、奧運來得緊密。

表 1 標籤次數表

	p_i	p_j	p_k	p_l	p_m	總數
奧運	40	20	0	2	0	62
中國	15	15	15	15	15	75
北京	10	11	0	6	8	35
政治	5	10	20	10	8	53
台灣	6	9	19	8	10	52

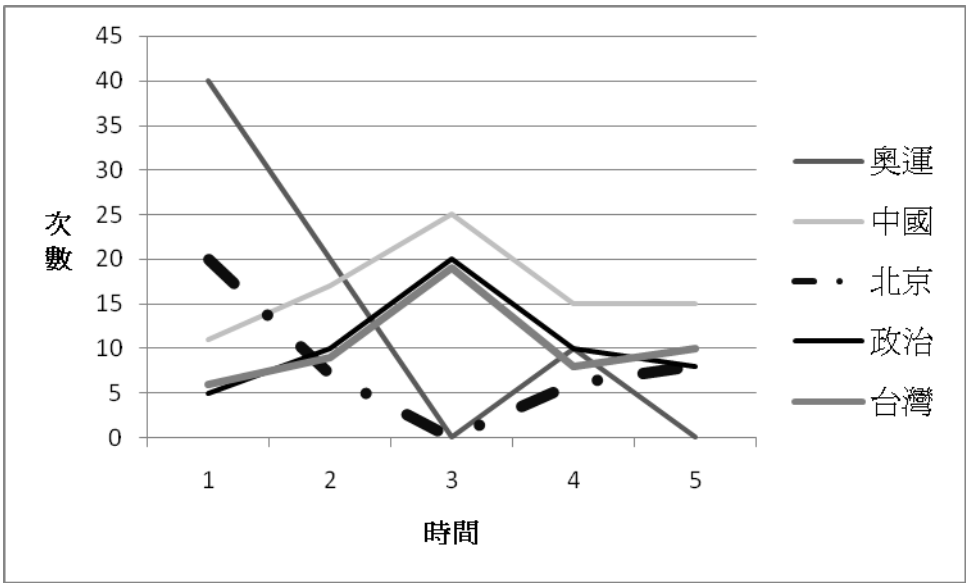


圖 1 標籤時間序列圖

1.2 研究目的

本研究之目的在於將時間序列分群運用於社會性標籤上，並希望藉此挖掘出不同的主題概念，再檢視不同時間區間社會性標籤的變化趨勢，瞭解社會現象的脈動。透過檢視社會性標籤在每個時間點出現情形的變化，社會性標籤的時間序列分群以標籤的時間序列走勢做為判斷不同標籤相似與否的依據。

本研究利用時間序列的分群演算法，處理網路使用者的社會性標籤和被標記的網頁內容，擷取出具有時間概念的主題，進而推薦給使用者。整體的研究方向如下：

一、 透過時間序列化的標籤，讓使用者瞭解在特定時間區間擁有相同走勢的標籤組合。

二、 針對單一標籤的搜尋結果，觀察與此標籤相關之標籤在不同時間區間的變化情形。

三、 計算不同時間區間的群聚相似度，推薦其他時間區間擁有相似主題概念的標籤給使用者。

1.3 論文架構

本論文在第二章將進行社會性標籤、時間序列分析、分群方式等三大主題的文獻回顧。第三章則詳細描述本研究如何進行文章前置處理作業，進而產生時間序列；如何運用產生的時間序列進行分群演算法，形成最後的分群結果推薦給使用者。第四章介紹本研究所開發之雛形系統，並透過質化與量化分析方式比較有無使用時間序列分群結果的差異。第五章總結本研究，並說明未來發展方向。論文整體架構如圖 2所示。

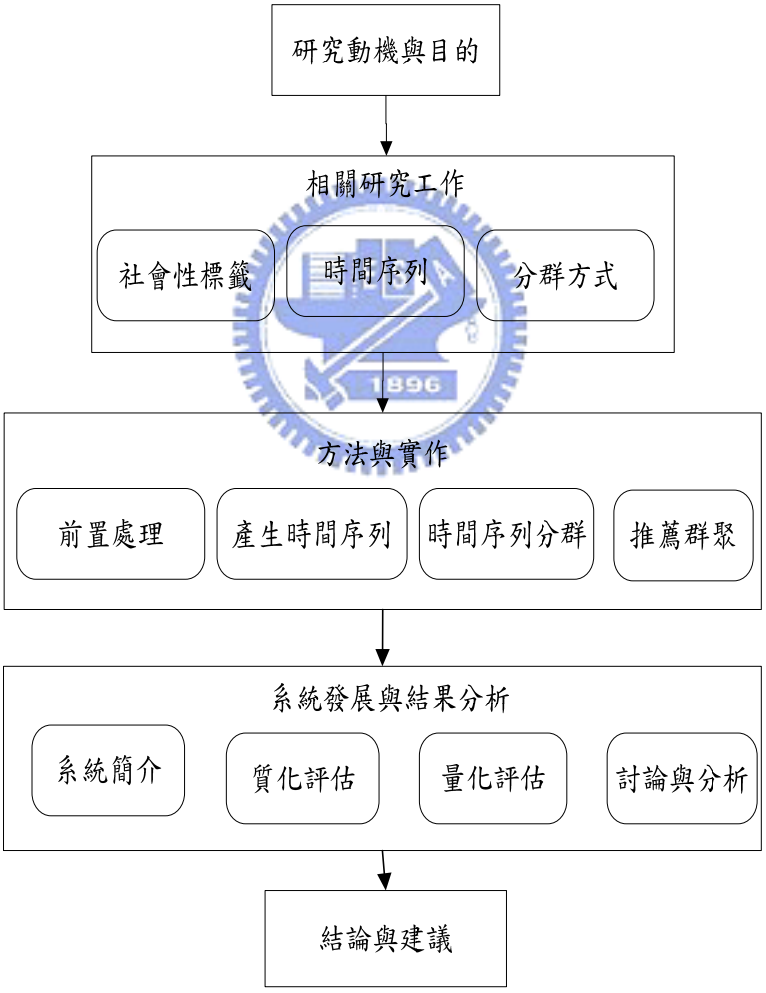


圖 2 論文整體架構

第二章　相關研究工作

本研究將時間序列分群演算法應用在社會性標籤上，本章收集與整理相關的研究工作，分別在2.1節探討社會性標籤的相關文獻，包含社會性標記目前的應用及標籤的未來趨勢。2.2節闡述何謂時間序列，以及時間序列的分析方法。2.3節介紹常用的分群演算法和評估方式。

2.1 社會性標籤

社會性標記(Social tagging)有很多別名，諸如：Sharing Tagging、Collaborative Tagging、Folksonomy、Social Classification或Social Indexing…等。簡言之，它就是每個人都可以為網路上的各式內容，如：網頁連結、音樂、照片、影片、文章書目等，加上自己的關鍵字(即標籤)。

2.1.1　社會性標籤的起源

說到社會性標籤就得先提到社會性書籤(Social Bookmarking)。社會性書籤是一種網路使用者可將自己喜愛的網頁超連結儲存、組織和管理的方式。有別於傳統分類的詞彙進行分類，多數的社會性書籤網站鼓勵使用者使用較不正式的「標籤」來分類，這標籤就稱為「社會性標籤」。將這些共同創作所形成的標籤分類系統，稱為大眾分類法「Folksonomy」，將會在下一小節作說明。這些被收集的網頁超連結因可以被許多人在網路上分享，因此又可以稱為「網路書籤」或「線上書籤」。它有點類似瀏覽器中的「我的最愛」，但不同於「我的最愛」之處在於：使用者可以隨時透過網路看到自己的收藏，也能透過搜尋和標籤分類等功能，快速地找到自己的收藏或別人分享的書籤。透過看到他人分享的書籤，使用者也能夠輕易地知道目前較為熱門或受好評的新聞與網頁文章。

具有分享性質的社會性書籤，早自1996年便已出現，itList，當時使用者就可以自行決定他收藏的書籤要公開或私有。莫約三年後，踏入線上書籤服務的競

爭者越來越多，例如：Backflip, Blink, Clip2, ClickMarks, HotLinks…等。然而缺乏有利的商業模式為基礎，在2000年爆發網路泡沫化後，這些早期社會性書籤的先驅者慢慢地淡出這舞臺。

2003年左右，隨著Blog的興起和個人化網路的發展，使用者從網路獲取資訊的習慣和方法也產生了很大的變化，不再只侷限於單向的資訊給予，雙方分享式的資訊互通交流，已漸漸成為主流。最早的社會性書籤網站del.icio.us也在此時推出。其簡潔的風格和各種互動分享式的功能立刻吸引了大量的使用者，再加上網友們的口耳相傳引爆了「del.icio.us」的流行。在台灣，類似的服務則有co 嘿米共享書籤HEMiDEMi、MyShare、和推推王funP…等。

2.1.2 大眾分類法

大眾分類法(Folksonomy)這個概念是由 Vander Wa l(2005)在討論 Flickr 和 Del.icio.us 所發展的資訊架構時，將「Folks」和「Taxonomy」組合而成的新詞彙[16]。照字面上的意思來看「folk」是指一群人、老百姓，屬於較口語化的方式，類似社會大眾；「Taxonomy」則是傳統的分類法；結合起來就是社會大眾、平民老百姓在進行分類。既然是大眾在進行分類，那分類的方式和使用狀況一定和傳統的方類方式大相逕庭。如[7]所說；「這種分類是平面化的(flat namespace)，沒有等級層次的劃分。」當然這一定有它的優缺點。根據[24]裡整理大眾分類法的優缺點，如表 2：

表 2 大眾分類法優缺點

優點	缺點
定義標籤者即為內容使用者,使用者認同感較高。	錯別字太多。
具回饋性,可幫助社群創造溝通與分享空間。	單詞索引(Single Word Indexing)：是googlemaps還是google-maps
具集體智慧：如透過字彙與概念的變化，可呈現流行主題。	無分類架構，較無脈絡可循。
具語言及文化的豐富性：例如同一網站，各國人士所訂定關鍵字，可以反映不同文化觀點	無標記指導原則:標籤格式及給定原則缺乏標準，不易達成一致性
具啟發性：可協助使用者發現、探究，以尋得原先未知的資源。	缺乏控制字彙,降低資料的有用性與接受度

雖然它相對的不夠嚴謹，缺乏準確度，但是在社會性軟體中，這種平面延伸的分類方法卻在無形之中成為形成了溝通的渠道和網路。這些社會性標籤包含的範圍很廣泛，但實際上還是可以大致歸類為以下幾種[19]

I. 以內容為主(content-based tags)：這類標籤以書籤的內容為主，或書籤主題的類別。通常會有較多精確的詞彙。例如：馬英九、奧運、消費券等。

II. 以文章脈絡、上下文為主(context-based tags)：這類標籤較傾向描述整體文章，例如：是描述一個觀光景點、或觀光勝地。

III. 以屬性為主(attribute tags)：這類標籤指出標記的書籤的類型，例如：是部落格文章還是一篇新聞。或者根據這書籤內容的作者、擁有者,例如：Mr.six’s blog。

IV. 主觀性的(subjective tags)：這類的標籤包含較多的個人情感和主觀意識在裡面，以形容詞為主，例如：悶、funny、stupid等。

V. 組織性的(organizational tags)：這類標籤較屬於「任務」型的，用來提醒自己待辦事項，例如：to-read，job-search；還有一種是拿來自我參考用的，例如：my stuff、 my work等。

具備Folksonomy功能的網站應用相當廣泛，舉凡有Web 2.0概念的網站，幾乎都有社會性標籤的影子。這類網站多以提供一般資訊為主；網站亦提供RSS訂閱服務，讓使用者隨時可以掌握最新、熱門書籤；大多也都有提供API外掛功能，開發小工具讓使用更方便收藏書籤。如圖 3 社會性標籤開發小工具的黑米共享書籤網站，使用者在瀏覽網頁時，想收藏該網頁時，只要在瀏覽器右上方點選「+黑米書籤」，就可以直接收藏該網頁到使用者所使用的社會性書籤網站；圖 3 社會性標籤開發小工具下方的加入書籤小圖示則常見於部落客的部落格，部落客除了喜愛將自己的文章發表到網路上外，促使他們有源源不絕發表的動力，更是網友們的推薦和鼓勵。當他們發現他們發表的文章有網友將之儲存於類似「我的最愛」的網路書籤上時，是一種滿足自我的榮譽感，讓他們持續有發表的動力。這也符合了Web 2.0的最初的精神：雙向、互動、分享。

圖 3 社會性標籤開發小工具

2.1.3 社會性標籤未來趨勢

標籤自2003年開始流行，在2007年有些學者發現標籤似乎停滯不前了，甚至認為標籤在未來已經沒有任何競爭優勢。然而，美國資訊科學與技術學會會報

(ASIS&T Bulletin, August/September 2008)上，Gene Smith發表了Tagging: Emerging Trends(標籤分類：新興的趨勢)一文[11]，文中提到標籤的未來有以下四個趨勢：

I. 更結構化

早期的標記系統相當具有自由的思想和擴充度，近年來一些創新的標記系統結合了更結構化的構想，但卻又不失標記原有的開放性和社交性特色，使標籤更有價值。例如：Zigtag這網站，它將標籤結合概念，當使用者下「apple」時，它可以分辨是指水果還是蘋果電腦，如圖 4所示。

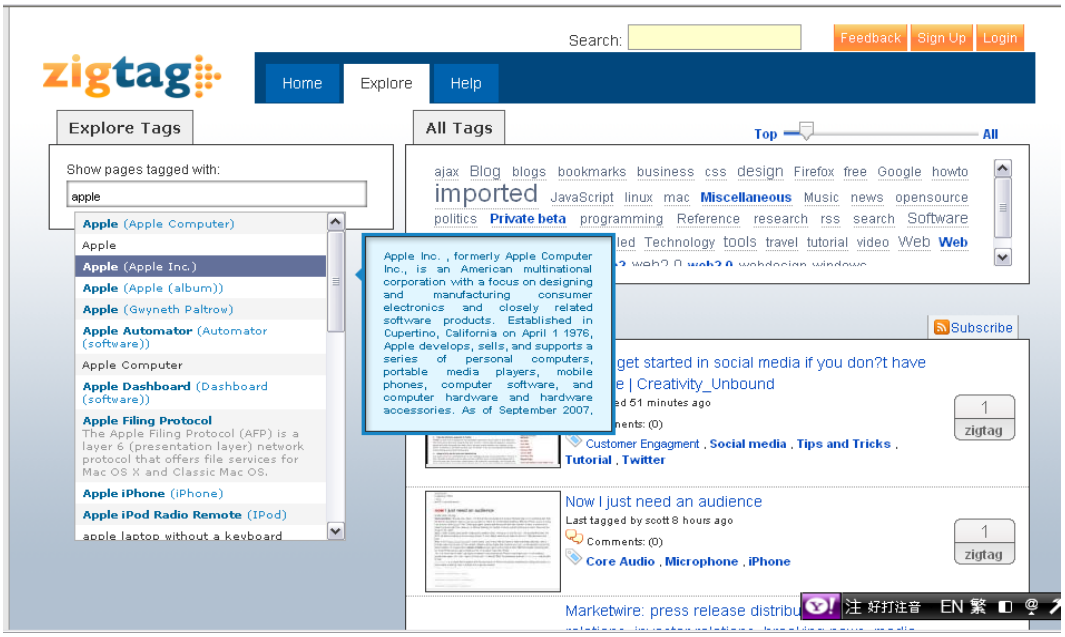


圖 4 Zig標籤網站

(資料來源：http://www.zigtag.com/)

II. 運用社群之力

有些系統利用使用者去幫助減少雜訊或者刪除一些無意義或重複的標籤。社群能夠協力改善他們所建的標籤集合(標籤 Collections)。例如：

a) 完全相同的同義詞，如WWII與World War 2(第二次世界大戰)。遇到這種情形，其中那個較常被大家使用的詞彙便會成為首選詞(preferred term)。這種同等詞彙(標籤

equivalency)的建立是由下而上、由社群驅動產生的控制詞彙。

b) 社群同時也能決定表面上同義，但實際上在社會語意方面差別很大的詞彙，如humor(美國慣用)與humour(英國慣用)。

III. 自動與人工結合的分類法

這是結合由上而下(Top-down)與由下而上(Button-up)制定出標籤一種自動大眾分類方式。有些社會性標籤系統結合演算法和人工方式，希望將傳統與新型資訊架構間的鴻溝降到最小。例如Library Thing的Tagmash，它是一種可結合多個標籤檢索出符合條件書單的搜尋方式。如圖 5，輸入了harry potter、england、fantasy，Tagmash會去對照資料庫中關於harry potter、england、fantasy的同義詞，再去對映到分類表或標題表的項目上，便可產生關於harry potter、england、fantasy的類別書單。雖然這種方式可能無法達到專業編目員這麼高的正確性及一致性，但對資訊架構師而言，如果想取標籤集合大眾智慧分類方式，但又不想完全將控制權交給使用者的話，這種結合自動與人工的標籤分類會是種好方法。

