四、實驗部份

4.1 系統實作概述

本章節描述所實作之系統的開發環境與各模組功能，並以此系統來實際分析幾個測試資料，以評估此演算法的區塊分割程度。

4.1.1 開發環境

本文的系統以微軟的Visual Studio 2005的C# 環境做為開發工具，而產生的資料則以XML格式來存放。當初會以微軟的工具來開發，主要就是為了能相容於實驗室所開發的網頁資料萃取系統BODE，故選擇同一平台進行開發，希望能以Plug-in的方式讓BODE使用本系統的功能為目標。

4.1.2 模組介紹

就功能面來說，系統可分為1. 人機介面、2.網頁蒐集、3.網頁區塊切割、4. 網頁區塊輸出檢視、5. 網頁區塊輸出統計等五個模組。以下分別就這幾個模組簡單介紹。

1. 人機介面

負責與使用者互動，提供使用者各項操作、設定的圖形化介面。

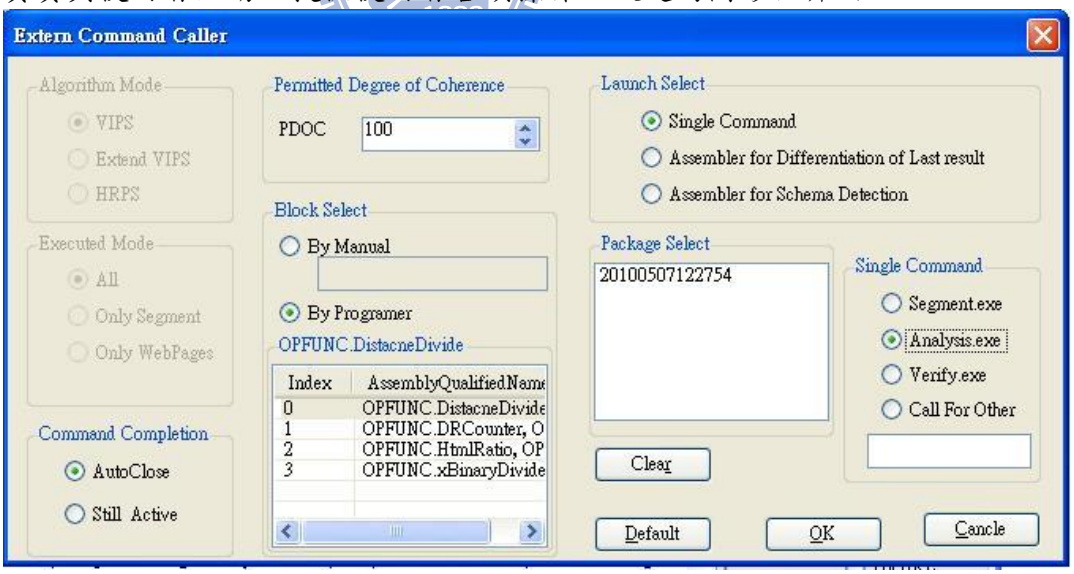


圖4.1 Command UI

2. 網頁蒐集

類似網頁爬行機器人(Web Crawler)，會從給定的網址所包含的超鏈結往下擷取網頁，但僅限於同一個Domain Name或是在抓取清單內與排除清單外的頁面。而爬行不限於網際網路，當然可將網址改為檔案位址，擷取完畢後，會將頁面送給網頁區塊切割模組進行區塊的切割。

