Intelligent Really Simple Syndication System Using Document Clustering Technology

凌士偉　邊國維

Department of Information Management, Huafan University, Taiwan

E-mail: b9110131@cat.hfu.edu.tw, gwbian@cc.hfu.edu.tw

摘要

由於網際網路的普及與個人出版網站的快速成長，大量的網路資訊進行傳遞與分享，包括: 新聞內容、網路日誌、個人相簿、網路留言版。 網站的管理者透過 Really Simple Syndication (RSS) 機制，利用 XML (eXtensible Markup Language) 所制定的格式，把最新網頁內容傳給訂閱戶，將豐富資料迅速地進行交換，並賦予使用者更大的自訂與整合能力。

使用者藉由 RSS Reader 匯入(訂閱)多個網站的 RSS 資訊，由於資料快速累積，使用者無法有效地從 RSS Reader 內閱讀與整理相關的資料，本論文整合文件分群技術，增強 RSS Reader 的功能，解決數位化資訊的組織與過濾問題。

關鍵詞：RSS, 文件分群

1. 前言

數位化時代的來臨與網際網路的快速發展，促進網路媒體的發展，各新聞網站的新聞不但數量大，且更新速度越來越快，如何提供即時且有效的資訊交換與傳遞，已成為重要的研究課題。除此之外，由於部落格(Blog)具有簡單的出版界面，提供一般使用者網路日誌、個人相簿、網路留言版的功能，部落格快速成長與流行，國內網站如: 天空部落、無名小站，不但產出大量的網路內容，更吸引與聚集大量的使用者。

一般而言，使用者必須要連線到網站後，才知道這些網站是否有更新並閱讀最新資訊，由於每個網站內容更新的頻率差異很大，如：一小時、數小時、每日、數日、每周，這樣的方式並不是一個有效率的方式。傳統上，有些網站使用電子報訂閱方式，通知使用者最新資訊。有些討論區網站與新聞群組網站，也使用訂閱方式，通知使用者每篇討論文章。

近年來，越來越多的網站管理者利用RSS，將最新資訊與網頁內容傳送給使用者。RSS是透過XML(eXtensible Markup Language)所制定的格式，網站的管理者可以將豐富的網頁內容傳給訂閱戶，概念上像電子報和新聞群組的整合，但是賦予使用者更大的自訂和整合能力[4,5]，RSS成為改變網路出版的新興技術，適合新聞性網站與個人Blog使用。

RSS Reader 提供使用者匯入(訂閱)多個網站的 RSS 資訊，目前有相當多的免費軟體，知名的有 MXIE（圖 1）[10]、RSS Bandit、Sharp Reader、KlipFolio 等。使用者能夠利用這些 RSS 資料，讀取相關網站的最新內容，由於資料快速累積，要從 RSS Reader 內閱讀與整理相關的網站資訊，變得相當困難。

圖 1　MXIE RSS Reader

本論文介紹利用中文自然語言處理(Chinese Natural Language Processing)與文件分群(Document Clustering)技術，建立一套具文件分析功能的 RSS Reader，解決大量數位化資訊的組織與過濾問題。第二節描述此系統的架構，第三節介紹 RSS 機制，第四節討論所使用的中文處理與文件分群技術，第五節將此具文件分析功能的 RSS Reader 運用於新聞網站，並展示實驗結果，第六節探討未來的發展。

2. 具文件分析功能的RSS Reader

當資料來源一個以上時，資料重複是不可避免的問題，例如多個新聞網站的電子報，同一新聞事件會有很多新聞網站報導，甚至同一新聞網站也有多篇報導。利用文件分群技術，可自動將多篇探討同一主題或事件的文件整理在一起，也可以過濾

重複的資料。

例如 GOOGLE News [7]自動收集各新聞網站的資料，將相關的新聞自動分群，相關的新聞在同一個頁面內呈現，如圖 2 所示。

圖 2　GOOGLE News 分群結果

本系統首先提供 RSS Feed 的機制，供使用者訂閱有興趣的網站，利用 RSS 資料，蒐集與彙整各網站的資訊，再將 RSS 資料經中文斷詞系統(Chinese Segmentation System)與詞性標記系統(Chinese Part-of-Speech Tagger)處理，利用文件分群模組，將近似的 RSS 資料整理為一群，自動將相關性的文章與新聞事件歸類，解決使用者閱讀大量資訊的組織與過濾問題，系統架構如圖 3 所示。