圖 5 自動與人工結合分類法例子

(資料來源：LibraryThing網站)

IV. 由使用者產生的創新

在網路服務與應用上，運用標籤可易於作混搭(Mashup)及創新。例如在Flickr，每個標籤皆有RSS feed，可供他人訂閱並供運用來做一些創新服務的實驗，Flickr上的地理標籤分類便是一種創新服務，圖 6。社會性標籤的價值不僅讓人們能與資訊互動，而且讓人們能改變他們的資訊環境以更符合自己的需求。

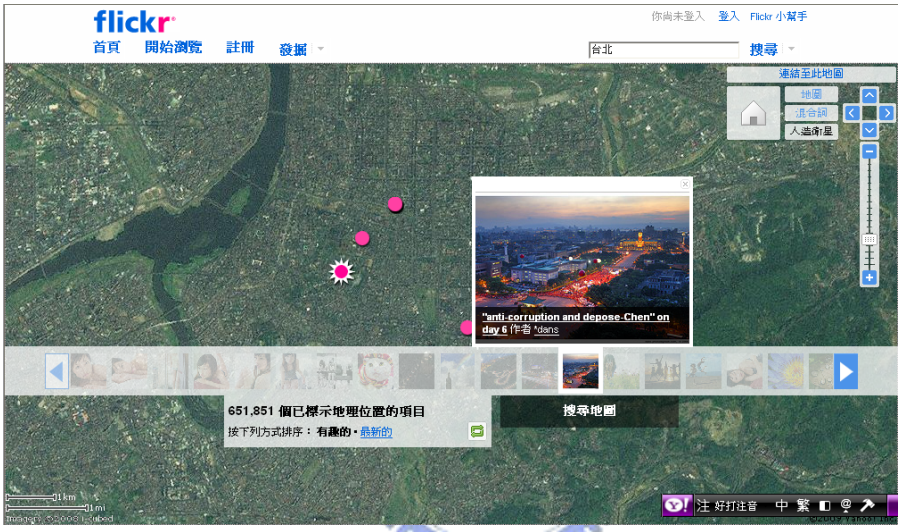


圖 6 使用者產生的創新

(資料來源：http://www.flickr.com/)

以上四個趨勢說明了社會性標籤仍繼續在演進。今天看到標籤、專家分類及多面分類等混合後產生的一種新的、有價值的資訊結構(Information Structure)。最重要的是，標籤正被用來解決資訊架構的一些古典問題(Classic Problems)－ 即幫助人們尋找及利用資訊，從語言的纏結中找出意義、降低因模糊不清所導致的認知和經濟上的成本[11] 。

2.2 時間序列分析

時間序列(Time Series)是依事件或資料發生的先後次序排列的一群統計數據[22]；時間序列資料(Time Series Data)指的是同一元素的同一特質(變數)於不同時

間點或不同期間的資料，包括逐日的日資料、週資料、月資料、季資料及年資料等。例如：2008年台北盆地的降雨量、某公司第一季股票價格的漲幅、醫院病患一週的心電圖…等；時間序列分析(Time Series Analysis)利用這些過去的歷史資料為依據，分析事情發展的前因後果，預測將來變動趨勢。常應用於：股票市場價格之漲跌、產品銷售量的成長、溫度、降雨量變化之類的資料。

如同一般分群方式，處理時間序列的分群也是需要一組分群演算法。因此要進行時間序列分群需要先分析兩兩時間序列的相似度，以下小節介紹如何計算時間序列的相似度。

2.2.1 歐幾里德距離

分析兩個時間序列之間的相似程度最簡單的方法就是使用歐幾里德距離(Euclidean distance)，此方法將長度為N的序列看成是N維的歐幾里德空間(Euclidean space)裡的一個點，而定義兩序列間的相似度為序列分別對應到空間裡之點的距離，而此法對於兩個測量目標有程度變形的關係以及偏移量的關係則無法有效地處理，因此[4]提出改進之方法：將序列正規化。

序列正規化之概念在於將每一個序列的平均值（mean）以及變異數（variance）正規化，以去除序列間不同的偏移量以及程度變形，之後再利用歐幾里德相似度量測法計算相似度。

但是此方法無法處理有關時間平移的問題。例如有兩個時間序列表現反應非常相似，但第二個時間序列的表現曲線比第一個時間序列延遲了一段時間，則直接使用距離量測或相關係數來計算相似度，不會得到此兩個時間序列表現相似的結果，圖 7表示時間序列平移的示意圖。

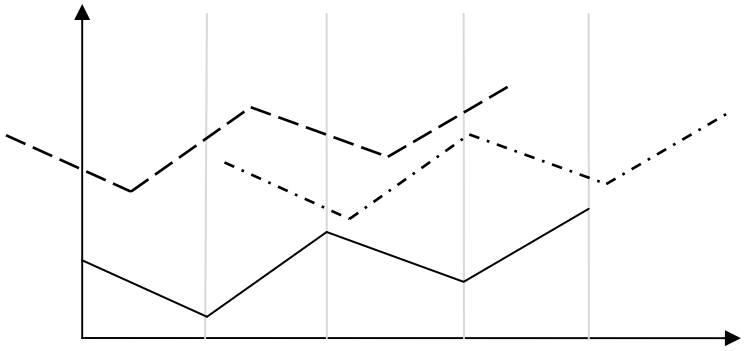


圖 7 時間序列平移

2.2.2 動態時間變形

動態時間變形(Dynamic time warping distance, DTW)的主要想法是允許序列中任一資料點重複使用以將序列延伸，試圖藉此解決時間平移的問題，如圖 8。

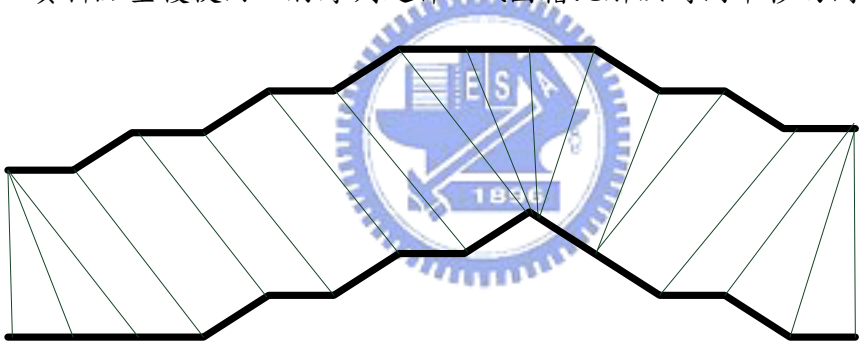


圖 8 兩個時間序列的變形

假設有兩個時間序列 $Q = q_1, q_2, ..., q_i, ..., q_n$ 和 $R = r_1, r_2, ..., r_j, ..., r_m$ ，DTW組合兩個序列使他們的距離差異最小化，為了達到這個結果，會形成 $n \times m$ 的矩陣，矩陣存放q_i和r_j的距離，常用距離計算方式為歐幾里德距離。一個變形路徑(warping path)為 $W = w_1, w_2, ..., w_K$ ， $\max(m,n) \le K \le m+n-1$ ， $w_k = (i, j)$ 。其中令人感興趣的是這兩個序列變形路徑中的最短距離，d_{DTW}如公式(1)，

$$d_{DTW} = \min \frac{\sum_{k=1}^{K} w_k}{K} \tag{1}$$

利用動態規劃的方式有效的計算兩個時間序列延伸後之距離，以找出此二序

列最佳之對齊方式。遞迴方程式如(2)[17]：

$$D(i,j)=d(q_i,r_j)+\min\begin{Bmatrix}D(i-1,j-1)\\D(i-1,j)\\D(i,j-1)\end{Bmatrix} \tag{2}$$

2.2.3 最長相同子序列

最長相同子序列(Longest Common Subsequence)之基本概念是找出兩序列間最長的相同序列部分，並以兩序列間相同部份長度之比例來代表此二序列間的相似度。若相似比例超過門檻值，則判定這兩個序列是相似的。

但這樣無法處理兩個時間序列有不同的偏移量和振幅，因此有些研究提出改進的方式 [1]，以圖 9作說明：(1)兩個時間序列Q 和R，(2)忽略它們之間有差距的部分（gaps），(3)調整R的偏移量使其在垂直虛線上能和Q對齊，然後再縮放Q和R間振幅的比例，(4) 最後觀察配對的子序列之總長度佔Q和R序列之長度的比例是否超過門檻，若是則此二序列便視為相似。

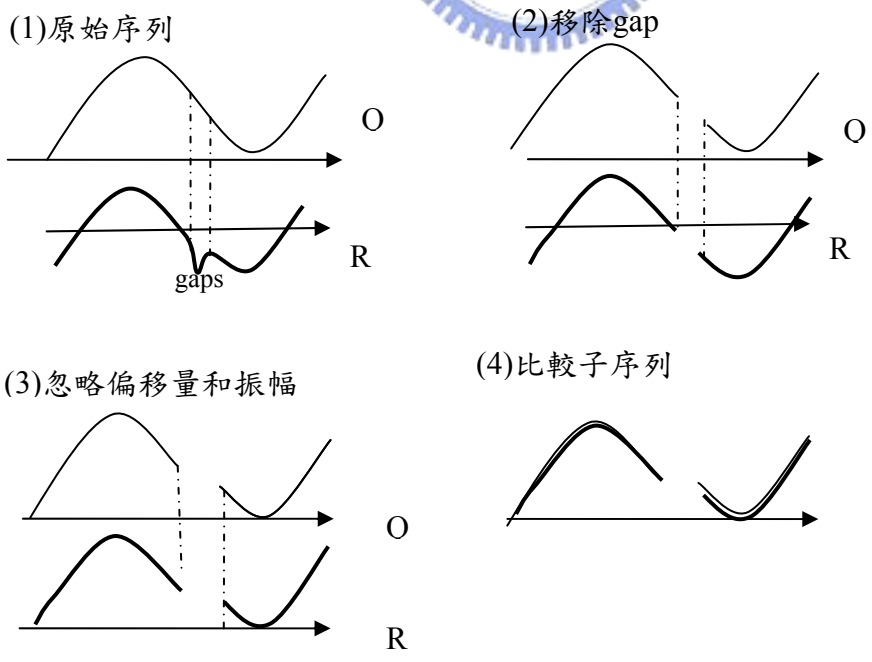


圖 9 子序列配對圖

要達到圖 9所述的概念有以下三種方式：

I. **Atomic Matching**：從O和R序列中找出沒有gap且長度為w的子序列之配對，稱之為" window pairs "。為了減少振幅和偏移量，再將每個window pair正規化。

II. **Window Stitching**：在任兩個window pair 之間可能有兩種情形：一種是在它們之間有重疊部分，而且沒有gap；另一種則是在它們之間沒有重疊部分；因此把這兩種情形都縫合起來則變成一個相似的子序列。

III. **Subsequence Ordering**：從上一步驟得到的相似子序列，依照時間順序將它們重新排列成沒有重疊且對應部分之長度為最長的新序列。計算這兩個時間序列的最長相同子序列的長度。

2.3 分群方式

一般分群演算法中的資料並不會隨著時間的變動而有大幅度的改變，這種資料也稱為「靜態資料」。[8]將處理靜態資料的分群演算法整理為以下五大類：

I. 分割式(Partitioning methods)：k-means；k-medoids；fuzzy c-means；fuzzy c-medoids

II. 階層法(Hierarchical methods)：agglomerative；divisive

III. 密度基礎法(Density- based methods)：DBSCAN

IV. 格子基礎法(Grid-based methods)：STING

V. 模組基礎法(Model-based methods)：statistical；neural network

然而，在處理時間序列分群時，資料因為會隨著時間的變動而有所改變，所以時間序列的資料通常會盡可能轉換為靜態資料的形式，以便直接使用現有的演算法來進行分群，尤其是Partitioning methods、Hierarchical methods、和Model-based methods。本研究中只介紹時間序列最常使用的partitioning methods和hierarchical methods做說明。

2.3.1 分割式分群法

分割式(Partitioning methods)分群法所產生的結果為一個個明確的分割(partitions)，使得每個群聚可以完全分離。一開始會先指定群聚的數目，然後藉著反覆疊代運算，逐次降低一個誤差目標函數的值，直到目標函數不再變化，達到分群的最後結果。一般而言，分割式群聚法的目的是希望盡量減小每個群聚中，每一點與群聚中心（cluster center）的距離平方差（square error）。基本的分割式分群演算法方法有k-means algorithm[13]和k-medoids[14]，其分群概念為下列步驟：

1. 決定預期分群的數目c，並隨機選取c個啟始點，將之分別視為c個群聚的群中心。
2. 對每一個資料點x，尋找與之最接近的群中心，並將x加入該群聚。
3. 計算目標函數，如果保持不變，代表分群結果已經穩定不變，便可結束此疊代方法。
4. 計算新的群聚中心，等於該群聚中所有資料點的平均向量。並跳回步驟2。

兩者最主要的差異在於k-means是計算物件平均值來表示群中心；k-medoids是計算群聚內最具代表性的資料表示該群聚，需要額外的計算成本。但一般而言不論是用k-means或k-medoids，要如何挑選一個適合的分群數目，使得所產生的群聚不會太相近且差異明顯，是個值得深思的問題。

2.3.2 階層式分群法

階層式分群法(Hierarchical methods)[6]主要透過分類樹狀圖(dendrogram)的建立，依其進行的方式分為聚合式(agglomerative)與分裂式(divisive)，以圖 10 描述聚合式與分裂式兩種方法，在一個包含五個物件的資料集合{a, b, c, d, e}上的處理過程。其概念為將所要處理之資料集合的資料點，利用聚合或分裂的方式，將彼此相似度高的較小群集合併成較大的群集，或者將較大的群集進行分離所產

生之樹狀結構，可以彈性地依據使用者不同的需求，對資料集合產生不同的群集數量。

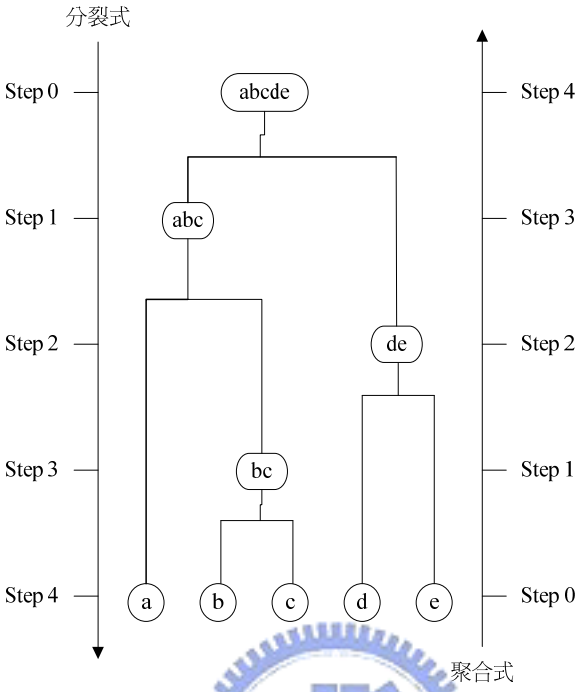


圖 10 階層式演算法處理流程[2]

I. 聚合式

聚合式(agglomerative)由樹狀結構的底部開始層層聚合。一開始將每一筆資料視為一個群聚，假設全部擁有 n 筆資料，則將這 n 筆資料視為 n 個群聚，亦即每個群聚包含一筆資料，再找出所有群聚間，距離最接近的兩個群聚，合併最接近的兩個群聚成為一個新的群聚，或直到滿足終止條件為止。

II. 分裂式

分裂式(divisive)方法所採用的策略與聚合式方法正好相反，由樹狀結構的頂端開始層層分裂。一開始時將全部的資料物件視為同一個群聚，然後尋找相異度最高的群集，再往下一個階層分裂成較小的子群聚；經由反覆進行群集分裂的步驟，直到每個子群集都只有一個物件，或是符合終止條件為止。

判斷兩個群聚是否可以合併或分裂，即計算群聚間距離方式，有以下四種方式：

I. 單一連結聚合(single-linkage agglomerative algorithm)：群聚與群聚間的距離可以定義為不同群聚中最接近兩點間的距離，圖 11 群聚間距離方式示意圖圖 11(a)：

$D(C_i, C_j) = \min_{\mathbf{a},\mathbf{b}} d(\mathbf{a}, \mathbf{b})$，其中 **a** 屬於 C_i 且 **b** 屬於 C_j。