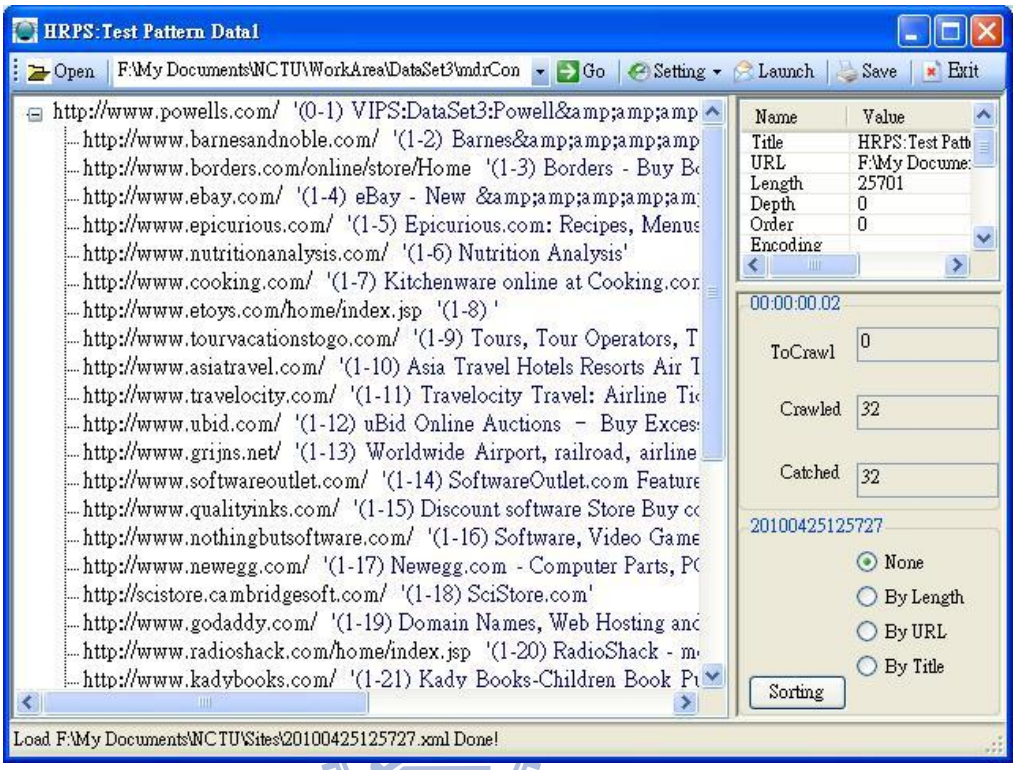


圖4.2 WebCrawler.exe

3. 網頁區塊切割

在每一個頁面擷取完畢後，會接手進行區塊的切割。會將過程中所產生的各項資料，包含最終的區塊資料以XML格式來存放。

圖4.3 WebSegment.exe

4. 區塊輸出檢視

得到網頁區塊後，我們還需要檢視工具，以確保區塊會依據設定的條件取出。

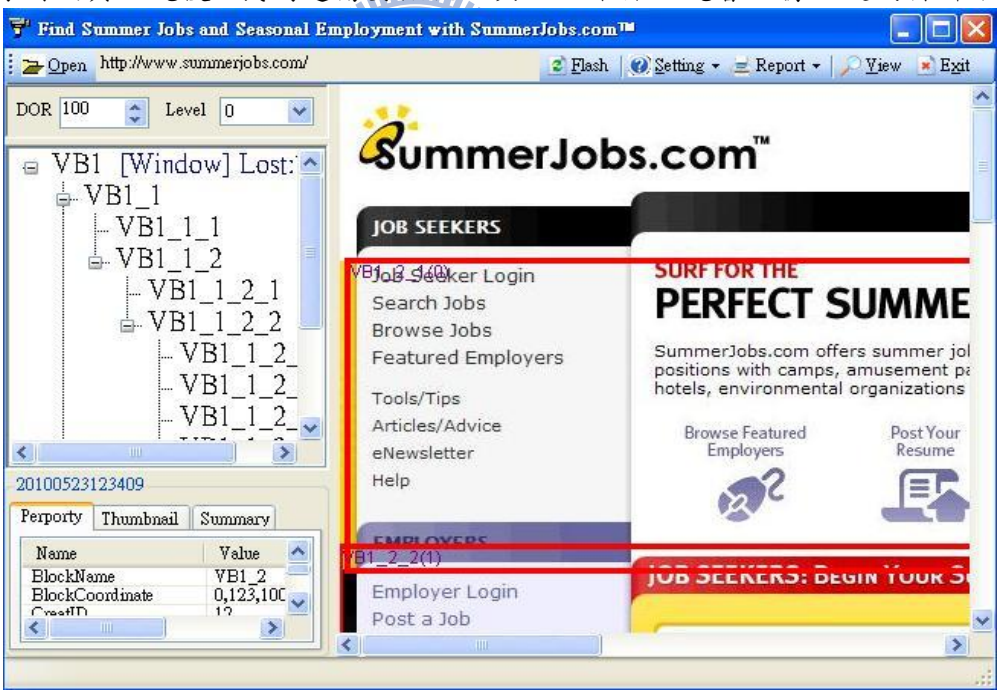


圖4.4 WebAnalyer.exe

5. 網頁區塊輸出統計

此模組會針對網頁區塊，進行統計。

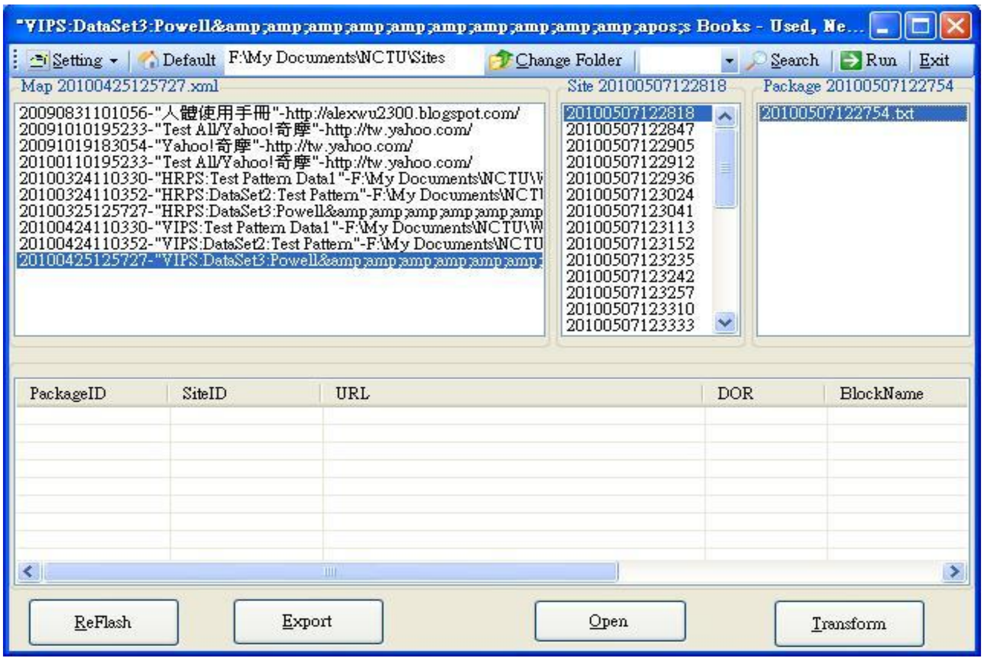


圖4.5 Web Verifyer.exe

4.2 實驗

看下節各圖例，代表我們所分析的例子其重疊區塊皆能解出，

4.2.1 範例

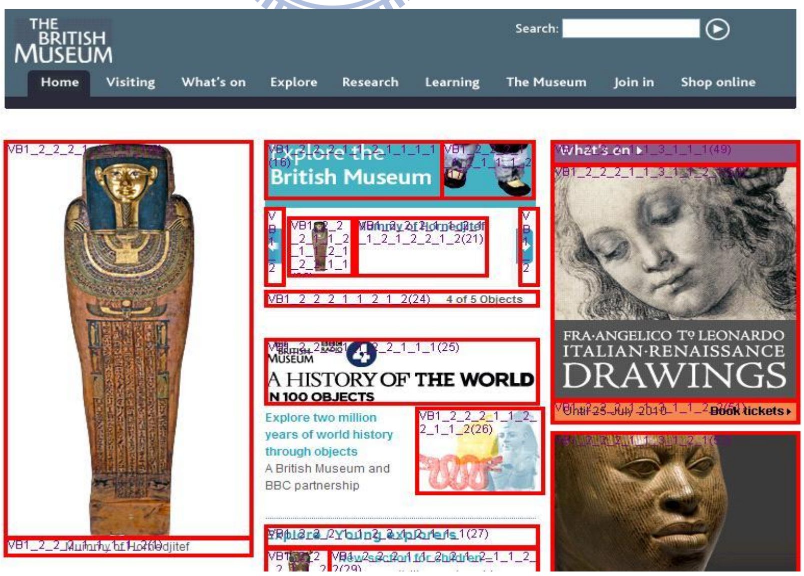


圖4.6 Figure 2.18的區塊分割

圖4.7　Figure 2.12的區塊分割

圖4.8　Figure 2.15的區塊分割

為了驗證我們所提出問題是一般化網頁現象並非特殊應用下的設計，我們選擇三個可公開下載的數據集，來比較我們的方法與VIPS演算法的差異。

4.2.2 資料來源

1. 第一個數據集（Data 1）是ViNTs[20] 使用的 DataSet 31。
   Download Address:
   http://www.data.binghamton.edu:8080/vints/testbed.html
   它是早期使用於自動化網頁擷取的測試資料庫。在使用Wrapper的早期頁面分割演算法大都採用此測試資料庫。
2. 第二個數據集（Data 2）是DEEP WEB[28]之TBDW版本。
   Download Address: http://daisen.cc.kyushu-u.ac.jp/TBDW
   TBDW擁有來自 51個搜索引擎的查詢結果，且每個搜索引擎，有五個查詢的結果頁。 在數據集 2，我們只收集各搜索引擎的第一個結果頁（1.html）為代表。
3. 第三個數據集（Data 3）是從MDR[12]文件中所列出具代表性的網站首頁。
   它是Data Mining的測試數據來源， 因為隨著網站的改變，其位址也有了變動，故列於表 4.1所示。

表4.1 Data Mining的測試來源表

Title	URL
Powell's Books - Used, New, and Out of Print - We Buy and Sell	http://www.powells.com/
Barnes & Noble-Books,Textbooks, eBooks, Toys, Games & More	http://www.barnesandnoble.com/