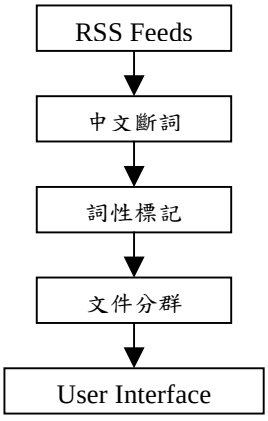

圖 3　系統架構

3.　RSS Feed 機制

RSS 的發展過程雖然具有相同的英文縮寫，但是其英文名稱與規格卻完全不同[12]，分別是 RDF Site Summary(RDF 網站摘要，版本為 RSS 0.90 與 RSS 1.0)、Rich Site Summary(豐富網站摘要，RSS 0.9x 版)與 Really Simple Syndication(非常簡易的資訊聯合， RSS 2.0 版)。屬於一種輕量級的 XML 技術，主要應用在各聯盟網站間，分享其新聞標題及其他內容提供服務網站的整合用途。透過嵌入 RSS 的資料，可以同步整合來自不同網站的資訊於單一網頁上，具有入口網站服務與個人化網頁(personalized web page)的功能。然而，RSS 並不只限於一般網站服務上，因為它的資料格式具有簡易解析及標準性，近年來也常應用於企業內部網站(Intranet)發佈最新消息與知識分享的傳遞媒介，行動服務業者也在簡訊服務(Short Message Service, SMS)中應用 RSS 為其交換標準格式，提供手機選單(Menu)個人化的行銷服務。

RSS 規格最早於 1999 年由 Netscape 公司發展出來，當時 RSS 的三個英文縮寫字母分別代表 RDF Site Summary (RSS 0.90) ，被應用於 My Netscape Network 入口網站，提供個人化設定，將其他網站的焦點標題與資訊連結呈現在單一頁面上。My Netscape 網站的這項功能受到網友們的歡迎，因為使用者可以透過簡單的設定來規劃與設定自己頁面版塊，將自己所關心的議題，例如氣象、交通等資訊整合於一頁，減少網站間資訊瀏覽與搜尋的時間成本。My Netscape 網站藉此不僅增加人氣，並從線上廣告中獲益，同時得到大量網路內容供應商(Content Provider)的支持與連結，因為 My Netscape 的 RSS 聯合機制讓它們的網站增加更多的使用者。

隨著 My Netscape 網站的成功經驗，RSS 逐漸受到重視，不過在 RSS 規格發展上卻發生分歧。由於 RSS 0.90 所使用的 RDF (Resource Description Framework)語法與資源描述架構過於複雜，有些發展者認為簡單的規格定義對於網站間的內容聯合相當重要，因此 UserLand Software 公司不採用 RDF 架構，從其所自訂的 Web 資料交換格式 Scripting News 中衍生出 RSS 0.91 規格，此版本的 RSS 縮寫改為 Rich Site Summary，所強調的是 RSS 在應用上的便利與簡單性，該公司亦提供許多 RSS 的相關應用軟體來推廣這樣簡單架構的版本。另一方面，My Netscape 入口服務網站因經營策略因素而關閉，具有 RDF 資源描述特性的 RSS 0.90 規格轉由非營利網際社群 RSS-DEV 工作小組來維護，並於 2000 年推出 RSS (RDF Site Summary) 1.0 規格，除了改善原 RSS 0.90 規格架構外，更推出具有延伸性的 RDF 模組，來加強內容聯合的應用與自訂性。堅持簡單架構發展路線的 UserLand 公司也於 2002 年推出 RSS (Really Simple Syndication) 2.0 規格，這一版本除了相容原有 RSS 0.9x 版本文件外，亦提供非 RDF 架構的延伸模組與支援 XML Namespace。從此區分出是否具有 RDF 架構的兩種 RSS 規格，不過 RSS 的應用卻廣泛運用在網際網路上，例如許多個人網誌系統（Blog）都預設使用 RSS 技術來進行站台文章的內容引用與聯合，並透過自動化程式同時提供不同 RSS 版本的匯入器（Feeds），RSS Feed 是一個 XML 特殊格式編碼後的網頁，提供給使用者