II. 完整連結聚合(complete-linkage agglomerative algorithm)：群聚間的距離定義為不同群聚中最遠兩點間的距離，圖 11(b)：

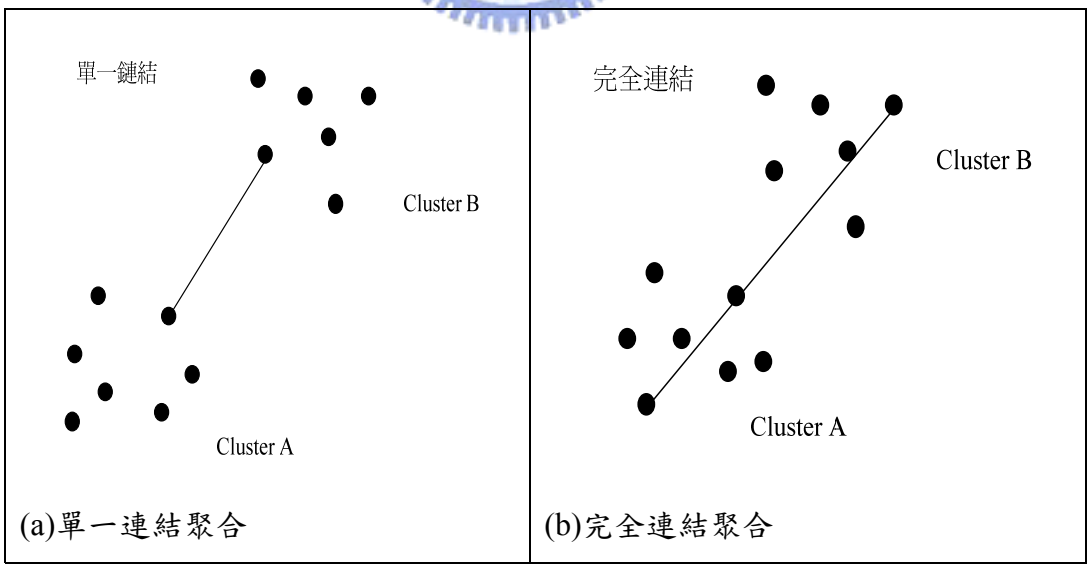
$D(C_i, C_j) = \max_{\mathbf{a},\mathbf{b}} d(\mathbf{a}, \mathbf{b})$，其中 **a** 屬於 C_i 且 **b** 屬於 C_j。

III. 平均連結聚合(average-linkage agglomerative algorithm)：群聚間的距離則定義為不同群聚間各點與各點間距離總和的平均，圖 11(c)：

$D(C_i, C_j) = \sum_{a,b} d(\mathbf{a}, \mathbf{b})/(|C_i||C_j|)$，其中 **a** 屬於 C_i 且 **b** 屬於 C_j。

IV. 沃德法(Ward's method)：群聚間的距離定義為在將兩群合併後，各點到合併後的群中心的距離平方和，圖 11(d)：

$D(C_i, C_j) = \sum_a |\mathbf{a}-\mathbf{m}|^2$，其中 **a** 屬於 $C_i \cup C_j$，**m** 表示 $C_i \cup C_j$ 的平均值。



(a)單一連結聚合　(b)完全連結聚合

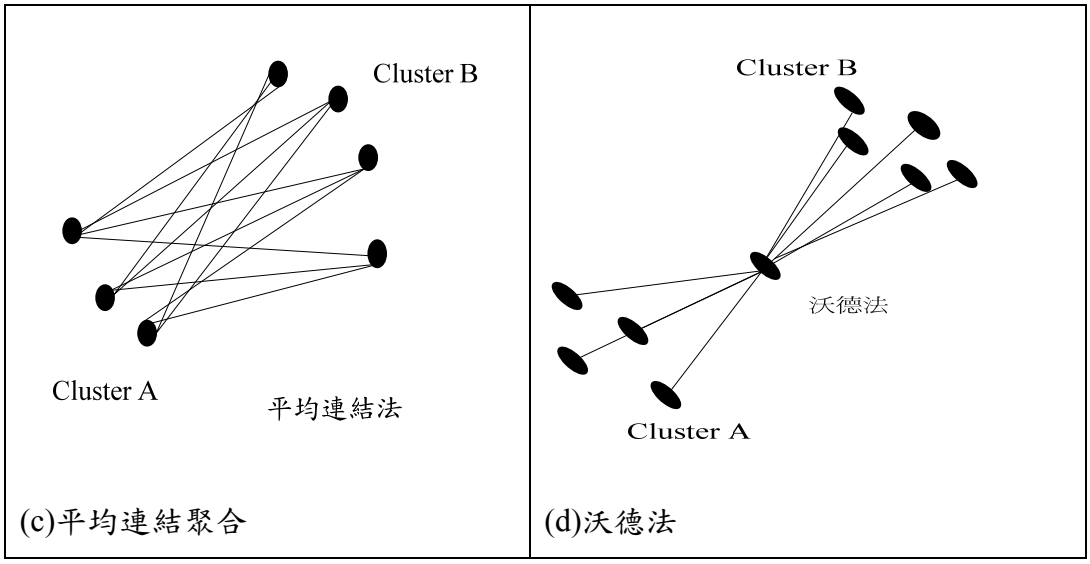


圖 11 群聚間距離方式示意圖

階層式分群法的概念雖然簡單，但是經常會遇到合併或分裂點選擇的困難。因此，若選擇了不適當的合併或分裂的條件，可能會導致最終分群結果不佳。

2.3.3 群聚量化評估方法

2.3.3.1 以群聚分佈評估分群結果

分群的結果無不希望得到高內聚力低耦合力。在[18]提到一般分群結果的良窳可以計算群的內聚力(cluster compactness)和群的分離度(cluster separation)以及綜合前兩個的整體分群品質(overall cluster quality)。

I. 內聚力：

內聚力的公式為(3)：

$$Cmp = \frac{1}{C}\sum_{i}^{C}\frac{v(c_i)}{v(X)} \tag{3}$$

其中v(X)為文件向量X的變異數，計算方式為公式(4)。

$$v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \bar{x})} \tag{4}$$

N為文件總數，$\bar{x}$ 表示所有文件向量的平均，$d(x_i, \bar{x})$ 為向量 x_i 和文件向量平均的

距離。$v(c_i)$為群c_i的變異數，計算方式和v(X)大致相同，以群為單位，計算群內文件和群中心的距離，計算公式為(5)。

$$v(c_i)=\sqrt{\frac{1}{c_i}\sum_{j=1}^{c_i} d^2(c_{ij},\overline{c_i})} \tag{5}$$

計算完$v(c_i)$和$v(X)$後，將每群的變異數除上$v(X)$後取平均，即得Cmp，為群的內聚力；Cmp的值介於0~1，當Cmp值越小，表示每一群的內聚力愈強。

II. 分離度：

分離度計算方式，為公式(6)

$$Sep=\frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\exp\left[-\frac{d^2(x_{c_i},x_{c_j})}{2\sigma^2}\right] \tag{6}$$

其中$d(x_{c_i},x_{c_j})$為群c_i和群c_j群中心距離，σ為Gaussian Constant，Sep介於0~1之間，當Sep愈小，表示群和群間的分離度愈大。

III. 整體分群品質(Overall Cluster Quality)

將上述得到的Cmp和Sep利用線性組合得到綜合分數，稱之為Overall Cluster Quality，來評估整體分群的品質。公式為(7)

$$Ocq(\beta)=\beta\times Cmp+(1-\beta)\times Sep \tag{7}$$

β介於0~1，當Ocq值愈小，表示群聚分佈效果愈好。

2.3.3.2 以專家評估推薦結果

在進行人工判別實驗結果時，必須評估專家對推薦結果的同意度，通常利用可信度(Reliability)及有效性(Validity)來區別。可信度指專家在評估過程中標示的一致性，而有效性是指專家評估的樣本中可用的樣本數。本研究採用Kappa Statistics.

[20]評估專家的同意度，以表 3為例說明之。假設有29位病人，分別由兩位醫生診斷病情，Yes表示診斷結果為不健康，No則表示健康，則Kappa

Statistics計算如下所示。

表 3 Kappa範例

		DoctorA		
		No	**Yes**	**Total**
DoctorB	**No**	10 (34.5%)	7(24.1%)	17(58.6%)
	Yes	0(0.0%)	12 (41.4%)	12(41.4%)
Total		10(34.5%)	19(65.5%)	29(100%)

Kappa = (Observed agreement - Chance agreement)/(1 - Chance agreement)

Observed agreement = (10 + 12)/29 = 0.76

Chance agreement = 0.586 * 0.345 + 0.655 * 0.414 = 0.474

Kappa = (0.76 - 0.474)/(1 - 0.474) = 0.54

並計算標準差得到0.134%，在信賴水準達95%時的信賴區間為(0.279, 0.805)，對照kappa評分表格，表 4，得知Kappa同意度介於Fair和Almost perfect。

表 4 Kappa參考對照表

Kappa	**Strength of agreement**
0.00	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

在評估推薦樣本時，採取專家一致性的意見當作樣本，計算推薦正確率(accuracy)。當被專家標示為相似，表示這兩群表達的是相似的概念；同理，標示不相似，表示表達的不為相似的概念。進而計算正確率。計算方式為公式(8)，事件分布為表 5。

表 5 專家標示兩兩群相似事件分布

		專家標示	
		Y	N
分群結果	Y	t_pos	f_pos
	N	f_neg	t_neg
		pos	neg

t_pos：專家標示為相似而分群結果也為相似。

f_pos：專家標示為不相似，但分群結果卻為不相似。

f_neg：專家標示為相似，但分群結果卻為相似。

t_neg：專家標示為不相似，分群結果也為不相似。

pos：專家標示為相似的樣本數。

neg：專家標示為不相似的樣本數。

$$sensivity = \frac{t\_pos}{pos}$$

$$specificity = \frac{t\_neg}{neg} \tag{8}$$

$$accuracy = sensitivity \times \frac{pos}{pos+neg} + specificity \times \frac{neg}{pos+neg}$$

第三章　時間序列分群方法與實作

在本章中將闡述本研究如何將時間序列分群應用在社會性標籤中。圖 12為研究步驟示意圖，首先說明本研究前置處理步驟，為圖 12虛線框部分；再來描述產生標籤時間序列的方法；接著進行標籤時間序列的分群演算法；最後進行標籤與相關網頁之推薦。

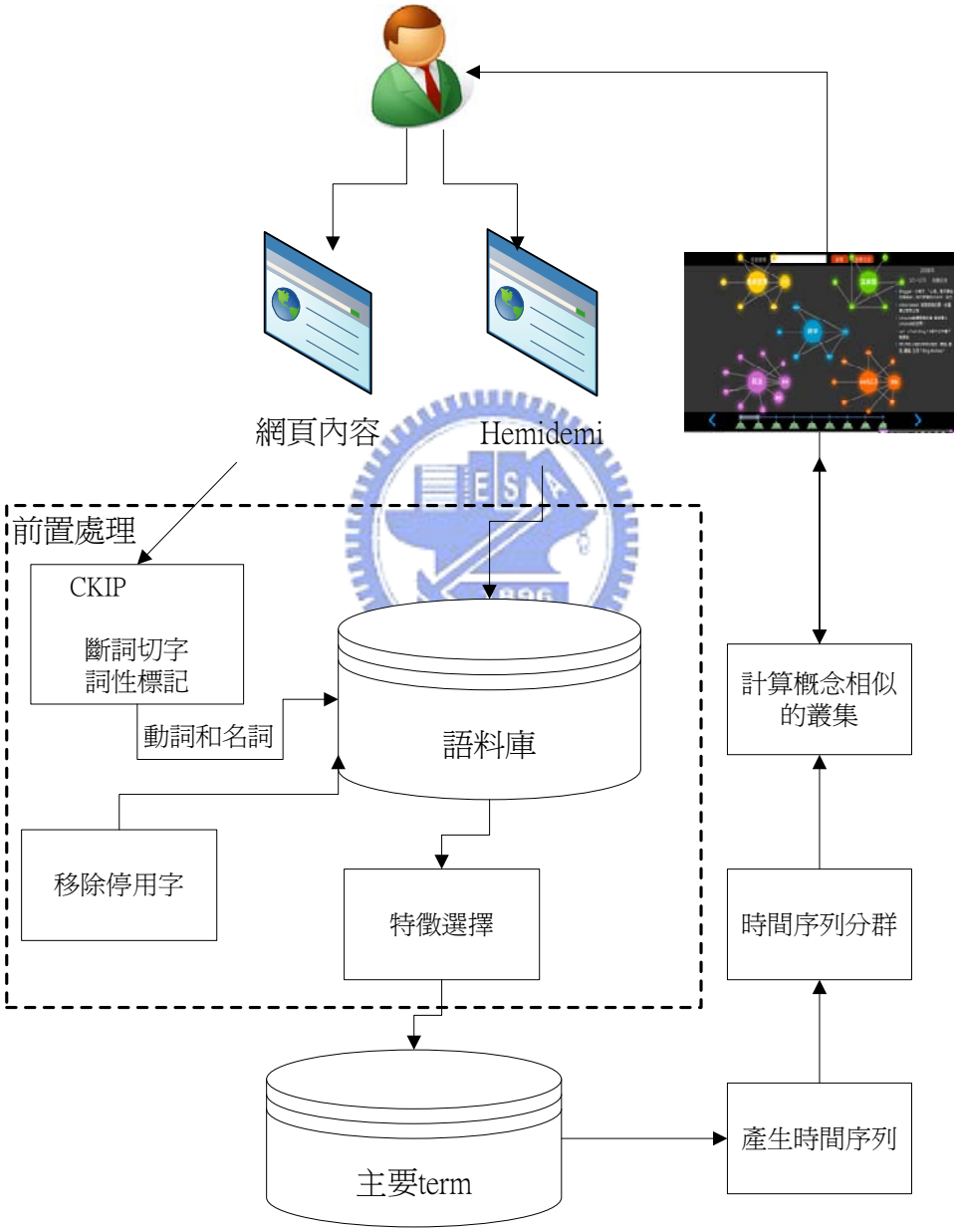


圖 12 研究步驟示意圖

3.1 資料收集

3.1.1 資料來源

本研究收集台灣線上書籤網站HEMiDEMi-黑米共享書籤[27]中的部分書籤以及這些標籤所標記的網頁內容為語料庫(Corpus)。

3.1.2 資料收集方式

圖 13為黑米共享書籤網站中某位使用者所收集的書籤，本研究係以使用者的標籤、收藏的網址和標題、收藏日期等做為資料的來源(如圖 13中黑線框所示)，並根據收集到的網址，到該網頁取回網頁內容資料。因此每收集一筆書籤資料，資料庫中就會記錄使用的標籤(tagID)、標記網頁網址(url)、網頁標題(title)、網頁內容的詞彙(token)、以及標記日期(createdate)這五大項。

圖 13 HEMiDEMi使用者收藏書籤資料

(資料來源：HEMiDEMi黑米共享書籤)

以圖 13為例，假設要蒐集為「Google不會翻譯古詩，那…你會嗎？-左右腦的對話」這則書籤。則資料庫中會記錄的資料如表 6：

表 6 蒐集書籤實際範例

Title	url	Tags	Createdate	token
Google 不 會 翻譯古詩，那…你會嗎？-左右腦的對話	http://chiron.nicetypo.com/contentbypermalink/b260ad8a6843a2525827f862df2e2967	tvbs 主播 媒體 影片 機器公敵 生活 科技 職場 葉慈 評論 辛棄疾 電影 電視	2009/5/25	…Google翻譯來翻『辛棄疾』的古詩文，翻的牛頭不對馬嘴云云... 主播們一副發現新大陸的樣子，…

本研究以2008/01/01~2008/12/31的資料為主，共蒐集了3,842個使用者標記的網頁，2707個不重複標籤。

3.2 前置處理

前置處理之主要目的在於過濾語料庫中不必要的字詞與雜訊，以避免這些低代表性的字詞降低分群結果的正確性。本研究的前置處理步驟包含：中文斷詞切字處理、詞性標記、刪除停用字、特徵選取等。

3.2.1 斷詞切字和詞性標記

斷詞切字(Tokenization)的目的在於找出文字的分界並從文字資料中擷取含有意義的詞。由於中文是以「字(Character)」為基礎的語言，包含單字詞或多字詞等型態，詞和詞間的界線不明確，不像英文以空格或標點符號來區隔句子和單字。本研究採用中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group, CKIP)所研發之中文斷詞系統同時進行斷詞切字及詞性標記，其斷詞正確率約為 95%-96%[25]。方法為利用中研院中文詞知識庫小組提供的API，資料的交換方式採用 XML，用戶端可自行撰寫程式經由 TCP Socket 連線傳送驗證資訊及文本至中研院中文斷詞系統的伺服器，伺服器經過處理後經由