Borders -Buy Books, Used Books, Music, DVDs & Blu-ray Online	http://www.borders.com/online/store/Home
eBay - New & used electronics, cars, apparel, collectibles, sporting goods & more at low prices	http://www.ebay.com/
epicurious.com Recipes,Menus, Booking Articles & Food Guides	http://www.epicurious.com/
Nutrition Analysis	http://www.nutritionanalysis.com/
Cooking	http://www.cooking.com/
Etoys	http://www.etoys.com/home/index.jsp
Tours, Tour Operators, Tour	http://www.tourvacationstogo.com/

Packages, Escorted Tours	
Asia Travel Hotels Resorts Air Ticketing Tours Packages Reservation	http://www.asiatravel.com/
Travelocity	http://www.travelocity.com/
uBid Online Auctions – Buy Excess Inventory from the World's Most Trusted Brands!	http://www.ubid.com/
SoftwareOutlet.com Features A Huge Selection Of The Most Popular Software Titles At The Guaranteed Best Price	http://www.softwareoutlet.com/
Discount software Store Buy computer software at cheap software prices	http://www.qualityinks.com/
Software, Video Games, PC Games, Electronics, Accessories - NothingButSoftware.com	http://www.nothingbutsoftware.com/
Computer Parts, PC Components, Laptop Computers, LED LCD TV, Digital Cameras and more!	http://www.newegg.com/