訂閱使用，網友可使用具有分析 Feed 編碼的 RSS Reader 軟體閱讀資料。

RSS Feed 資料中，有 description 的機制，系統可利用 description 內的資料代表每個項目的摘要資料，單純分析此資料可做出分群；或利用 description 內的連結，捉取完整文章資料，但是抓回來的資料，可能含有非屬於文章的其他文字，這將會造成雜訊，圖 4 為一個 RSS Feed 資料範例。

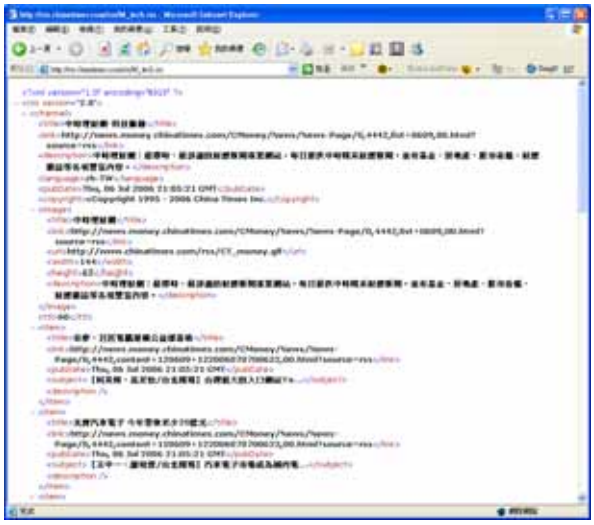

圖4　RSS Feed 資料範例

4. 中文處理與文件分群技術

文件分群的技術可以將文件集合依其相關程度分為許多群組，使用者透過分群的結果，可以快速地找出相關的文件。一般而言，分群的技術分為階層式分群法與非階層式分群法[11]。Jardine & Rijsbergen [9]利用單一連結法(single linkage method)作為文件分群的方法。Griffiths, Luckhurst, and Willet [8] 比較單一連結法之外、完整連結法(complete linkage method)、平均連結法(average linkage method)、與沃德法(Ward's method)，結果顯示單一連結法最差，而沃德法與完整連結法較好。

Zamir, Etzioni, Madani, and Karp 的研究[13]也指出完整連結法的計算複雜度較高($O(n^3)$)，分群的結果也最好；並且分別以單字和斷詞做分群比較，結果顯示以斷詞做分群有較好的結果。黃聖傑[3]透過斷詞處理與詞性分析，利用名詞與動詞來計算文件之間的相似度，其實驗結果的新聞分群準確率可達 90%。洪鵬翔[1]利用計算字串相似度的方式計算新聞標題之間的相似度，再以「階層式聚合演算法」加以分群，結果顯示只利用新聞標題分群的精確率與召回率各約 60%。

4.1 中文斷詞與詞性標記

系統利用所訂閱的 RSS Feeds 收集各網站的資料，將每個資料項目的描述(Description)利用中文斷詞系統與詞性標記系統加以處理，此部分係透過中央研究所開發的 CKIP 中文斷詞系統[6]。

例如: 去年每名員工替公司創造 630 幾萬的產值，而公司只發給(VD) 二十幾萬年終獎金，實在不合理。

經過斷詞與詞性標記處理的結果：

去年(Nd) 每(Nes) 名(Nf) 員工(Na) 替(P) 公司(Nc) 創造(VC) 630(Neu) 幾萬(Neu) 的(DE) 產值(Na) ，(COMMACATEGORY) 而(Cbb) 公司(Nc) 只(Da) 發給(VD) 二十幾萬(Neu) 年終(Nd) 獎金(Na) ，(COMMACATEGORY) 實在(D) 不(D) 合理(VH) 。(PERIODCATEGORY)

4.2 文件分群

若兩篇文章或新聞文件描述的人、事、時、地、物越相近，則屬於同一個主題的機率越大；本系統以各文章的名詞(N)與動詞(V)為基礎，將每篇文章的名詞與動詞組合成一個向量(vector)，文件(D)以此向量表示為 **D = (N1, N2, N3, ..., V1, V2, V3, ...)** ，其中 Ni 和 Vj 分別代表名詞和動詞的特徵(features)。例如挑出以上範例內的動詞與名詞，變成以下結果：