原連線傳回結果。

經由CKIP斷詞系統處理過的文件會以XML格式傳回結果，結果中會對每個切割出來的詞以括弧標示詞性(part-of-speech)，每個詞之間以全形空白隔開。表 7為原文輸入CKIP中文斷詞切字系統，表 8為輸出結果。詞性對照表詳見附錄一。

表 7 CKIP原文輸入實例-原文

《貧民百萬富翁》的拍攝地點和故事背景設於印度，敘述一名來自貧民窟的青年，到孟買參與遊戲問答節目《百萬富翁》，當中過程非常順利，他答對了一條又一條的難題，卻被節目主持及警方懷疑其作弊。之後在警方的拷問下，他道出一段段與題目相關的往事。

表 8 CKIP原文輸入實例-斷詞切字與詞性標記

《 (PARENTHESISCATEGORY) 貧 民 (N) 百 萬 富 翁 (N) 》 (PARENTHESISCATEGORY) 的(T) 拍攝(Nv) 地點(N) 和(C) 故事(N) 背 景 (N) 設 (Vt) 於 (P) 印 度 (N) ， (COMMACATEGORY)</sentence><sentence> ?(QUESTIONCATEGORY) 述 (Vt) 一(DET) 名(M) 來自(Vt) 貧民窟(N) 的(T) 青年(N) ， (COMMACATEGORY)</sentence><sentence> 到(P) 孟買(N) 參與(Vt) 遊戲(N) 問答(N) 節目(N) 《(PARENTHESISCATEGORY) 百萬富翁(N) 》 (PARENTHESISCATEGORY) ，(COMMACATEGORY)</sentence><sentence> 當 中 (N) 過 程 (N) 非 常 (ADV) 順 利 (Vi) ， (COMMACATEGORY)</sentence><sentence> 他(N) 答對(Vt) 了(Di) 一 (DET) 條(M) 又(ADV) 一(DET) 條(M) 的(T) 難題(N) ， (COMMACATEGORY)</sentence><sentence> 卻(ADV) 被(P) 節目(N) 主持(Vt) 及(C) 警方(N) 懷疑(Vt) 其(DET) 作弊(Vi) 。 (PERIODCATEGORY)</sentence><sentence> 之後(N) 在(P) 警方(N) 的 (T) 拷問(Vt) 下(POST) ，(COMMACATEGORY)</sentence><sentence> 他 (N) 道出(Vt) 一(DET) 段段(DET) 與(C) 題目(N) 相關(Vi) 的(T) 往事(N) 。(PERIODCATEGORY)

經過詞性標記後，並非所有的詞性都是需要的，本論文只保留精簡詞類中的名詞(N、Nv)和動詞(Vt、Vi)。其結果如表 9。

表 9 CKIP原文輸入實例-擷取動詞與名詞的結果

貧民(N) 百萬富翁(N) 拍攝(Nv) 地點(N) 故事(N) 背景(N) 設(Vt) 印度(N) 述(Vt) 來自(Vt) 貧民窟(N) 青年(N) 孟買(N) 參與(Vt) 遊戲(N) 問答(N) 節目(N) 百萬富翁(N) 當中(N) 過程(N) 順利(Vi) 他(N) 答對(Vt) 難題(N) 節目(N) 主持(Vt) 警方(N) 懷疑(Vt) 作弊(Vi) 之後(N) 警方(N) 拷問(Vt) 他(N) 道出(Vt) 題目(N) 相關(Vi) 往事(N)

3.2.2 刪除停用字

所謂「停用字(Stopword)」，是指某一些在資料中出現頻率極高的字，其在語料庫中的資訊鑑別能力不強，但容易造成資訊檢索的雜訊，例如英文中的：the、a、of、by、for；中文的：的、但是、你、我...等。若不將停用字刪除，則後續計算字詞的重要程度時，有些停用字會因此突顯出來，會誤以為有相當程度上的重要性。本論文的停用字列表係參考Oracle Text Reference[21]、中央研究院平衡語料庫詞集及詞頻統計[26]頻率最高的前100個詞中的動詞和名詞，再加以補強。表 10為本研究的停用字列表。

表 10 停用字範例

說 看 聽 讀 你 它 他 我 我們 你們 阿 妳 妳們 他們 自己 她 人 是 上 後 到 無 小 們 今 好 後 者 大 那 年 時 說 有 個 這 種 中 讓 此 做 沒有 位 想 其 高 沒 何 不同 一 兩 各 每 次 三 目前 希望 有關 包括 最近 是 引起 最後 加強 繼續 有 了解 過去 任 左右 經過 使得 相關 當時 進入 現在 需要 原因 如此 什麼 問題 學生 表示 公司 大家

3.2.3 特徵選擇

本研究以向量空間模型(Vector Space Model)來表示標籤和其所標記網頁間的關係，一個字詞就表示一個空間維度(Dimension)。若產生的向量空間相當大，但真正重要的詞彙不多，這樣除了會產生相當稀疏的向量外，也會浪費大量的時間在處理不重要的字詞上，甚至導致分群的正確性降低。因此為了節省運算時間及增加未來分群的正確性，將重要性不高且不具資訊鑑別能力的字詞刪除是相當重要的。本研究依照以下幾項規則來進行維度縮減[9]：

I. 刪除在語料庫中出現次數小於3篇文章的字詞；

II. 刪除在語料庫中出現次數超過5%文章的字詞；

III. 在剩下的文章中，若一個字詞在一篇文章中出現次數小於2次給予刪除，因為這個詞彙不足以來表示該篇文章概念。

IV. Log Likelihood Ratio

除了上述三種法則外，本研究還加入了Log Likelihood Ratio (LLR)作為選取特徵的方法。Log Likelihood Ratio (LLR)是Likelihood Ratio衍生的統計方法利用機率、統計的方式來測試兩個假設:虛無假設(Null Hypothesis(H_1)) 和對立假設(Alternative Hypothesis(H_2))何者發生的機率比較大。其優點為經由數學函數轉換後，可以產生一個易於計算的統計函數分布。在本研究中，H_1是指一個詞彙($term_i$)出現在該篇文章(d_x)的情況，和其他的詞彙一樣； H_2則是會有所不同。如果有個詞彙使用的情況和其他詞彙很不一樣，就代表這個詞彙應可拿來代表該篇文章，公式表示為(9)(10)(11)。

$$H_1 : P(term_i \mid d_x) = p = P(term_i \mid \overline{d_x}) \tag{9}$$

$$H_2 : P(term_i \mid d_x) = p_1 \neq p_2 = P(term_i \mid \overline{d_x}) \tag{10}$$

$$p = P(term_i \mid d_x) = P(term_i \mid \overline{d_x}) = P(term_i)$$
$$p_1 = \frac{P(term_i \cap d_x)}{P(d_x)} \tag{11}$$
$$p_2 = \frac{P(term_i \cap \overline{d_x})}{P(\overline{d_x})}$$

$term_i$和d_x的事件分布用表 11表示：

表 11 詞彙與文章的關係狀況

	d_x	$\overline{d_x}$
$term_i$	O_{11}	O_{12}
$\overline{term_i}$	O_{21}	O_{22}

O_{11}表示$term_i$在d_x中出現的頻率；O_{12}為$term_i$在d_x以外的出現頻率；O_{21} 是d_x中所有非$term_i$的詞彙的出現頻率；O_{22} 是所有非$term_i$的詞彙在d_x以外的出現頻率。

假設機率分佈是二項式分佈(Binomial Distribution)，如公式(12)

$$b(k;n,x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{12}$$

可得H_1和H_2的likelihood為公式(13)：

$$L(H_1) = b(O_{11}; O_{11}+O_{12}, p)\, b(O_{21}; O_{21}+O_{22}, p)$$
$$L(H_2) = b(O_{11}; O_{11}+O_{12}, p_1)\, b(O_{21}; O_{21}+O_{22}, p_2) \tag{13}$$

$$\lambda = \frac{L(H_1)}{L(H_2)} \tag{14}$$

$-2\log\lambda$就會等於(15)：

$$\begin{aligned} -2\log\lambda &= -2\log\frac{L(H_1)}{L(H_2)} \\ &= -2\log\frac{b(O_{11}; O_{11}+O_{12}, p)\, b(O_{21}; O_{21}+O_{22}, p)}{b(O_{11}; O_{11}+O_{12}, p_1)\, b(O_{21}; O_{21}+O_{22}, p_2)} \\ &= -2((O_{11}+O_{21})\log p + (O_{12}+O_{22})\log(1-p) - \\ &(O_{11}\log p_1 + O_{12}\log(1-p_1) + O_{21}\log p_2 + O_{22}\log(1-p_2))) \end{aligned} \tag{15}$$

將每個詞彙經過公式(15)運算後，可以求得每個詞彙的$-2\log\lambda$。

根據Koller etal.[5] D. Koller and M. Sahami, "Hierarchically classifying

documents using very few words," Stanford InfoLab, 1997.[5]認為在階層式分群中，一篇文章中取10～20個關鍵詞彙最能表示該文章主題概念。在[9]中也提到特徵值愈多愈會降低特徵值和原文的一致性(Coherence)，因為特徵值愈多愈會增加干擾(Noise)。因本研究中使用者所標記之網頁其主題包羅萬象，用字較為通俗且廣泛，所以本研究每篇文章取$-2\log\lambda$最高的50個字當做特徵值。表 12為使用LLR的選取範例，從這50個詞彙中，可以看出該文章是在描述一位來自奧地利之奧運舉重金牌選手其背後感人的故事。

表 12 特徵選取實例

金牌 德國 背叛 妻子 機器人 奧運會 舉重 奪得 動人 相片 來到 觀眾 獻給 出場 雅典 心願 亞軍 冠軍 奧地利 選手 背後 禮物 舉起 裏 服務站 訓練 家族 維也納 緣起 電視機 車禍 參觀 拿到 奧運 吻 力量 感動 淚水 光明 書籤 頒獎 離開 感人 成績 窩 口袋 大全 運動員 這時 賽

表 13為整理到現階段前置處理詞彙的結果：

表 13 前置處理作業結果

	總數	不重複
詞彙總數	1,760,840	123,830
刪除停用字後的詞彙	1,699,352	123,757
刪除df<3的詞彙	1,602,877	44,064
刪除df>5%的詞彙	0	0
刪除tf<2的詞彙	555,163	34,650
LLR特徵選取後	402,319	20,371

3.2.4 權重計算

本研究以向量空間模型來表示語料庫中的文件，從一文件擷取出能夠代表該

文件的特徵，並用這些特徵以向量的方式來表示此文件。在此模型下，每一個網頁內容的表示法為 $d_x = \{w_{x1}, w_{x2}, ..., w_{xi}, ..., w_{xn}\}$ ， w_{xi} 表示在文件中該詞彙($term_i$)的權重，當此詞彙不存在時，權重則設為零。本研究中權重計算方式為該詞彙($term_i$)的 $tf_{xi} \times idf_i$ ， tf_{xi} 為該詞彙在網頁 d_x 出現的次數， idf_i 為總文件數目除以包含該詞彙之文件的數目、再將得到的商取對數。每份網頁可能會被數個標籤標記，因此任一個標籤 tag_j 可以用所有被其所標記的網頁來表示。每一個標籤 tag_j 之向量維度為語料庫中經過特徵選取後詞彙的總數(n)。假設將時間因素加入考慮，在 p 這天， tag_j 標記於網頁 d_x ，利用空間向量模型可以將此表示為：

$$tag_{j,p,x} = \{w_{x1}, w_{x2}, ..., w_{xn}\},\ 若 tag_j\ 於 day\ p 標示於 d_x \quad (16)$$

$tag_{j,p,x} = \{0, 0, ..., 0\}$ 若 tag_j 於 $day\ p$ 未標示於 d_x .

若 tag_j 在 p 被標記於 d_x 、 d_y 等二網頁，當中又有相同的詞彙，則權重相加，表示為：

$$tag_{j,p} = tag_{j,p,x} + tag_{j,p,y} = \{w_{x1} + w_{y1}, ..., w_{xn} + w_{yn}\}; \quad (17)$$

綜合上述， tag_j 在某一天(p)可表示為(17)：

$$tag_{j,p} = tag_{j,p,1} + tag_{j,p,2} + ... + tag_{j,p,q} = \{W_{p1}, W_{p2}, ..., W_{pk}\} \quad (18)$$

其中 $W_{pk} = \sum_{x=1}^{q} w_{xk}$ (k=1, 2, …, n)；q為標籤 tag_j 所標示之文章總數。

3.3 時間序列表示法

時間序列是一群依其發生時間的先後順序排成序列的資料。在本研究中，認為事件發生應該都有它的延續性，藉由使用者標記的社會性標籤，可以看到一個事件的發展，例如：奧運的標籤可能每隔四年會有一比較頻繁段時間出現得，而與2008年奧運相關的「北京」、「水立方」這些標籤雖然在2008年時會頻繁出現，但是在2004年時並不會出現；因此，在分群時若能加入時間趨勢，便更能看出北

京、水立方和奧運間的時序關係。

3.3.1 產生時間序列資料

為了之後要產生時間序列的向量，本研究先進行每個標籤的正規化(Normalization)，目的是為了降低之後時間序列可能產生的偏移量，讓差異大但走勢相似的時間序列可以分在同一個群聚，如圖 14。

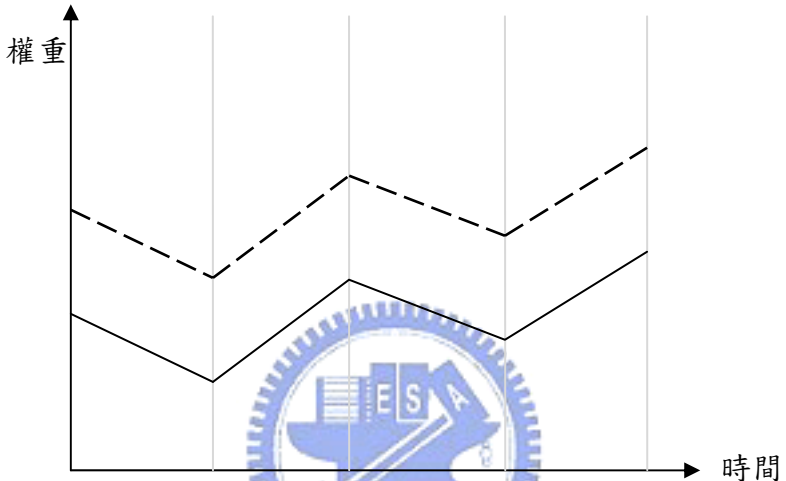


圖 14 偏移量

正規化方式為將每個標籤向量的各元素值除以各元素值平方和的平方根，公式如(18)。

$$tag_{j,p}=\left\{W_{p1}\Big/\sqrt{\sum_{k=1}^{n}W_{pk}^{2}},\ldots,W_{pn}\Big/\sqrt{\sum_{k=1}^{n}W_{pk}^{2}}\right\} \tag{19}$$

若要產生標籤tag_j在p的時間序列資料，則要把相鄰兩天的$tag_{j,p}$、$tag_{j,p+1}$，以大的日期減去小的日期，產生$v_{j,p}=tag_{j,p+1}-tag_{j,p}=\left\{W_{v_{jp},1},W_{v_{jp},2},\ldots,W_{v_{jp},k}\right\}$的方向向量，以一年365來說，$tag_j$的時間序列資料為這364個方向向量的集合。

根據[12]，時間序列資料可定義為包含N對的序列，(y_i,t_i)，i=1,…,N，y_i為在時間點t_i的值。本研究將每個tag_j產生的時間序列資料分割成M個區間，以大

寫V表示一個區間之時間序列的集合，$V_{j,m}$，$m=1,...,M$，每個區間中包含N對的時間序列資料$(v_{j,p},t_p)$，p=1,…,N (每個區間由N個時間片段組合而成)。