SciStore.com	http://scistore.cambridgesoft.com/
Domain Names, Web Hosting and SSL Certificates - Go Daddy	http://www.godaddy.com/
mobile phones, MP3 players, laptops, and more	http://www.radioshack.com/home/index.jsp
Kady Books-Children Book	http://www.kadybooks.com/

Publisher & Childrens Bookstore	
Kids Athletic Shoes Kidsfootlocker.com	http://www.kidsfootlocker.com/
Online Shopping, Compare Products and Save - Lycos Shopping	http://shopping.lycos.com/
Laptops, Desktop Computers, Monitors, Printers & PC Accessories Dell	http://www.dell.com/
TVs, Computers, Cameras, GPS, Home Audio, Desktops, Laptops, Consumer Electronics, and More at CircuitCity.com	http://www.circuitcity.com/
Online Shopping - Bedding, Furniture, Electronics, Jewelry, Clothing & more	http://www.overstock.com/
Digital Cameras, Camera Accessories, Printers, Ink & more	http://www.kodak.com
Find Local Jobs & Employment Listings - FlipDog Job Search	http://www.flipdog.com/
Find Summer Jobs and Seasonal Employment	http://www.summerjobs.com/
Lycos Search	http://search.lycos.com
Northern Light Strategic Research Portals	http://www.northernlight.com/
Mamma.com The Mother of All Search Engines	http://www.mamma.com