員工(Na) 公司(Nc) 創造(VC) 產值(Na) 公司(Nc) 發給(VD) 獎金(Na) 合理(VH) 。

要將文章加以分群，必須判斷文章間的相似度(similarity)，兩篇文章 Di 和 Dj 的相似度定義如下：

$$SIM(Di,Dj) = \frac{|Di \cap Dj|}{\sqrt{|Di|} * \sqrt{|Dj|}}$$

有了任兩篇文章的相似度後，再使用完全鏈結分群演算法 (Complete-Link Clustering Algorithm)，把同一主題的文章分為相同的一群。翁鴻加[2]指出由於較長的標題通常有較多的資訊，本系統選擇擁有最多字元的文章標題，代表同一群的標題。

5. 系統實作

我們實作出一個具文件分析功能的 RSS Reader，由於系統需要快速地處理完所有的新聞並提供服務，目前本系統只採用新聞標題來計算相似度，並提供使用者設定文件分群相似度的分群門檻值（clustering threshold）。

使用者可以自行訂閱各網站的 RSS 資料，透過抓取 RSS Feeds 的子系統，將收集的資料加以分群，自動將相關的文章連結在一起，閱讀文件時也可以自動列出相關文章，我們以國內的新聞網站為例(中時電子報、東森電子報、聯合報等)，示範此系統的結果。

圖 5 為程式的起始畫面，在畫面左方為使用者設定的 RSS Feeds，若要再增加新的 Feed，只需按 F2，即會跳出輸入 RSS Feed 的視窗。當 Feed 有更新資料進來後，在該 Feed 後方顯示更新數量，點選該 Feed 後可在右上方看到所有更新資料的標題，如圖 6 與圖 7 所示。選取新聞標題後，所連結出來的資料顯示在程式的右下方，如圖 8 所示。

在新聞標題上按右鍵之後將可以選取顯示所有相關新聞的選項，點取之後便會跳出一視窗，顯示所有的相關新聞，如圖 9 所示。本系統也提供使用者設定文件分群相似度的門檻值，可在程式工具屬性的其他選項內調整其值，如圖 10 所示。