圖 15為tag_1在區間1中(1/1~1/15)的時間序列表示($V_{1,1}$)：

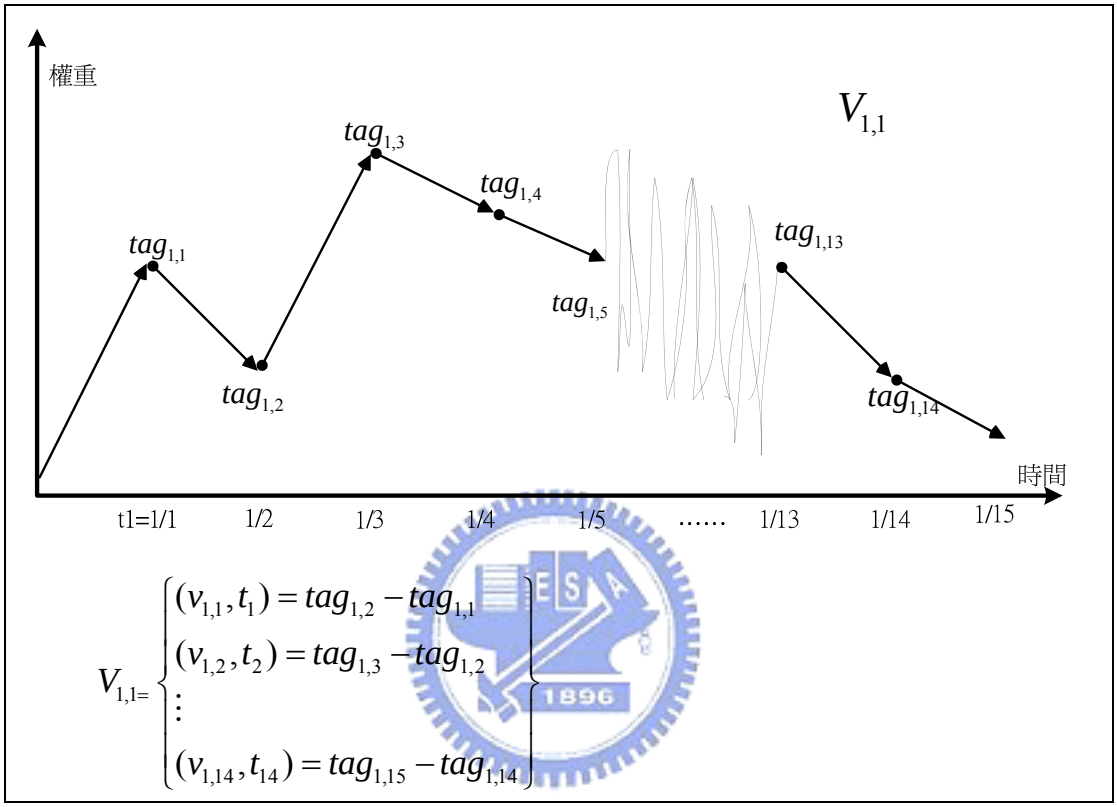


圖 15 時間序列資料表示圖

其中$v_{1,1}$表示tag_1在1/1~1/2這個片段(t_1)產生的方向向量；$v_{1,2}$表示tag_1在1/2~1/3這個片段(t_2)產生的方向向量。因此$V_{1,1}$表示tag_1在第一個時間片段(m=1)的產生的時間序列；$V_{j,m}$表示tag_j在第m個時間片段產生的時間序列。在本研究中M涵蓋的日期範圍係以每兩個星期、15天為一個分界線，因此一年364個時間序列裡有M=26個區間、每區間裡有N=14對序列。

3.3.2 計算時間序列相似度

本研究採用餘弦相似度(Cosine Similarity)計算任兩個標籤(tag_i, tag_j)間的相似度。每個標籤的時間序列是由前述N對序列向量所組成，因此各區間相似度為這N對序列的相似度平均，表述如下。

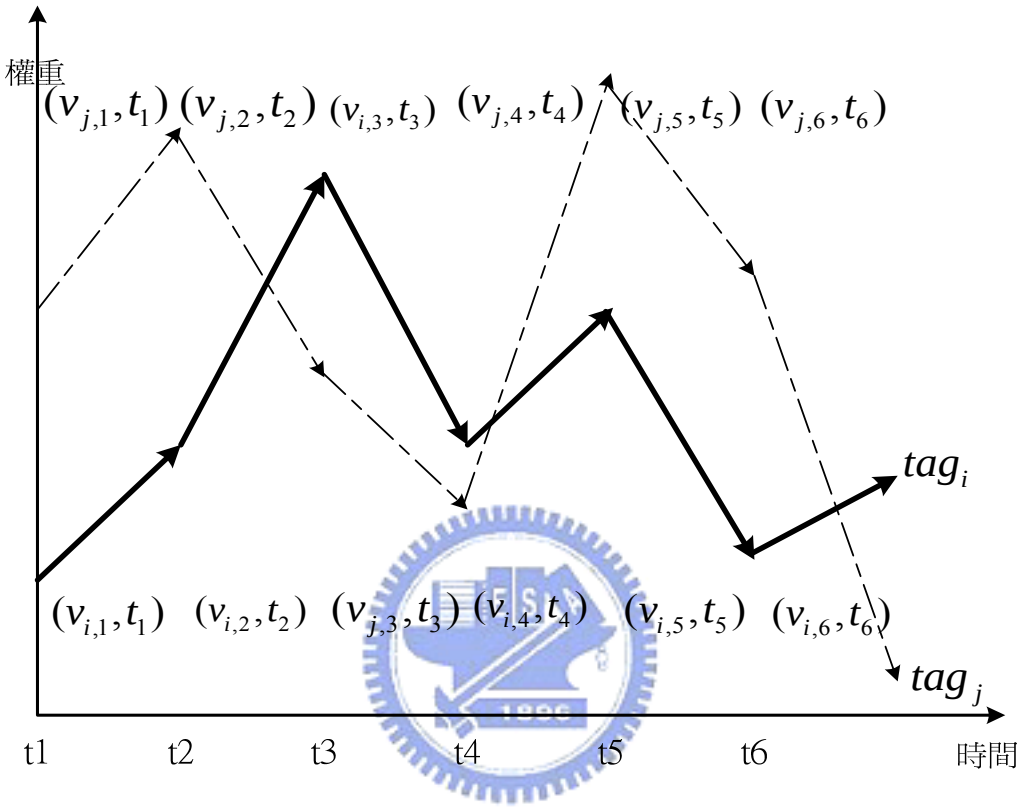


圖 16 某時間區間，兩個標籤向量

圖 16 中 tag_i 以實線粗體文字表示，$tag_i = \{v_{i,1}, v_{i,2}, ..., v_{i,N}\}$；$tag_j$ 以虛線表示，$tag_j = \{v_{j,1}, v_{j,2}, ..., v_{j,N}\}$；

$$sim(tag_i, tag_j) = (simlarity(v_{i,1}, v_{j,1}) + ... + simlarity(v_{i,N}, v_{j,N})) / N \tag{20}$$

$$simlarity(v_{i,p}, v_{j,p}) = \frac{\sum_{k=1}^{n}(W_{v_{ip},k} \times W_{v_{jp},k})}{\sqrt{\sum_{k=1}^{n} W^2_{v_{ip},k}} \times \sqrt{\sum_{k=1}^{n} W^2_{v_{jp},k}}} \tag{21}$$

又因方向向量有正有負，計算出的similarity會介於-1～1之間。負的向量解釋為

這兩個標籤走勢呈現反方向，因此不將此類列為考慮，將其相似度設為零。

然而在計算兩個標籤時間序列表示的相似度時，可能會有平移(Shift)的問題，如圖 17，某一標籤之表現走勢比另一標籤表現走勢慢上幾個時間點，例如有兩個標籤的表線走勢非常相似，但第二個標籤的時間序列曲線比第一個標籤慢了一個時間點。

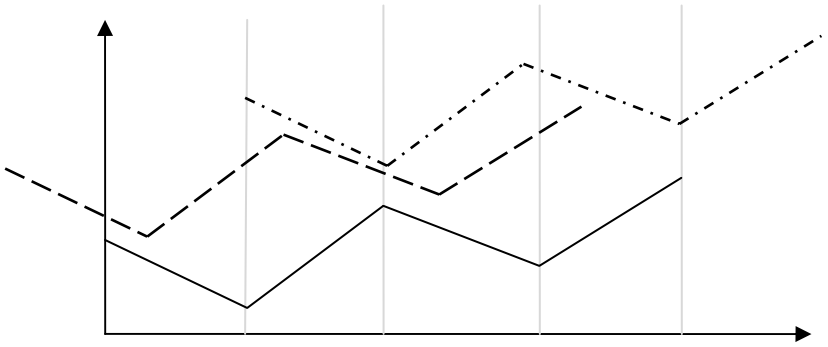


圖 17 時間序列平移

因此在計算相似度時，還要考慮將它納入考量。將原本的兩個標籤時間向量，一個為基準曲線，另一個為移動曲線，分別往前即往後平移1～4天，取相似度最高的當做平移的相似度，Sf，再將原本的相似度和平移的相似度加權計算，新的相似度公式如(20)，其中w為加權的權重，在本研究中設為0.5。

$$sim''(tag_i, tag_j) = w \times sim'(tag_i, tag_j) + (1-w) \times sf(tag_i, tag_j) \quad (22)$$

本研究共產生581,423組相似度，其中包含負向量有420,925組；剩下160,498組大於零的組，其分布狀況以表 14示之，平均值為0.00233：

表 14 相似度分布統計表

相似度區間	組數
0.5~0.851	1,079
1.71E-02~0.5	17,888
3.41E-03~1.71E-02	19,713
2.33E-03~3.41E-03	9,221

6.81E-04~2.33E-03	43,557
1.36E-04~6.81E-04	49,478
2.72E-05~1.36E-04	15,461
0.0~2.72E-05	4,101

3.4 時間序列分群

由前一小節利用餘弦公式算出標籤和標籤間的相似度後，要找出在各個區間內標籤的分群，分群的目標是達到同群內擁有高度的內聚力、而群間分離度也達到最大。本研究採用聚合式階層式分群法，採用平均連結聚合演算法計算群聚間的距離，如圖 18。

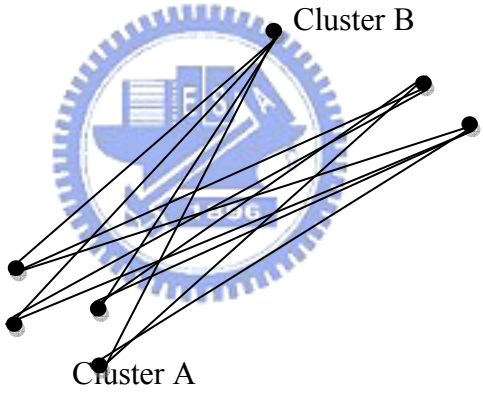


圖 18 平均連結聚合分群示意圖

步驟如下：

1. 選取某個時間區間(M)，在這時間區間裡，每一個標籤都視為獨立的一群。
2. 計算所有群聚間的平均值，群聚平均值公式為(22)。

$$D_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{a \in C_i, b \in C_j} d(a,b) \qquad (23)$$

$|C_i|$ 為群 C_i 裡的標籤數，C_j 亦同；d(a,b)為公式(19)計算的餘弦相似度。

3. 找出計算後平均值最高的兩群(C_i, C_j)，合併 C_i 與 C_j，成為新的群聚(C_{ij})。

4. 重複進行步驟2和步驟3，直到所有的標籤都歸到同一群，或群間的所有相似度平均值都小於門檻值。本研究所設定之門檻值為每個區間標籤相似度加總的平均值。

5. 回到步驟1，選取下一個時間區間。

3.5 推薦群聚

當分群完成後，分在同一群的標籤表示具有相同的概念外，也表示這些標籤的時間序列走勢是較為相像的。除了專注在同時間區間的分群結果，還考慮事件發生會有其延續性，因此將其他時間區間的分群結果納入考量。因此在這階段，本研究的推薦模式分為基於相同時間區間和不同時間區間的推薦方式：

3.5.1 同時間區間

在同一時間區間裡，本研究推薦給使用者和該群聚最相關之資訊以及說明群聚內標籤鏈結之關係，分述如下：

I. 推薦和該群聚最相關的文章

推薦給使用者的文章並不是只要推薦有被該群聚裡的標籤所標記的網頁，而是要從被標籤標記的網頁中，找出和此群聚相似度最高的網頁，依序推薦給使用者。計算方式同3.3.2計算標籤間相似度之方式，將該網頁(d_X)所產生的向量和此群聚的群中心做餘弦計算，$sim(d_x, \overline{C_i})$；其中群中心$\overline{C_i}$的計算方式為將群內所有標籤的時間序列相加再除以總標籤個數(z)，$\overline{C_i} = \sum_{j=1}^{z} V_{j,m} \Big/ z$。

II. 群聚內標籤和標籤相關文章

分在同一群聚的標籤表示他們有相近的概念和相似的時間序列，然而這些標籤所標記的文章可能有數筆，使用者要怎麼知道標籤tag_i和tag_j是有相關的？因此要找出tag_i和tag_j標記文章中，與tag_i、tag_j相似度最高的文章。

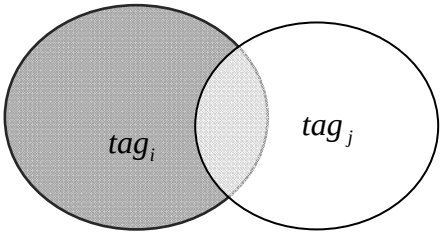


圖 19 tag_i和tag_j標記文章

tag_i和tag_j所標記的文章，並不會都相同，如圖 19，有時候甚至會沒有交集，會相似的原因為所標記的文章中有重複的詞彙。首先找出tag_i和tag_j所標記的文章，取聯集，再將tag_i和tag_j在這時間區間(m)的維度合併，($V_{ij,m} = V_{i,m} + V_{j.m}$)，重新計算有被$tag_i$和$tag_j$所標記的文章和合併產生的新向量的相似度，相似度最大者為最能涵蓋tag_i和tag_j概念的文章。

3.5.2 不同時間區間但概念相似的群聚

計算不同時間區間內群聚間之相似度，以萃取出擁有相似主題概念的標籤，推薦給使用者。首先找出每個群聚的群中心，利用餘弦公式計算群聚間的相似度，$sim(\overline{C_i}, \overline{C_j})$。因此當兩個群聚間的相似度大於門檻值時，則認為這兩個群聚只是在不同區間表達同一概念。在本研究中門檻值設為0.07，因每個時間區間由14個時間片段組成，認為群聚間的相似性起碼要涵蓋一個時間片段，才能表示是相似。

第四章　系統發展與結果分析

本章將首先介紹本研究開發的系統；其次以案例說明的方式，探討有無使用時間序列的差異，以及利用時間序列分群所發現之具有時間性的案例。在探討完有無使用時間序列的差異後，比較使用時間序列分群品質良窳。最後，再針對分群結果的一些例外狀況進行討論。

4.1 系統簡介

4.1.1　系統資料

本研究所開發之系統其資料來源為HEMiDEMi黑米共享書籤，由於黑米共享書籤之資料量甚多，本研究僅利用介於2008/01/01~2008/12/31的部分資料，使用了2,707個不重複的標籤，涵蓋了3,629個使用者標記的網站。

4.1.2　系統介面

系統呈現分兩層，第一層用以檢視不同時間區間中重要的標籤群聚，以圖 20為例，呈現出2008/01/29～2008/02/12這區間中資料集所分析出來的前五大群聚，前五大群聚表示為在這時間區間含有標籤數最多的前五大群，每一個群聚用一個顏色表示之，例如：左下角粉紅色表示政治群聚、左上角黃色表示體育群聚、右上角綠色表示閱讀群聚、右下角橙色表示旅遊群聚、中間天空藍表示Web 2.0群聚；在每一群聚中，首先找出該群聚下相似度最高的標籤組合，形成最大的節點，再以最大的節點當做主節點，根據關係程度強弱決定其他節點的大小，節點愈小表示和主節點的相似度愈小。於右方框中顯示和這五大群聚最相關且被使用者標記的網頁。

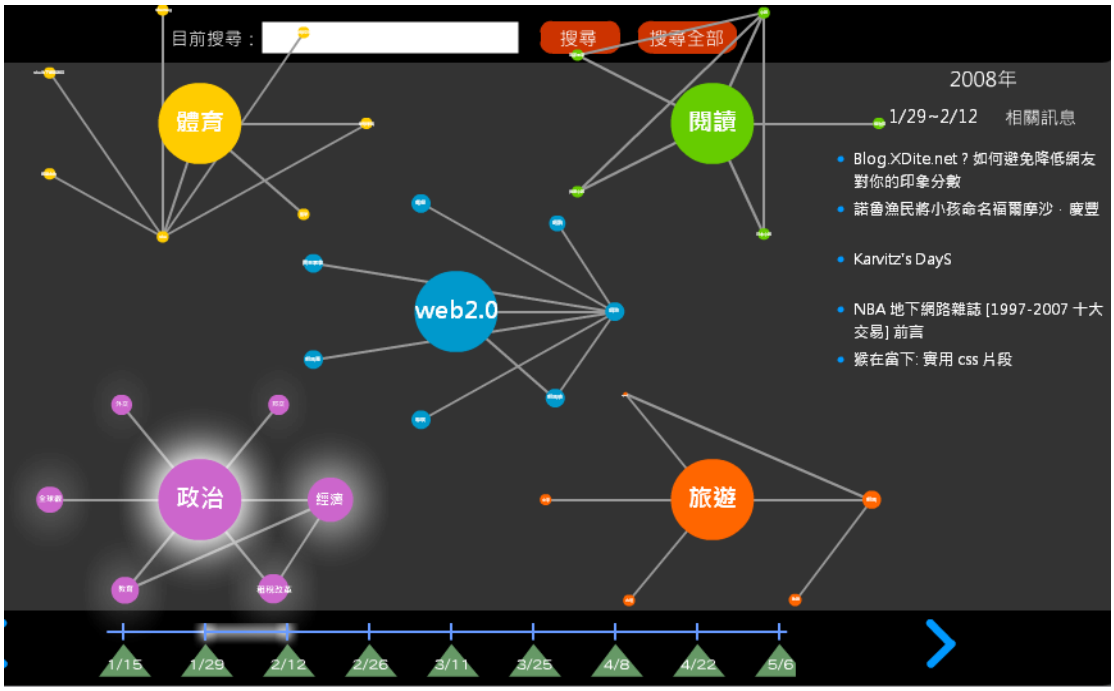


圖 20 2008/01/29～2008/02/12的主要群聚

點選任一節點或者在上方目前搜尋欄裡輸入想知道的標籤，進入到第二層，提供更詳細的節點關係。假設使用者想知道關於「電影」這標籤的變化趨勢，在上方搜尋鍵入「電影」後搜尋結果如圖 21。搜尋結果分四部分說明。