4.2.3 實驗結果

我們所引用的網頁區塊切割演算法 VIPS[11] 並無公開其原始碼，而是根據馬維英博士團隊所發表的論文內容實作而來的，在這部份，因為並沒有修正的論文出現，我們是以自訂的規則處理，但這不影響之後的探討。在解釋實驗結果之前，先介紹為比較視覺區塊(Visual Cues，VCs)分割狀態，我們在輸出程式上額外增添的功能。在圖4.9 中 分割程式在輸出的樹狀結點上，會標示一個 Empty的字樣 ，告訴我們在此視覺區塊上是找不到視覺上的切割線(Empty Separator) ，所以由此標示可知其對應視覺區塊的大小與 DOM Tree上的結點資訊，而在此例上，分割程式是整頁無法分割。而在圖4.10 中 是 HRPS 的執行結果，我們把分割後的樹狀圖中，所有的葉節點的總合稱為視覺區塊總數 (Extracted Visual Cues，X-VCs) ，如果葉節點帶有 Empty 的標記，其總合稱為無法切割的視覺區塊總數(Empty Separator Visual Cues，ES-VCs) 。

我們實作了 VIPS 演算法來與 HRPS 比較，藉由網頁區塊切割程式的標記，經網頁區塊輸出統計程式整理後得到區塊切割失敗的統計資料，如表4.2所示。在表4.2第三行的 WebPages 是指測試資料的頁面總數，由頁面總數來看這三組測試資料相差不會太大，恰好適合看出彼此間差異處。第四行的Data Records由原來的測試資料記錄得出， 是指如果只計算本文(Text Area)頁面中區塊有多少個，可看到頁面總數與本文區塊總數比較皆在20多倍左右。第五行的Extracted Visual Cues (X-VCs)是由程式擷取的視覺區塊總數，可看到加上 Image，Form，Object 以後， 視覺區塊總數比只擷取本文區塊要大上很多，也看出頁面總數與視覺區塊總數固定比值的關係已打破，反應這三組測試資料是不同特性的網頁，可看出 HRPS 切割出的視覺區塊總數是大於 VIPS。第六行的Empty Separator Visual Cues (ES-VCs) 是頁面切割程式計錄下來的視覺區塊，可看出無法切割的視覺區塊並不多，列出第七行 ESVCs/XVCs比值，可看出在視覺區塊總數的比例上，最多到 4%。但把視覺區塊數量改為視覺區塊大小(第八九十行)，可看到 DataSet1 在 VIPS 下，無法切割的視覺區塊達總面積 28% ，DataSet2 在 VIPS 下，無法切割的視覺區塊達總面積 31% ，DataSet3 在 VIPS 下，無法切割的視覺區塊更高達總面積 43% ，所以 VIPS 遇到無法切割的區塊面積是佔相當大的比重，但反觀 HRPS 中 DataSet1 無法切割的視覺區塊達總面積 0.78% ， DataSet2 無法切割的視覺區塊達總面積 1.11% ，DataSet3 無法切割的視覺區塊達總面積 9.23% ，可看出切割的視覺區塊面積比重是大幅的降低，所以可證實HRPS對提高在區塊上找到視覺上切割線的機會確實有效。但 HRPS 中 DataSet3 的比值為 9.23%，是非常的大於其他兩組測試資料，經探討與觀察網站區塊狀態後，發現這組測試資料中，有為數不小的網頁設計，是重覆設計了某些HTML資料，所以頁面切割程式會標示為無法切割（重覆的HTML資料卻代表相同位置），經由人工篩選後此比率已大幅降低。所以整個來看，HRPS 可在保持視覺分割的基礎上，又大幅降低無法切割的區塊面積。