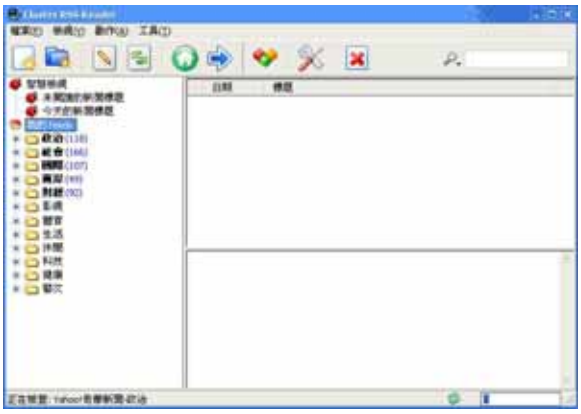

圖 5 起始畫面

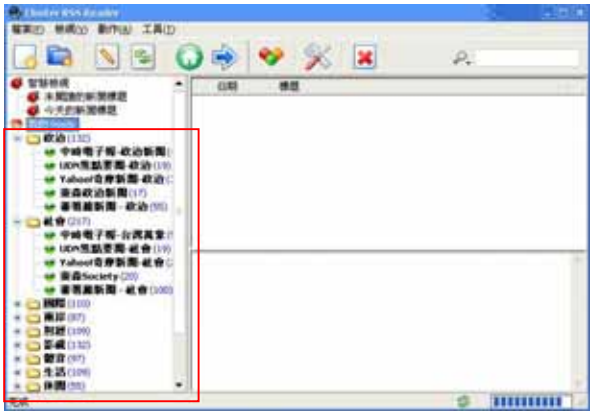

圖 6 Feed 更新數量

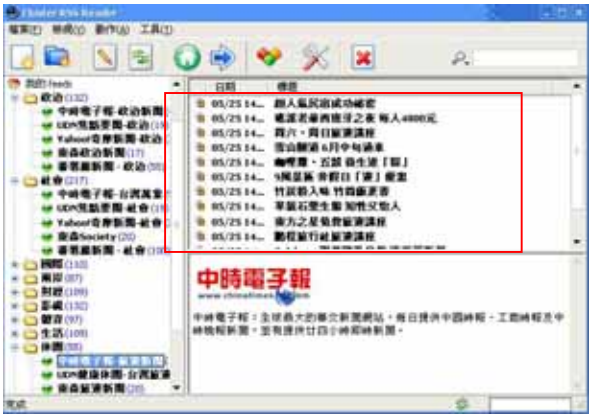

圖 7 更新資料的標題

圖 8 新聞資料

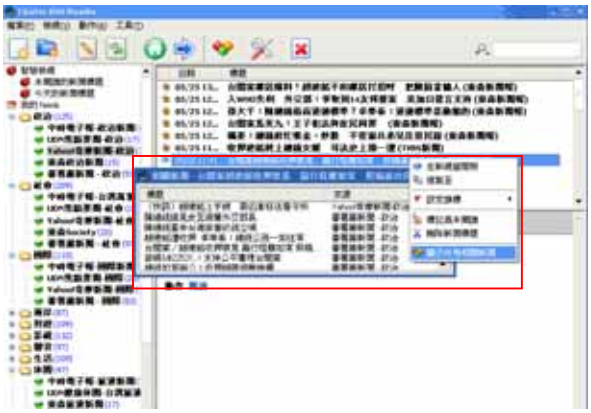

圖 9 相關新聞

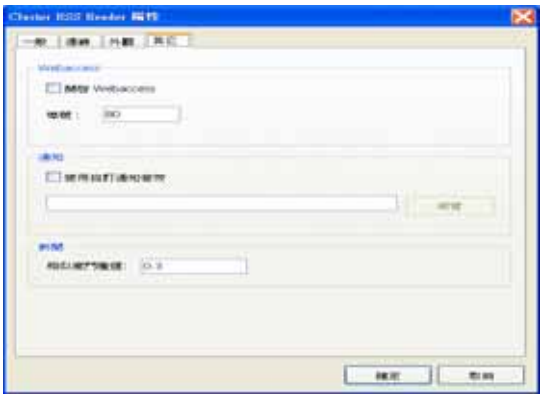

圖 10 文件分群相似度的門檻值屬性設定

6. 結論與未來發展

本論文整合中文自然語言處理與文件分群技術，建立一套具文件分析功能的 RSS Reader，增強目前 RSS Reader 的功能，使用者藉由 RSS Reader 匯入(訂閱)多個網站的 RSS 資訊，包括: 新聞內容、網路日誌、個人相簿、網路留言版。本系統自動將收集的資料加以分群，自動將相關的文章連結在一起，使用者閱讀文件時也可以自動列出相關文章，解決大量數位化資訊的組織與過濾問題。

未來將研究如何整合多國語言的資訊來源，將不同語言的資料但主題相關的文章加以分群，提供多國語言與跨語言檢索的功能。

參考文獻

[1] 洪鵬翔，中文新聞自動群聚，國立清華大學資訊工程研究所, 2000。

[2] 翁鴻加，多文件摘要一些新技術及評估模型之建立，國立臺灣大學資訊工程學研究所, 2001。

[3] 黃聖傑，多文件自動摘要方法研究，國立臺灣大學資訊工程學研究所, 1999。

[4] 歐坤宗，建構以 RSS 為基礎的新聞內容聯合機制-以《大學報》為例，國立台北大學資訊管理研究所, 2004。

[5] 謝良奇編譯，Web 內容串聯技術 RSS 標準之爭，自由軟體鑄造場計畫網站，九十二年八月十日, http://www.openfoundry.org/archives/000164.html。

原文：http://simon.incutio.com/archive/2003/06/25/moreOnRss

[6] CKIP，中文詞知識庫小組，http://rocling.iis.sinica.edu.tw/CKIP/

[7] Google NEWS, http://news.google.com.tw/。

[8] A. Griffith, H. C. Luckhurst, P. Willet, "Using Inter-Document Similarity Information in Document Retrieval Systems", Journal of the American Society for Information Science, Vol. 37, pp. 3-11, 1986.

[9] N. Jardine, and C. J. van Rijsbergen, "The Use of Hierarchical Clustering in Information Retrieval", Information Storage and Retrieval, Vol. 7, pp. 217-240, 1971.

[10] MXIE RSS, http://www.rss104.com。

[11] E. Rasmussen, "Clustering algorithms", Information Retrieval, Editors: W.B. Frakes and R. Baeza-Yates, pp. 419-442., Prentice Hall, Eaglewood Cliffs, N.J., 1992.

[12] Wiki: RSS, http://zh.wikipedia.org/wiki/RSS

[13] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp, "Fast and Intuitive Clustering of Web Documents", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp. 287-290, 1997.