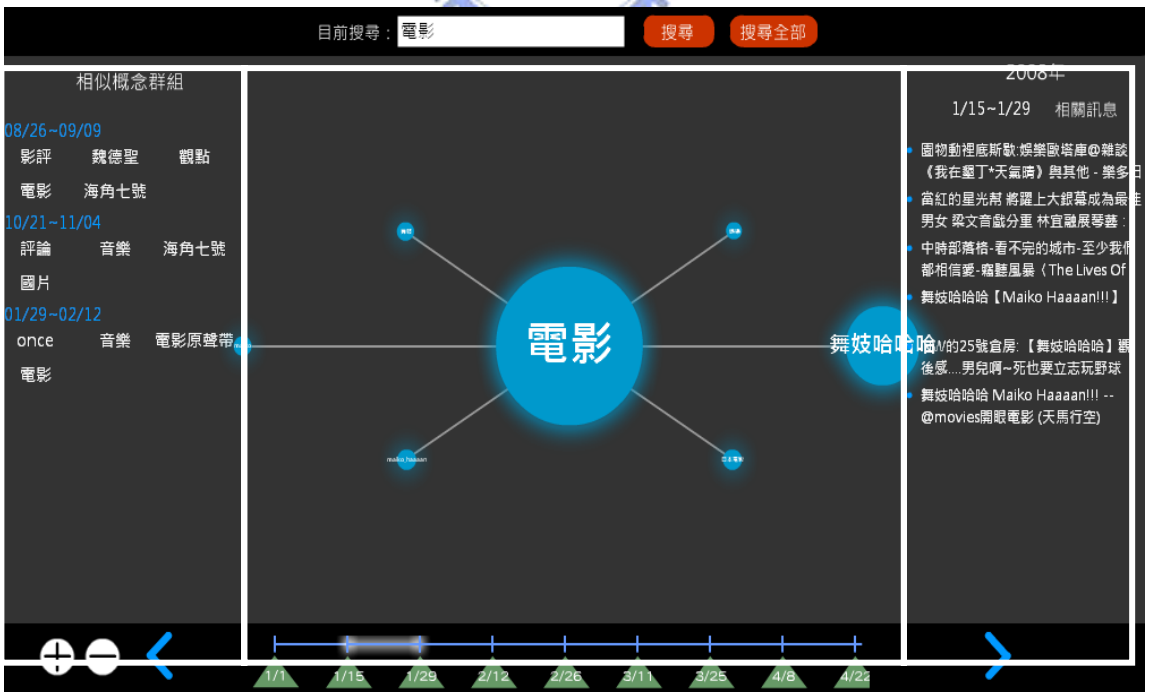


圖 21 搜尋「電影」標籤後結果

I. 右邊相關訊息：顯示電影標籤所在的這一群聚中，和此群聚最相關的網頁，並非只有和電影標籤相關的網頁。

II. 中間主體：顯示此時間區間中和電影標籤相似度最高的其他標籤。節點愈大表示該標籤和電影之關係愈強；節點間的連線，會顯示和這兩節點最相關的網頁標題。以圖 21為例，當滑鼠移動到電影標籤和舞妓哈哈哈標籤時，會顯示「舞妓哈哈哈【Maiko Haaan】」網頁標題(如圖 22)，表示這一篇網頁文章最能涵蓋這兩個標籤的概念。

圖 22 電影標籤和舞妓哈哈哈標籤

III. 左邊相似概念群聚：根據3.5.2的計算後，顯示其他時間區間和此群聚概念相似的群聚，供使用者參考，圖 23顯示其他區間和電影標籤群聚最相似群聚的標籤。

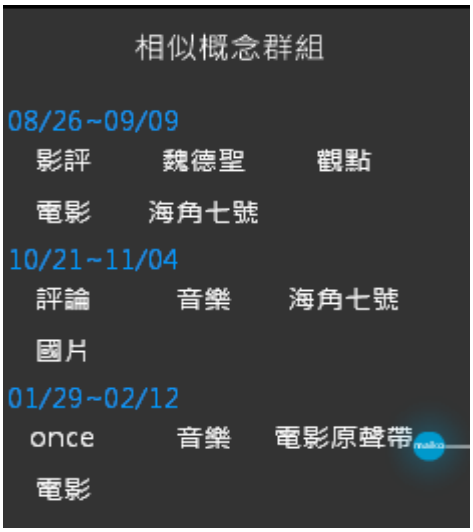


圖 23 與電影標籤群聚相似的其他區間群聚

IV. 下方時間軸：點選不同時間區間，檢視同一標籤因時間點不同，其群聚概念的變化情形。

4.2 質化評估

本節首先比較使用時間序列與否對分群結果的差異，模擬出可能的分群結果，再實際舉證本研究的結果。接著討論，同一個標籤因為時間點不同，所描述之事件所產生的變化。

4.2.1 一般分群與時間序列分群之比較

本研究將所蒐集之資料分別進行一般分群與時間序列分群，以下舉例說明二者分群結果所產生之差異。

由於本研究是以向量來表示標籤，當以一般分群方式計算標籤和標籤間的相似度時，會受以下兩種情形影響：

I. 若 tag_i 出現在多時間片段、tag_j 只出現在單一時間片段，但 tag_i 和 tag_j 重複的維度卻很多。在計算相似度時，因重複維度很多，相似度高的機會也增大，使得 tag_i 與 tag_j 容易分在同一群。假設 tag_i 在三個時間片段分別有50、10、15個不重複的詞彙；tag_j 在一個時間片段出現55個和 tag_i 重複的詞彙。計算相似度時，tag_i 和 tag_j 會有高相似度。

II. tag_i 出現在多時間片段，tag_j 只出現在單一時間，tag_i 詞彙所占的維度也很平均。假設 tag_i 出現在三個時間片段分別有10、15、17個詞彙；tag_j 有12個詞彙。但 tag_i 的詞彙權重遠遠超出其他標籤，即某一標籤在單一時間的權重太高，因此在計算和 tag_i 的相似度時，有機會造成相似度高於其他標籤。

若以時間序列分群來看，每一個時間片段所占的重要性是一樣的，不會受到是否有某個標籤獨占鰲頭的影響，或者單次出現頻率太高的影響。假設tag_j和tag_i在第一個時間片段相似度為0.9，但因為只占這時間區間的1/14(因一個時間區間有14個時間片段)，因此相似度只為0.06。若tag_k和tag_i在第一個時間片段相似度為0.2，在第二個時間片段相似度為1，在第三個時間片段，相似度為0.8，則在此時間區間的相似度為0.14(2/14)，高於tag_j的相似度。

取本研究中2008/7/29～2008/8/12這時間區間與中國標籤相關的其他標籤，包含奧運、北京奧運、bbc、台灣、政治、中國人物...等，如表 15黑色陰影表示它們有出現的時間片段。表 16顯示在這時間區間和中國標籤最相似的其他標籤。

表 15 標籤分布狀況

中國	■	■	■	■	■	■	■
奧運	■	■	■	■		■	■
北京奧運			■	■			■
bbc	■	■	■	■			
台灣			■	■	■		■
政治			■	■	■		
新聞自由	■						
	7/31	8/1	8/5	8/6	8/7	8/9	8/10

表 16 和中國最相似的其他標籤

一般分群方式		時間序列分群方式	
奧運	0.729	台灣	0.316
bbc	0.671	政治	0.270
新聞自由	0.563	奧運	0.265
台灣	0.547	bbc	0.205
政治	0.394	北京奧運	0.118
北京奧運	0.356	新聞自由	0.052

分群結果圖 24，在一般分群方式，中國、奧運、bbc、新聞自由等標籤分在同一個群組，因奧運標籤的詞彙和中國標籤詞彙的重複性很高；台灣標籤和中國標籤詞彙的重複性雖是次高，由於其在這時間區間涵蓋的範圍較廣，加總起來的權重卻沒有bbc和新聞自由標籤在單一時間點的權重來得高，因此沒有被分在同一群。而北京奧運，因為跟其他標籤相關性更高，所以沒有被分在這兩群中。

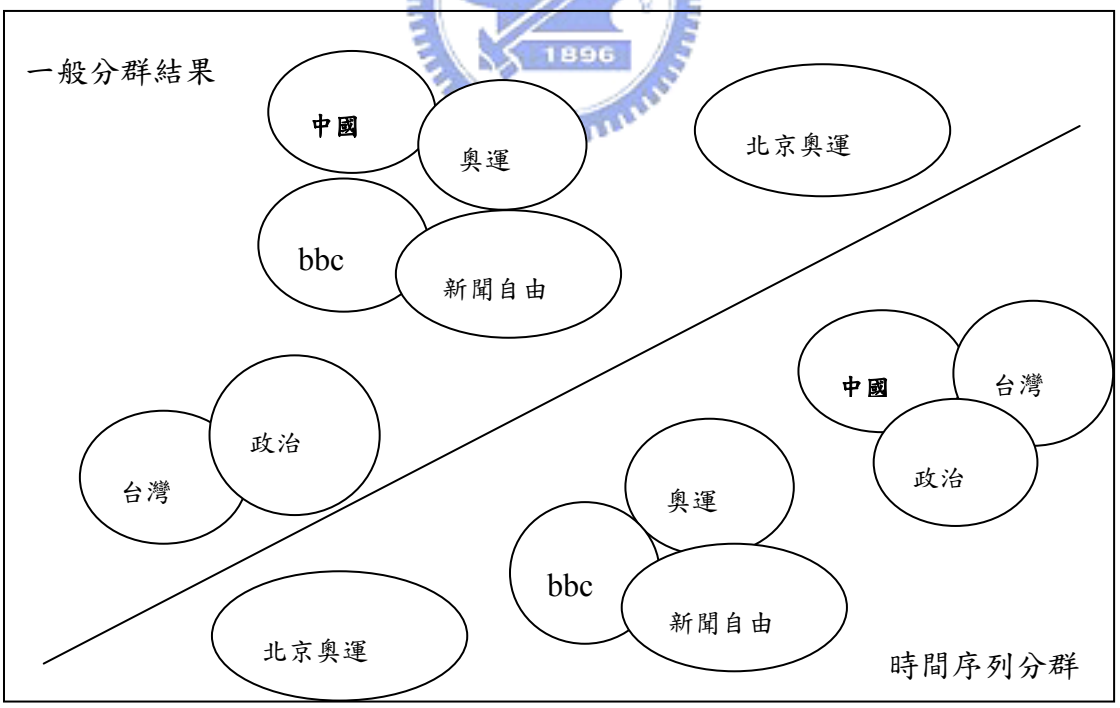


圖 24 分群結果

使用時間序列分群的結果認為台灣標籤和政治標籤對中國標籤其時間序列

走勢較為相像，如圖 25。因此將中國和台灣、政治等標籤放在同一群組；而奧運、bbc、新聞自由的時間序列較為雷同，故分在另一群。

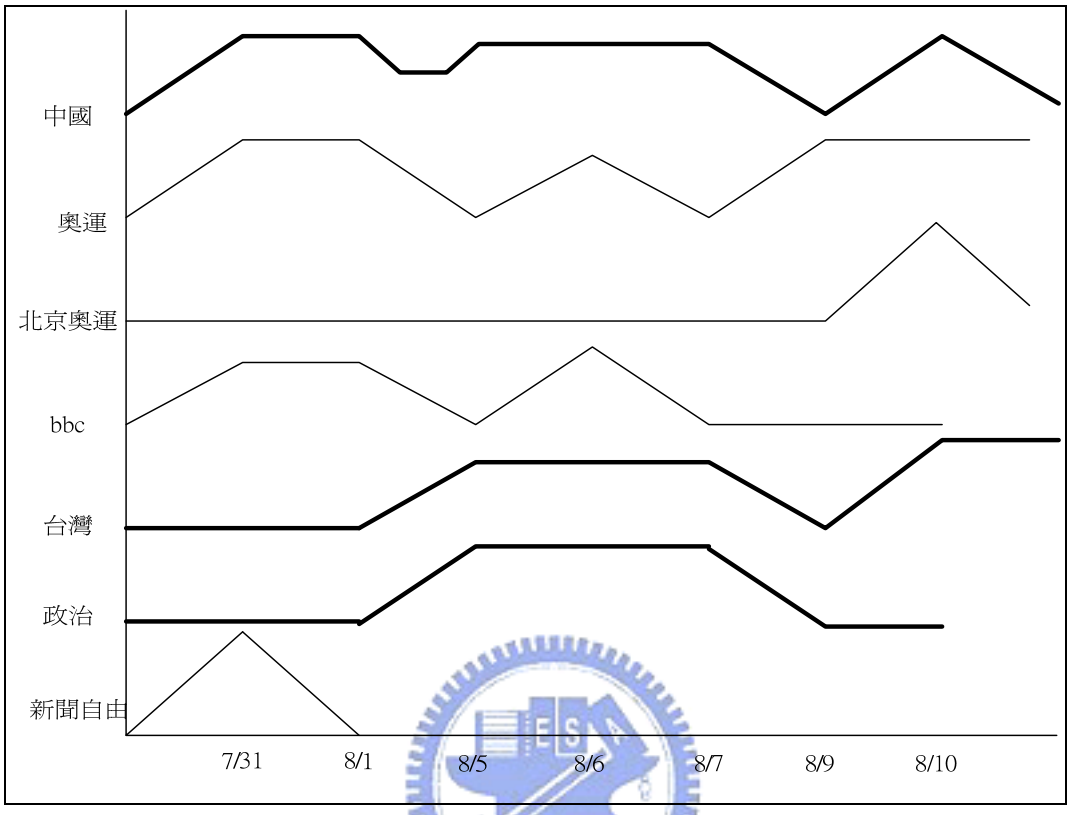


圖 25 標籤時間序列

4.2.2 個案分析

本節以個案說明的方式，說明相同的社會性標籤，隨著時間演變，所表示的群組概念所產生之變化。

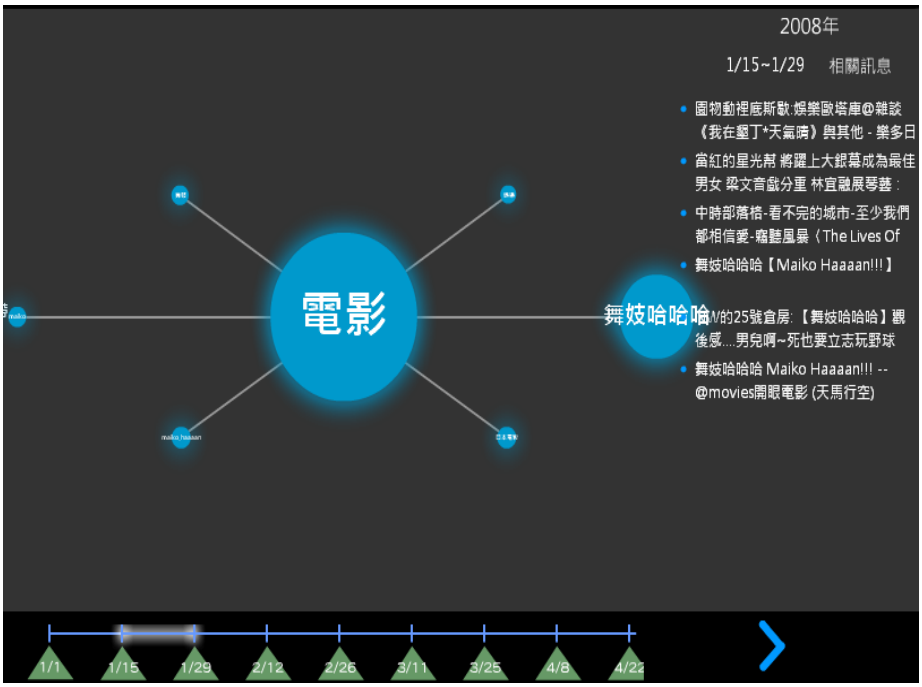
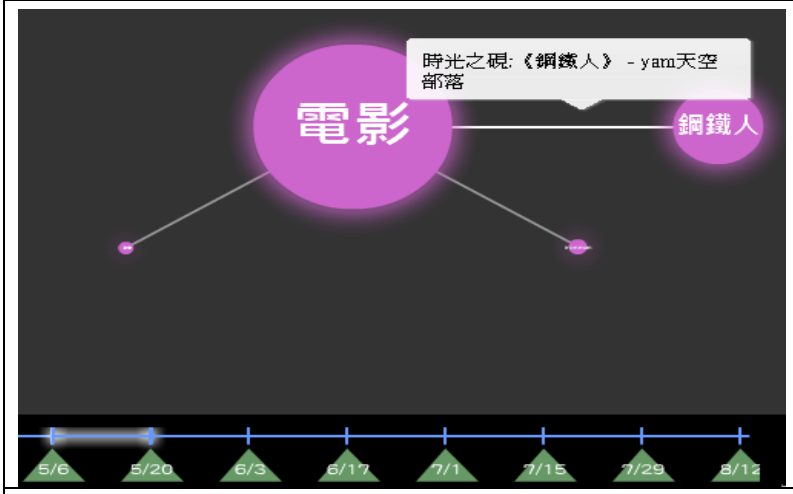