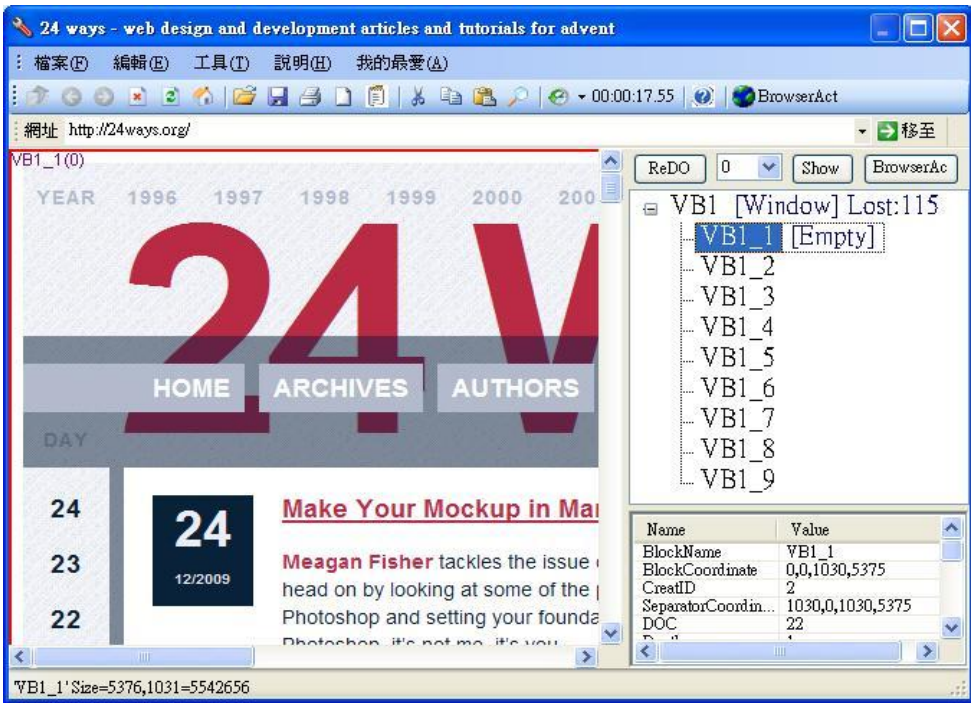


圖4.9 VIPS 區塊分割結果

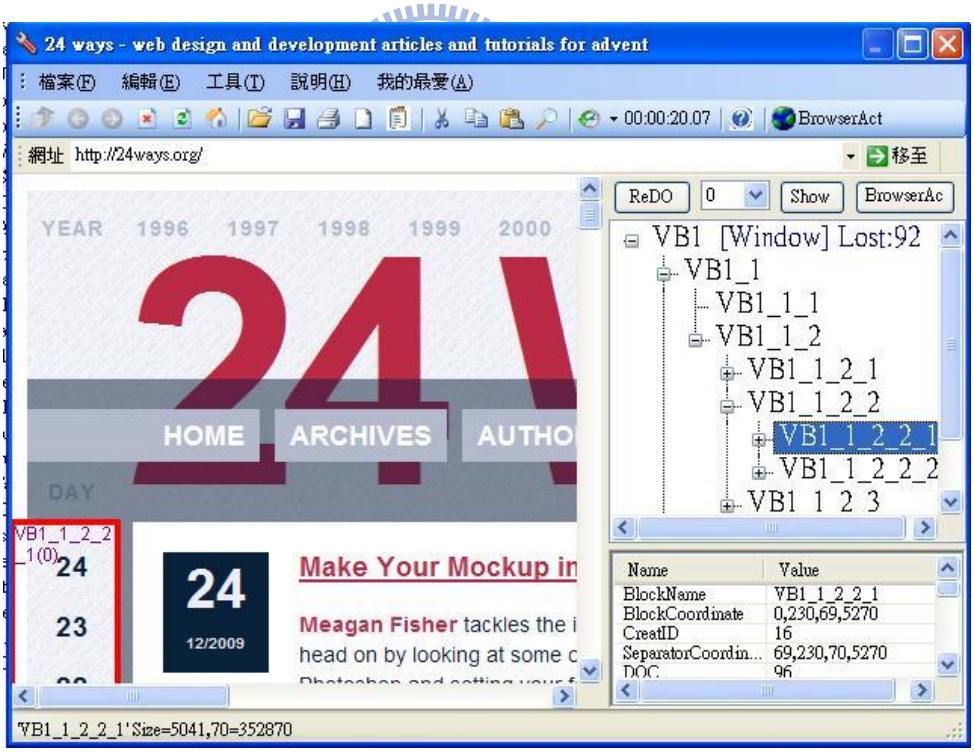
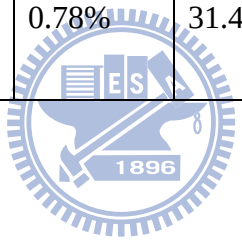


