國立交通大學

資訊學院 資訊學程

碩士論文

基於HTML文件佈局之網頁分割演算法

A HTML Rendering-Based Page Segmentation Algorithm (HRPS)

研 究 生：余提梵

指導教授：吳毅成 博士

中華民國九十九年六月

基於HTML文件佈局之網頁分割演算法

A HTML Rendering-Based Page Segmentation Algorithm (HRPS)

研　究　生：余提梵　　Student: Ti-fan Yu

指導教授：吳毅成　博士　　Advisor: Dr. I-Chen Wu

國立交通大學

資訊學院 資訊學程

碩士論文

A Thesis

Submitted to College of Computer Science

National Chiao-Tung University

in partial Fulfillment of the Requirements

for the Degree of Master of Science

in

Computer Science

June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

基於 HTML 文件佈局之網頁分割演算法

學生：余提梵　　指導教授：吳毅成 博士

國 立 交 通 大 學　資 訊 學 院　資 訊 學 程 碩 士 班

摘　　要

依據統計資料，截自 2010 年來共有 1.13 億個網站存在過，其中有 99.9%是在近 15 年間成立的，面對這樣龐大又高替換的網頁資料，如何有效地使用是一件很重要的事。對於大量變動態的資料，通常尋求搜尋引擎的協助來正確定位資料；對於已知位址的資料，為了增加使用效率，則會使用資料萃取的技術。而不管是搜尋引擎或資料萃取工具，要對複雜的網頁進行分析，首要就是要對網頁作區塊分類或標記，以濾除噪音（ Noise ）區塊及提取各主題(Topic)區域之本文區塊，也就是網頁區塊分割(Page Segmentation) 。

2003 年微軟團隊發表視覺化網頁分割演算法(Vision-based page segmentation: VIPS)後，很多網頁分割研究多參考了視覺化分割技術。但在近幾年來，越來越多網頁的頁面框架設計，採用 DHTML 技術為主時，原始的 VIPS 的方法在使用上，便出現當初設計時沒有顧及的小缺陷，雖然 之後的研究，出現很多組合型態的頁面分割演算法來彌補使用上的不足。但因為是採用其它特性的演算法來彌補 VIPS， 所以這部份切割區塊也就喪失視覺化分割的特性。

本文提出一個方法，在以視覺化分割為基礎上，帶入網頁文件布局特性(HTML Rendering-Based)，以解決視覺化區塊分割在 DHTML 網頁上，可能找不到視覺化分割線(Separator)的問題。

A HTML Rendering-Based Page Segmentation Algorithm (HRPS)

Student: Ti-fan Yu Advisor: Dr. I-Chen Wu

Degree Program of Computer Science
National Chiao Tung University

ABSTRACT

According to the statistical datas, Up to 2010, a total of 113 million websites existed，of which 99.9% was established nearly 15 years, the face of such large and high replacement page data， how to effectively use is a very important matter。 For the information that we don’t know its location，we usually use search engine to help us to find it out。And for the information that we do know where it is， we use data extraction to increase the efficiency。And whether it is a search engine or information extraction tool， to analyze the complex web，the first steps is to split the Web Page to provide subject area of this location，It’s a important thing that how to use this huge database efficiently。

Since 2003 the team released Microsoft Visual Web segmentation algorithm (Vision-based page segmentation: VIPS)， many papers are mostly used segmentation based on visual segmentation，However, in recent years，more and more web page Layout design，using DHTML technology-based, the original method of VIPS in the use， they are in the original design did not take into account small defects， though after the study， there are many page segmentation algorithm combined patterns to make up for the use of deficiency。

But since they are using other features of the algorithm to make up for VIPS, so this part of the Visual cues is losing the characteristics of visual segmentation，This paper presents a method， in order to split based on visualization， into the HTML document Rendering features， to solve the visual segmentation in DHTML pages， you may not find the visual Separator problems。

誌　　謝

首先要感謝我的指導教授，吳毅成博士，由於他的細心指導與協助，本論文才得以順利完成。

另外也要感謝陳隆彬學長、孫德中學長，總在每次的討論之中，給我正確的方向。再來就是實驗室BODE組的各位伙伴，有你們的加油打氣，我才能一路堅持至此，謝謝你們。

最後要感謝的就是我的爸媽，為了這篇論文，犧牲了很多陪伴他們的時間。他們不曾抱怨、始終要我以學業為重，並在我低潮時給我最大的支持與鼓勵，讓我可以安心的做研究、完成學業。謹以此論文，獻給我最愛的爸媽。

目 錄

表目錄

圖目錄

符 號 說 明

DoC ：Degree of Coherence

DoI ：Degree of Information