圖 26 電影標籤圖示

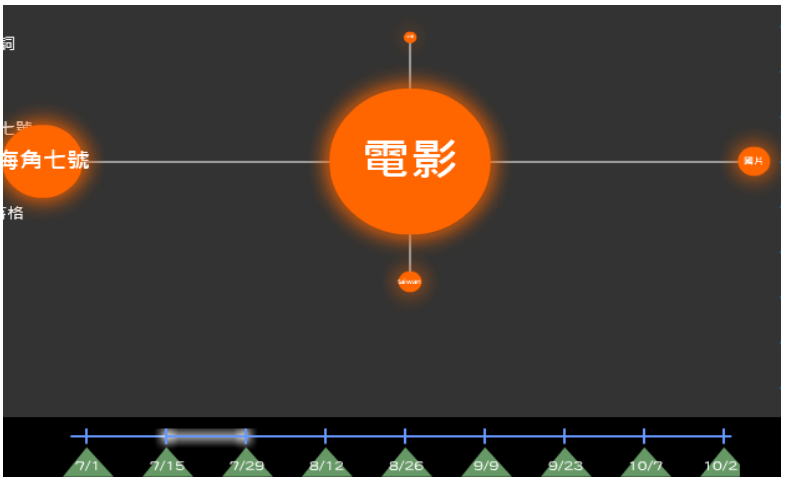
以圖 26電影標籤為例：

1. 在 1/15～1/29 中和電影標籤相關的標籤有「舞妓哈哈哈」、「日本電影」、「娛樂」…等。右邊框顯示和此群聚最為相關之被標記的網頁，有「舞妓哈哈哈 Maiko Haaaan!!! -- @movies 開眼電影 (天馬行空)」 、「NW 的 25 號倉房: 【舞妓哈哈哈】觀後感....男兒啊~死也要立志玩野球拳!!」…等網頁內容。移動下面時間軸可以看到電影標籤因為時間點不同，所描述事件的改變。
2. 當時間區間為 5/6~5/20 時，所引起討論的是鋼鐵人，如圖 27(a)。
3. 當時間區間為 7/15~7/29 時，電影海角七號上映了，如圖 27 (b)，但似乎沒有引發熱絡的討論，因為下一期 7/29~8/12 大家隨之討論的是動畫片一瓦力，如圖 27 (c)。
4. 但在上映將近一個多月(8/26~9/9)後，海角七號開始產生熱烈迴響，如圖 27 (d)，且熱度持續了一段時間。
5. 當時間區間為 9/23~10/7 時，海角七號和旅遊產生了關聯性，因為海角

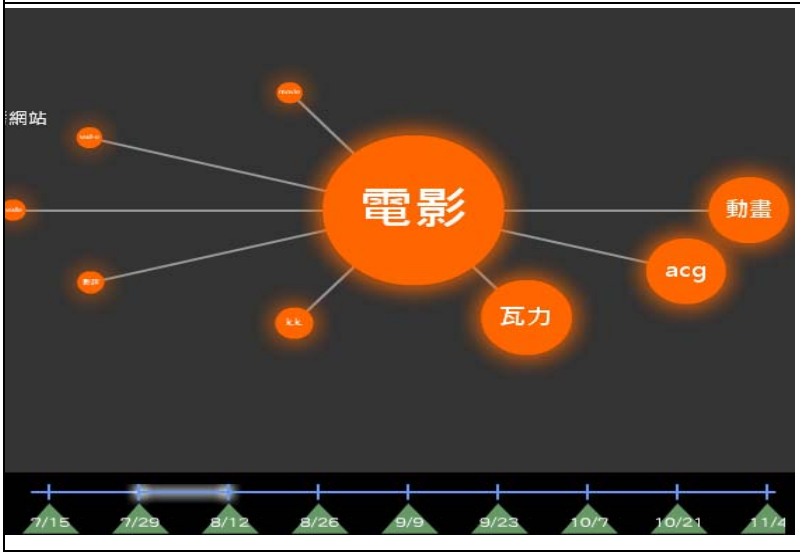
七號的效應，帶動了屏東一帶關於海角七號拍攝地點的旅遊觀光熱潮。右方相關網頁中也透露出網路使用者拜訪了劇中哪些景點。



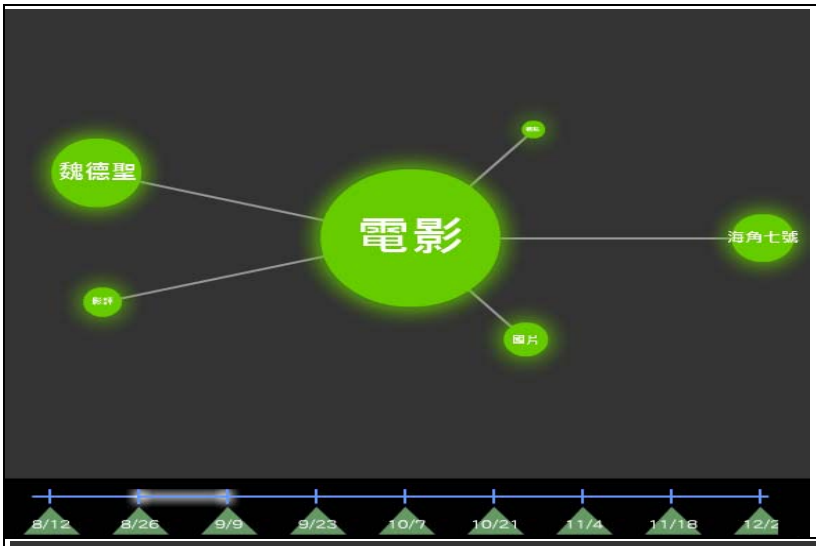
(a)5/6～5/20電影節點示意圖—以鋼鐵人與電影的相似度最高



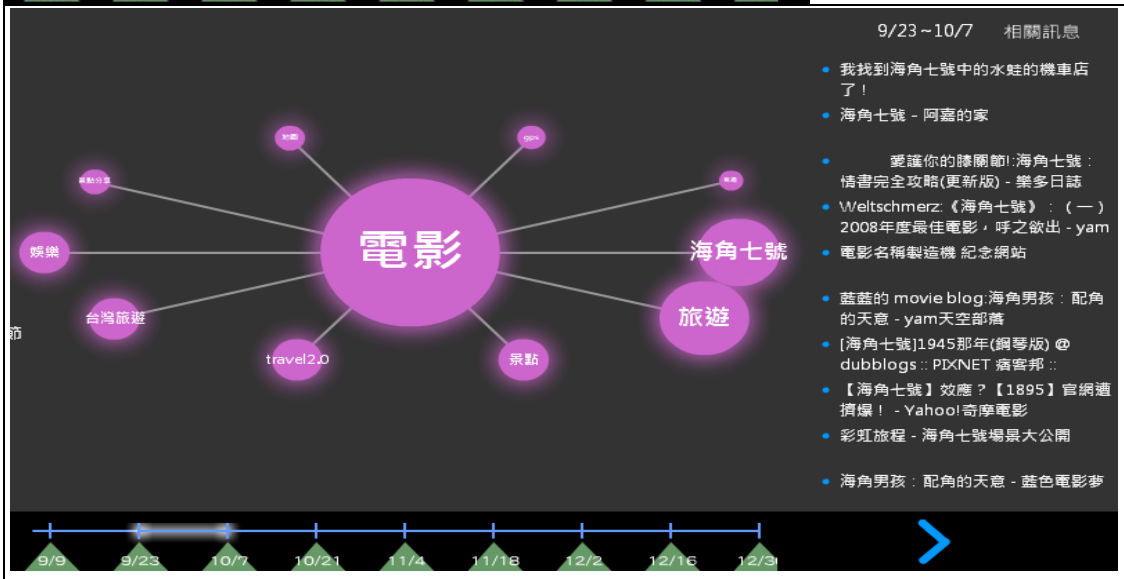
(b) 7/15～7/29電影節點示意圖—以海角七號與電影的相似度最高



(c) 7/29～8/12電影節點示意圖—以動畫、瓦力與電影的相似度最高



(d) 8/26~9/9電影節點示意圖—海角七號、魏德聖與電影的相似度最高



(e) 9/23~10/7電影節點示意圖—海角七號和旅遊產生了關連

圖 27 電影相關標籤示意圖

從這例子，利用社會性標籤和時間序列，使用者在網路上的互動得以彼此銜接，更將議題延伸到真實生活互動中，促成彼此相互的理解與知識的增長；並透過一些管理工具，達到動態及參與式的知識管理，能夠適應使用者，而非強迫使用者改變。

4.3 量化評估

要比較分群效果的好壞，方法很多種，本研究選擇使用計算群的內聚力和分離度。在評估系統推薦資訊的好壞上，使用Kappa Statistics.

[20]評估專家同意度，再以此作為計算推薦正確性之依據。

4.3.1 以群聚分佈評估分群結果

根據2.3.3.1，以距離表示相似度，故內聚力、分離度以及整體分群品質都是愈小愈好。本文使用時間序列分群結果和一般分群(沒有使用時間序列)的結果做比較。表 17統計這兩種方式分群結果的群數。

表 17 分群分布結果統計

一般分群		時間序列分群	
群中標籤數	群數	群中標籤數	群數
1	185	1	221
2	260	2	284
3	175	3	182
4	171	4	160
5	115	5	106
6～10	216	6～10	212
11～20	32	11～20	52
21～	8	21～	8
總數	1162	總數	1225

表 18列出在不同群聚個數下，一般分群和時間序列分群的內聚力、分離度。因本研究是採用聚合式階層式分群法，一開始就將每個標籤視為一群聚。當分群條件終止時，有些標籤還是自成群聚，當一個標籤為一群聚佔多數時，會影響計算內聚力和分離度的品質。因此本實驗分別計算當每個群聚至少要有幾個標籤才

能為一群聚計算其內聚力和分離度。表中「**群聚中至少有幾個標籤數**」，即為此意思。

表 18 群聚分佈評估結果

	一般方式			時間序列		
群聚中至少有幾個標籤數	**內聚力**	**分離度**	**整體分群品質 (β=0.5)**	**內聚力**	**分離度**	**整體分群品質 (β=0.5)**
1	0.2866	0.0044	0.1455	0.2730	0.0029	0.1380
3	0.3704	0.0074	0.1889	0.3645	0.0058	0.1852
4	0.4023	0.0077	0.2050	0.4055	0.0061	0.2058

經觀察知，本研究不論群聚中至少有幾個標籤數，使用時間序列分群方式在內聚力、分離度及綜合表現上和使用一般分群方式的結果，其差異並不會太大。因此使用時間序列分群並不會造成分群結果差異的偏頗。

表 18中，其分離度值蠻低的，推測原因為本研究蒐集的資料類型可能太過鬆散，符合門檻值條件的標籤數不多，因此在分群時容易界定出群聚間的界線，形成群聚間彼此較不相關。

4.3.2 以專家評估推薦結果

在本研究中，將語料庫內推薦結果的群聚隨機抽取250組配對組合，再由兩位有資訊背景的專家根據使用者使用的標籤和標記的網頁，判斷這兩群聚是否相似，表 19為系統及專家都判斷為相似的群聚範例；表 20為判斷不相似的群聚範例：

表 19 專家標示相似之群聚範例

標籤	觀光 全球觀 政治 國防 透視中國 租稅改革 邦交 呼籲藍綠同撤入聯 公投 經濟 教育 直航 外交	民生 政治 物價 退稅 經濟 公民 多元 中國 減稅
標記網頁	-[資訊視覺化] 石油與政治獻金 -馬英九:二○○八政黨輪替　經濟成長率百分之六?! -呼籲停收北一高過路費! 馬謝哪一個答應?! -史上最大 大陸觀光團明抵台 -科索沃將於2008/02/17日獨立 -展現國會新面貌 國民黨責任重大 -馬英九發誓說:展望2008,思考台灣的出路 -蕭萬長：民間競爭力強 問題是政策失誤	-「台灣稅制是不公不義」會計師：台灣富豪繳稅 不到收入1成 -別再叫我中華台北　四成想以「台灣」為名 -馬英九施政不滿意度暴增至4成2 -南韓退稅10兆韓元打賞75%的勞工! 馬退稅補貼中低收入「四六八方案」該不會也跳票吧!!! -馬上蓋殯儀館? 583億拿來凍漲! -苦熬140年 日本愛奴族獲原住民地位 -民運人士仍對中國民主有期待　公民運動可助力 -馬政府不管物價高漲?! 也沒事?! -馬小心點! 就任百日！ 李明博內閣擬總辭! 南韓物價7年新高 韓圜連6升

表 20 專家標示不相似群聚範例

標籤	體育 棒球 物價 政治 經濟 教育	國際觀 經營 降稅 國防 稅改 外交
標記網頁	-紅襪主場 Fenway Park 的 Ted Williams Seat -棒球小教室-投手犯規 @ 編織年少棒球夢 ::PIXNET 痞客邦:: -中時部落格-熱天午後-達賴喇嘛教馬英九的事 -郭泓志中繼3.2局飆8K・完全宰制大	-「台灣稅制是不公不義」會計師：台灣富豪繳稅 不到收入1成 -孬種馬! 萬船齊發包圍釣魚台! -重申降低營所稅到15%!!! 堅決減稅!!!