圖4.10 HRPS 區塊分割結果

表4.2 VIPS與HRPS測試結果評估表

	DataSet1		DataSet2		DataSet3	
Algorithm	VIPS	HRPS	VIPS	HRPS	VIPS	HRPS
Web Pages	41	41	48	48	31	31
Data Records	833	833	1004	1004	605	605
Extracted Visual Cues (XVCs)	5113	6316	3346	4390	2425	3764
Empty Separator VCs (ES-VCs)	71	75	110	99	88	18
ES-VCs/XVCs	1.38%	1.18%	3.28%	2.25%	3.62%	0.47%
Size of XVCs	57318495	48756794	63582504	57899345	49985922	30144870
Size of ES-VCs	16156791	380672	19981462	645507	21882462	278335
Size Ratio (ES-VCs/XVCs)	28.18%	0.78%	31.42%	1.11%	43.77%	9.23%



五、結論

5.1 總結

本文提出了適用於DHTML網頁文件的視覺化網頁分割演算法，並藉由實作的系統，來驗證理論是否正確。基本上網頁分割演算法，相當於在未標誌的空白網頁上打上可辨識的底稿格線， 如果格線太過粗糙，那會影響解析度，進而使後續動作失真，所以一個高解析度的網頁分割，將會使後續的動作有很大的正確性。這也是網頁分割演算法的最重要特點。接下來說明此演算法可應用的領域。

5.2 相關應用

1.搜尋引擎精確度提昇

被搜尋網頁中免不了夾雜著一些導覽或是廣告連結等不適用的資訊， 所以搜尋引擎要先行將每個網頁中，這些不重要的部份去除， 因為多了這層計算工作，勢必對效能有所影響， 故搜尋引擎在技術上有兩大關鍵，一是提升搜尋精確度，二是降低平均搜尋成本， 而這兩點又互為翹翹板的兩端， 因為提升搜尋精確度也會提高平均搜尋成本， 而降低平均搜尋成本可能會降低搜尋精確度，而高解析的網頁分割演算法可提高精確度， 且我們也可藉由高解析的網頁分割演算法，透過統計上的估算， 先行刪除網頁中不適用的資訊，即可儘量降低搜尋成本。

2.網頁過濾及更新追蹤

目前一般的網站內容都相當豐富，而且幾乎天天都會有更新，若網站沒有提供RSS的訂閱服務，使用者就必須要先花時間找尋網站中有興趣或重要的網頁，再來更是要自己定期追蹤是否有相關的更新發生，無疑又是一件費時的差事。因此我們可以運用此演算法，配合上特定的關鍵字協助使用者過濾出相關的重要內容，並做為基準，爾後定時再依同樣的條件對網站進行分析，藉由Diff演算法判斷出是否有更新，自動蒐集並呈現給使用者相關的資料，增加資訊利用的效率。如此，就算網站沒有提供RSS的服務，也可以達到類似的效果，並事先將資料下載至本機端，節省網路傳輸的等待時間。

5.3 未來工作

1.區塊邊界判斷

基本網頁分割演算法只是為網頁上打上可辨識的底稿格線， 並未對區塊大小與邊界提出定義， 所以對一些導覽或是廣告連結的區塊邊界， 還需要額外的演算法來判斷， 而在 HRPS 能完整反應其階層式區塊的基礎下，我們很容易加上以判斷區塊內容的方式，得到主要的區塊邊界，也就是 Page Layout 的偵測， 但同一個網頁 Page Layout，可能會有不同的組合方式來呈現， 所以可能要訴諸統計的方法，來決定其網頁類型。

2.資料路徑自動辨識

網頁資料的萃取，始終需要人力的介入，如撰寫所需的腳本或是設定所要擷取的資料區塊等動作，因而無法達到自動化，在實用上便有一定的限制。所以資料路徑的自動辨識仍是一個重要的問題，否則空有資料區塊，但沒有相對於首頁的路徑及區塊內容的資料項目化，對於網頁資料萃取系統的自動化，仍無太大的助益。

參考文獻

1. "Document Object Model-W3C Recommendation." Available: http:// www.w3.org/DOM.
2. "Google " Available: http://www.google.com
3. "The Internet Economy 25 Years After .com." Available: http://www.itif.org/publications/internet-economy-25-years-after-com
4. "W3C-Cascading Style Sheets." Available: http://www.w3.org/Style/CSS/
5. "W3C-The global structure of an HTML document." Available: http://www.w3.org/TR/REC-html40/struct/global.html
6. "Z-Index And The CSS Stack: Which Element Displays First?". Available: http://www.vanseodesign.com/css/css-stack-z-index
7. Arasu, A and Garcia-Molina, H, "Extracting structured data from web pages," pp. 337-348, 2003.
8. Baluja, S, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", p. 42, 2006.
9. Bjork, S, et al., "WEST: a Web browser for small terminals", pp. 187-196, 1999.
10. Cafarella, MJ, et al., "Web-scale extraction of structured data", ACM SIGMOD Record, vol. 37, pp. 55, 2009.
11. Cai, D, et al., "VIPS: a visionbased page segmentation algorithm," ed: Citeseer, 2003.
12. Chang, CH and Kuo, SC, "Olera: Semisupervised web-data extraction with visual support," IEEE Intelligent systems, vol. 19, pp. 56, 2004.
13. Chang, CH and Lui, SC, "IEPAD: information extraction based on pattern discovery",pp. 681-688, 2001.
14. Chen, J, et al., "Function-based object model towards website adaptation" , p. 596, 2001.
15. Cho, HGM, et al., "Extracting Semistructured Information from the Web" , pp. 18-25, 1997.
16. Embley, DW, et al., "Record-boundary discovery in Web

documents" , ACM SIGMOD Record, vol. 28, pp. 467, 1999.
17. Gupta, S, et al., "Automating content extraction of html documents", World Wide Web, vol. 8, pp. 179, 2005.
18. Gupta, S, et al., "DOM-based content extraction of HTML documents", p. 214, 2003.
19. He, B, et al., "Accessing the deep web", Communications of the ACM, vol. 50, p. 101, 2007.
20. Li, Longzhuang, et al., "Visual Segmentation-Based Data Record Extraction From Web Documents", pp. 502-507, 2007.
21. Liu, B, et al., "Mining data records in Web pages", pp. 601-606, 2003.
22. Liu, W, et al., "Vision-based Web Data Records Extraction" ,2006.
23. NANNO, T, et al., "Structuring Web Pages Based on Repetition of Elements" , Transactions, vol. 45, pp. 2157, 2004.
24. Reis, DC, et al., "Automatic web news extraction using tree edit distance," pp. 502-511, 2004.
25. Simon, K and Lausen, G, "ViPER: augmenting automatic information extraction with visual perceptions", pp. 381-388, 2005.
26. Wu, I, et al., "A Web Data Extraction Description Language and Its Implementation", pp. 293-298, 2005.
27. Yang, Y and Zhang, HJ, "HTML Page Analysis Based on Visual Cues" , pp. 859-864, 2001.
28. Zhai, Y and Liu, B, "Structured data extraction from the web based on partial tree alignment", IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 1614, 2006.
29. Zhai, Y and Liu, B, "Extracting web data using instance-based learning" , World Wide Web, vol. 10, pp. 113, 2007.
30. Zhao, H, et al., "Fully automatic wrapper generation for search engines", pp. 66-75, 2005.
31. Zhao, H, et al., "Automatic extraction of dynamic record sections from search engine result pages",pp. 989-1000, 2006.