	都會！ -陳添枝：6%經濟成長率 4年內達成 -民進黨還有一點氣息! 蔡英文當選民進黨主席 -贊成彈劾陳水扁 還全民公道(發動連署) -【議題】典範之死！誰謀殺了童玩藝術節？！	

以Yes表符合使用者需求之推薦，No表不符合需求之推薦，表 21為專家評估推薦結果：

表 21 專家評估推薦結果

		Expert A		
		No	**Yes**	**Total**
Expert B	**No**	53(21.2%)	22(8.8%)	75(30%)
	Yes	21(8.4%)	154(61.6%)	175(70%)
Total		74(29.6%)	176(70.4%)	250(100%)

Kappa=(Observed agreement-Chance agreement)/(1-Chance agreement)

Observed agreement=(53+154)/250=0.828

Chance agreement= $0.296 \times 0.3 + 0.704 \times 0.7 = 0.5816$

Kappa= $(0.828 - 0.5816) / (1 - 0.5816) = 0.589$

接著計算Kappa Statistics進行同意度分析，並計算標準差為0.828，當信賴水準達95%時，信賴區間為(0.479,0.699)，對照表 4 Kappa參考對照表，同意度介於Moderate和Substantial。表示專家A和專家B對於群和群之間的相似概念同意度介於中間值以上。

專家A和專家B具有一致性意見的結果共207筆，再以這207筆配對組合當作樣本，計算本文的正確率。將專家標示為相似概念的群和本研究判別為相似的群做比對。結果得表 22。

表 22 專家標示兩兩群相似度結果

		專家標示	
		Y	**N**
分群結果	**Y**	130	37
	N	13	27
總數		143	64

經計算運算後，$sensitivity = 130/143 = 0.909$；$specificity = 27/64 = 0.422$；$accuracy = 0.909 \times \frac{143}{207} + 0.422 \times \frac{64}{207} \cong 0.758$。

4.4 討論與分析

本研究的分群結果中，有些群聚十分匪夷所思，探究其可能原因涵蓋了多面向，以下就幾個案例探討之：

I. 案例一：因重複標籤而被分在同一群

圖 28為某時間區間，性/別人權、全球化、女性、奧運、kmt…等隸屬同一群。

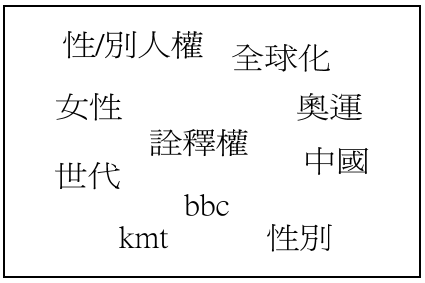


圖 28 案例一分群結果

仔細檢視與這一群聚最相關的網頁，如表 23，若僅觀察網頁標題，尚無法理解北京奧運和女性間的關聯性。然而，觀察使用者使用的標籤，發現原來二者間的共通點都是「bbc」，原來這三個網頁都來自於bbc新聞。從圖 29可以清楚看到標籤bbc將奧運、女性、中國…等標籤串接起來。

表 23 案例一相關網頁和使用標籤

網頁標題	使用標籤
北京奧運將首次准許運動員寫博客2008年02月16日	bbc 奧運 中國 blog 公民新聞 言論自由 新媒體 數位媒體 網際網路
BBC 中文網 中國報導 人權觀察年度報告批評中國人權狀況2008年01月31日	bbc 人權 奧運 中國 社會關懷 全球化
台灣來鴻：馬英九的女人們2008年02月21日	bbc 觀點 女性 選舉 馬英九 世代 性別 kmt 性／別人權 詮釋權

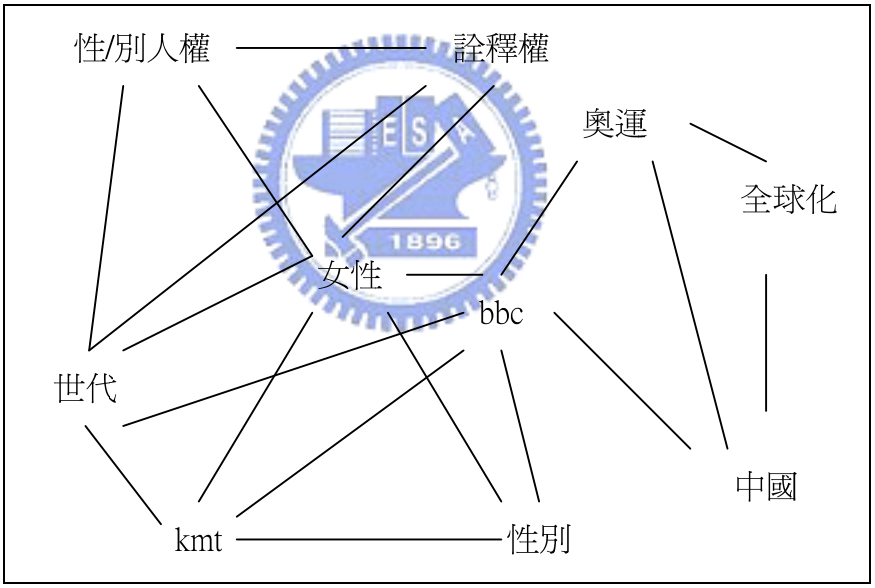


圖 29 案例一分群結果連結

受使用者在主題差別甚大之網頁中標記相同標籤的影響，使得在計算相似度時，這些不是很相關的網頁有了某種程度的關聯；這種情況發生在使用者想辨別所標記的書籤是何種類型，例如：是一則新聞或一篇部落格文章；或定義較模糊不清的標籤上，例如：地名、心情類的標籤。因此在進行分群時，容易使得原本

分離的兩群，合併在一起。容易造成分群結果整體而言不是很相關，但細看卻又似乎有點關連性。當要評斷分群結果良窳時，會不知怎麼拿捏這樣的分群結果。

但上述分群結果也跟分群時使用的距離演算法有關，因本研究使用的是平均連結聚合方式，為計算不同群聚間各點與各點間距離總和的平均。方法不夠嚴謹，使得有些點和點間彼此是沒有關連的，以表 1 表 24 標籤奧運與標籤女性及標籤 kmt 彼此之間的相似度為零，而有些卻有高度關連，標籤 bbc 與標籤奧運及標籤女性、標籤 kmt，彼此平均後，造成整體而言經過計算後是有相關連進而分在同一群聚中。假設換用完整連結聚合方式計算群聚間的距離為不同群聚中最遠兩點間的距離，以這例子而言，標籤女性與標籤奧運就不會分在同一群聚，因為不同群聚中最遠的兩點相似度為零。

表 24 案例一標籤相似度

標籤	標籤	相似度
bbc	奧運	0.107142857142857
bbc	女性	0.0714285714285714
bbc	kmt	0.0714285714285714
奧運	女性	0
奧運	kmt	0

在上述的例子中，也發現標籤blog、公民新聞、言論自由、新媒體、 數位媒體、網際網路、 社會關懷、選舉、馬英九等標籤不屬於這一群聚。這牽涉到這時間區間其他標記的網頁也重複使用這些標籤，使得這些標籤和其他標籤相似度較高。且本研究使用階層式分群法，在進行分群時相似度高的標籤組合就會先合併，等輪到和此範例的群聚時，相似度已沒有當初來得相關，故無法分在同一群聚。

II. 案例二：標籤間沒有重複卻被分在同一群

有些分群結果會發現，被標記的網頁間並沒有相關聯的標籤，以圖 30 的標籤為例：紅襪隊、四川地震，以及涼麵分在同一群聚，然而檢視這些標籤所標記的網頁(如表 25)，卻未發現有重複的標籤。

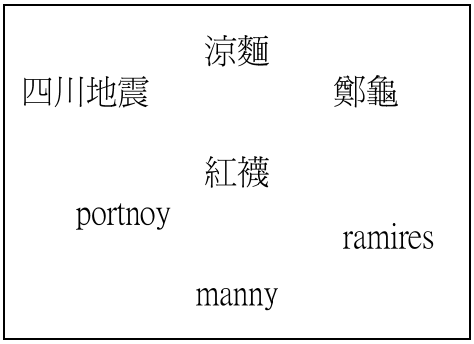


圖 30 案例二分群結果

表 25 案例二標籤標記網頁

網頁標題	使用標籤
I'm Vlog-涼麵不是簡簡單單就可以吃的	portnoy 涼麵 鄭龜
I'm Vlog-Manny Ramirez 耍寶集	manny ramirez 紅襪
I'm Vlog-【失敗的教育】遼寧少女痛罵地震災民〈繁體字幕〉	四川地震

再看標籤間的相似度表 26，卻發現兩兩標籤間的相似度竟都大於零。

表 26 案例二標籤相似度

標籤	標籤	相似度
四川地震	manny	0.0366
portnoy	涼麵	0.0357
manny	portnoy	0.0251
紅襪	涼麵	0.0251

因這幾個標記的網頁都來自同一個網站─I'm Vlog。I'm Vlog是一個影音網站，當在蒐集此網頁資料時，影音內的資料是無法被收集起來的，只能收集到該影音

檔的標題或使用者的敘述，以及其網頁架構的資料。這些網頁架構的資料幾乎不會隨著網頁內容的不同而有所變更，因此在產生標籤的向量時，就會有相同的詞彙，即使使用者使用的標籤不同，也會因而產生關聯，造成案例二這種分群結果。

換個方式來看，此現象可延伸到，假設有概念相同的網頁，可能都是在敘述同一主題，則一定會頻繁地使用某些關鍵詞彙，因此即使使用者下的標籤可能是跨語言或跨領域...等，但標籤所產生的維度一定會有所重複，使得它們有機會被分在同一群。如此可以看到來自四面八方不同性質的使用者，利用社會性標籤產生的社會互動。

第五章　結論與建議

本章總結本論文，說明運用時間序列分群的概念社會性標籤的效益，以及未來可改進的地方。

5.1 結論

本研究之主要目的在於利用社會性標籤及其標記的網頁內容，藉由時間序列分群，發掘出主題概念，並檢視不同時間區間社會性標籤的變化趨勢。本研究首先收集黑米共享書籤網站裡的標籤，根據其所標記之網頁內容轉換為時間序列的形式，找出在同一時間區間擁有相似走勢的標籤群聚，形成主題概念；接著計算不同時間區間所形成之群聚的相似度，以萃取出所有時間區間中擁有相似主題概念的群聚和包含於內的標籤；此外，對於同一標籤，分析其在各時間區間的變化趨勢，以及相關連的標籤和網頁。最後，透過開發之雛型介面將前述研究成果整合。經實驗與分析，本研究的結果整理如下：

一、　本研究透過時間序列化的標籤，利用餘弦相似度比對和階層式分群演算法找出在特定時間區間擁有相同走勢的標籤組合。在蒐集2008年HEMiDEMi中的部分資料，共計有3,629個網頁，以兩個星期為一單位，將一年分為26個時間區間，使用2707個不重複標籤，共產生1225群。

二、　本研究利用計算群的內聚力、分離度和整體分群品質來評斷分群的品質，經實驗結果發現，使用時間序列與否的分群結果在前述三項指標上並無太大差異。

三、　本研究將2008年以兩個星期為一單位分成26個時間區間，藉由視覺化的介面，讓使用者檢視同一標籤在不同時間區間的主題概念變化情形，如4.2.2的電影標籤之變化趨勢。

四、　本研究利用餘弦相似度計算不同時間區間中所構成之標籤群聚的

相似度，推薦擁有相似主題概念的標籤及相關網頁給使用者，藉著這些鏈結的資訊，提供使用者延伸閱讀的參考。本研究採用Kappa Statistics評估專家對推薦結果的同意度，同時考慮可信度及有效性。實驗結果推薦之準確率為0.758，仍有進一步改善的空間。

5.2後續研究建議

本研究將時間序列分群法應用在社會性標籤上，期能看到社會性標籤隨時間的變動趨勢。經過實驗結果的討論與分析，本研究仍有值得進一步改善與探究之處，茲闡述如下：

I. 分群演算法

使用其他群聚距離計算方式，能降低因受到單一重複標籤而被分在同一群的機率。

II. 標記網頁之內容資料

影響標籤分群結果最重要的為標籤間的關聯度，關聯度的來源為標籤各維度所代表之關鍵字的權重，而此權重乃是由標籤所標記之網頁的內容所產生。本研究中所蒐集之網頁內容五花八門，較多屬於跟日常生活相關的文章。由於每個網頁之架構都不盡相同，擷取網頁內容資料時，要正確地知道標題、內容的位置，並不容易。本研究是先統計擷取結果的TF和DF再加上LLR取前50名的詞彙代表一份網頁，但這樣做不盡然完美。以影評網頁為例，假設使用者想收藏的是「金鋼狼」這部電影的影評，但因為在網頁中可能包含了其他近期電影的資訊(如標題等)，在蒐集資料時，這些其他近期電影的資訊，就容易被拿來代表該網頁。因此，如果在擷取使用者標記的網頁時，可以準確地判斷標題、內容的位置，在後續步驟中將能夠提高準確度。

III. 標籤一致性

由於標籤是個人自由的創作，在缺乏訓練和規範的情境下，較不容易達到用法的一致性，例如：是Web20還是Web2.0、消費卷還是消費券；以及中

英文對照的問題，例如：是pixnet還是痞客邦。若能產生權威控制(Authority Control)資料庫或知識本體(Ontology)來表示同一概念的不同用詞，便可增加後續分群結果的準確性。

IV. 書籤分類

本研究在蒐集黑米書籤時，並沒有進行如圖 31的書籤分類動作。若能在一開始蒐集資料時，就先將書籤分類，則在後續分群結果上，可以更精確地觀看不同類別隨時間變化的標籤應用趨勢。例如：在運動類別裡，就可以看不同主題(籃球、棒球、網球)在不同時間區間的相關標籤。

圖 31 HEMiDEMi書籤分類

參考資料

[1] Agrawal, R., Lin, K. I., Sawhney, H. S., and Shim, K., "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," *In Proc. the 21st Int'l Conf. on Very Large Data Bases, Zurich, Switzerland,* pp. 490-501, Sept. 1995.

[2] A. K. Jain, M. N. Murty, & P. J. Flynn, "Data clustering: A review, " *ACM Computing Surveys*, vol. 31, pp. 264-323, 1999.

[3] B. Bollobás, G. Das, D. Gunopulos and H. Mannila, "Time-series similarity problems and well-separated geometric sets," in *SCG '97: Proceedings of the Thirteenth Annual Symposium on Computational Geometry,* 1997, pp. 454-456.

[4] D. Goldin and P. Kanellakis, "On similarity queries for time-series data: Constraint specification and implementation," *Principles and Practice of Constraint Programming — CP '95,* pp. 137-153, 1995.

[5] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," Stanford InfoLab, 1997.

[6] E. M. Voorhees, "Implementing agglomerative hierarchical clustering algorithms for use in document retrieval," *Information Processing & Management, vol.* 22, pp. 465-476, 1986.

[7] T. Hammond, T. Hannay, B. Lund, and J. Scott, "Social bookmarking tool (1) – A general review," D-Lib Magazine, vol.11, 4, Retrieved June 1, 2009, from http://www.dlib.org/dlib/april05/hammond/04hammond.h

[8] Han J., Kamber M., " Data Mining: Concepts and Techniques, " Morgan Kaufmann, San Francisco. pp. 346–389, 2001.

[9] Hsi-Cheng Chang and Chiun-Chieh Hsu, "Using topic keyword clusters for automatic document clustering," *Information Technology and Applications, 2005.*

ICITA 2005. Third International Conference on, vol. 1, pp. 419-424 vol.1, 2005.

[10] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis,* vol. 11, pp. 561-580, 01/01. 2007.

[11] Smith Gene, "Tagging: Emerging Trends, " http://www.asis.org/Bulletin/Aug-08/AugSep08_Smith.html , August/September 2008

[12] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," *Information Visualization, 1999. (Info Vis '99) Proceedings. 1999 IEEE Symposium on,* pp. 4-9, 140, 1999.

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations, " in *Proc. 5th Berkeley Symp.*, vol. 1, 1967, pp. 281-297.

[14] L. Kauffman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* J. Wiley & Sons, 1990.

[15] T. O'Reilly, " What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *COMMUNICATIONS & STRATEGIES,* no. 65, p. 17, 1st quarter 2007.

[16] T. W. Wal, "folksonomy,"*Online Information Conference* 2005 in London, Retrieved June 3, 2009, from http://vanderwal.net/essays/051130/folksonomy.pdf

[17] Warren, "Clustering of time series data--a survey," *Pattern Recognition,* vol. 38, pp. 1857-1874, 11. 2005.

[18] Wu, Ji He, Ah-hwee Tan,Chew-lim Tan,W., "On Quantitative Evaluation of Clustering Systems," 09/04. 2002.

[19] Z. Xu, Y. Fu, J. Mao and D. Su, "Towards the semantic web: Collaborative tag suggestions," in *WWW2006: Proceedings of the Collaborative Web Tagging*

Workshop, 2006.

[20] Kappa Statistics - http://www.dmi.columbia.edu/homepages/chuangj/kappa

[21]Orcale Text Reference-
http://download.oracle.com/docs/cd/B19306_01/text.102/b14218/astopsup.htm#sthref2545

[22] Wikipedia, "Time series", Retrieved June 7, 2009, from
http://en.wikipedia.org/wiki/Time_series

[23] 卜小蝶, "淺談社會性標記之意涵與應用",論文發表於淡江大學圖書館舉辦之「Web 2.0 與圖書館」研討會,臺北市,2006年12 月。

[24] 卜小蝶 , "使用者導向之網路資源組織與檢索",2007年.

[25] 中文斷詞系統簡介說明,http://ckipsvr.iis.sinica.edu.tw/

[26] 中央研究院平衡語料庫詞集及詞頻統計,
http://www.aclclp.org.tw/doc/wlawf_abstract.pdf

[27] 黑米共享書籤HEMiDEMi,http://www.HEMiDEMi.com/home.

[28] 資策會, "2008年12月底止台灣上網人口", Retrieved June 3,2009, from
http://www.find.org.tw/find/home.aspx?page=many&id=219

[29] 陳建誌 , "Web 3.0 時代來臨　是好是壞？",電子商務時報, Retrieved June 3, 2009, from http://www.ectimes.org.tw/shownews.aspx?id=081012153724 , 2008年

[30] 鄧兆旻, "Social Tagging火紅新網路商機逐漸成形",數位時代，Retrieved June 3, 2009, from http://www.bnext.com.tw/LocalityView_7648 , 2006年。

附錄一、中研院平衡語料庫詞類標記集

簡化標	對應的CKIP詞類標記[1]	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	*Daa*	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh,	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/

[1] 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE	/*的, 之, 得, 地*/	
SHI	/*是*/	
FW	/*外文標記*/