Content Extraction from HTML Documents

A. F. R. Rahman, H. Alam and R. Hartono
Document Analysis and Recognition Team (DART)
BCL Computers Inc.
990 Linden Drive, Suite # 203, Santa Clara, CA 95050, USA.
Tel: +1 408 557 5279, Fax: +1 408 249 4046, Email: fuad@bcl-computers.com

Abstract

In recent times, the way people access information from the web has undergone a transformation. The demand for information to be accessible from anywhere, anytime, has resulted in the introduction of Personal Digital Assistants (PDAs) and cellular phones that are able to browse the web and can be used to find information using wireless connections. However, the small display form factor of these portable devices greatly diminishes the rate at which these sites can be browsed. This shows the requirement of efficient algorithms to extract the content of web pages and build a faithful reproduction of the original pages with the important content intact.

1. Introduction

The problem of content extraction is very important not only from the point of view of managing the amount of content, but other important issues are associated with this. Some of which are:

- Viewing any website: Pattern recognition systems that use document analysis techniques can be employed for displaying web pages on small screen devices by extracting and summarizing their content. These systems have to be generic enough so that they can work with any web site, not only the well laid-out ones.
- High Speed Access: The transformation of web pages has to take place on the fly and therefore should be fast.
- Network Usage: The schemes employed for the transformation should not decrease network traffic.
- Easy Configurability: Any such scheme should be easily configurable within existing systems by System Integrators (SI) and end users.
- Rapid Deployment: This is also a very important factor in software development and deployment.
- Non-Intrusive Design: Any such translation scheme should be built on top of a web site without modifying the actual web site.
- Multiple Views: This scheme should also allow the SIs and end users to develop custom views.

2. Research Direction

The importance of efficient content extraction from HTML pages for wireless access of World-Wide-Web becomes clear, especially in the context of the issues discussed in the previous section. There are several ways of addressing this problem. One of the ways is to segment the web pages into *zones* based on its HTML structure [1]. Once these zones are identified, attribute based analysis of the content can be carried out. This can result in the extraction of content that is *relevant* and *important* [2].

However, extraction of content from individual zones is not the complete solution. These zones can have content that are related and it may not make much sense in displaying these contents separately. So the next stage in this process is the analysis of relationship of these zones. This can be achieved in three ways.

- Proximity Analysis: This approach involves a relational analysis based on proximity. The natural order of these zones can sometimes be used as strong indicators to establish relationship.

- Content Classification: Content extracted from individual zones can be classified into various types, and this classification, taken with the context of proximity can be a powerful tool to establish a logical map between various zones.
- The second analysis involves using content understanding methods to approximate the content flow between zones. This analysis is to be based on natural language processing involving contextual grammar and vector modeling [3]. This would involve knowledge models and information retrieval techniques to define the relationship between various zones [4].

Once relationships between various zones are established, this can be used to reflow the content into a more meaningful and efficient manner that suits the requirements of smaller display devices. Various methods can be applied to combine the information thus collected, some of which can be found in [5], [6], [7]. Although primarily developed for character recognition, these techniques are generic enough to be applied to this particular task domain with little or no modification.

The stages required to implement this are the following:

- *Structural Analysis*: Analysis of the structure of a web document
- *Decomposition*: Decomposing a web document based on the extracted structure
- *Contextual Analysis*: Once decomposed into constituent sub-documents, analyze each document for its context.
- *Summarization (Labeling)*: This contextual analysis of each sub-document produces a summarization, which can be expressed as a sentence or sub-sentence (a *label*) indicating the content of this sub-document.
- *Table of Content (TOC)*: Since each of these sub-documents are summarized with an "intelligent" summary, these can be put together as a summary of the whole document, giving rise to a Table of Content (TOC). Each entry into this TOC points to specific sub-documents within each document.
- *Order of TOC*: The order in which the TOC is extracted depends on the "natural" order of the sub-documents extracted from the main document. However, this "natural" order is often misleading as the main "interesting" or "important" message of the document can be lost in the TOC. So it is important to analyze the content of each sub-document and display the TOC by re-ordering them based on their relative "importance".

3 Results

The performance of the system is best described on a real life application. Figure-3 shows the first page of the web page www.bcl-computers.com. As is clearly seen, this web site has a complicated multi-column layout. The content in presented in multiple segments with an implied relationship between these segments. For example, a story segment might be followed by a segment providing additional links to similar stories.

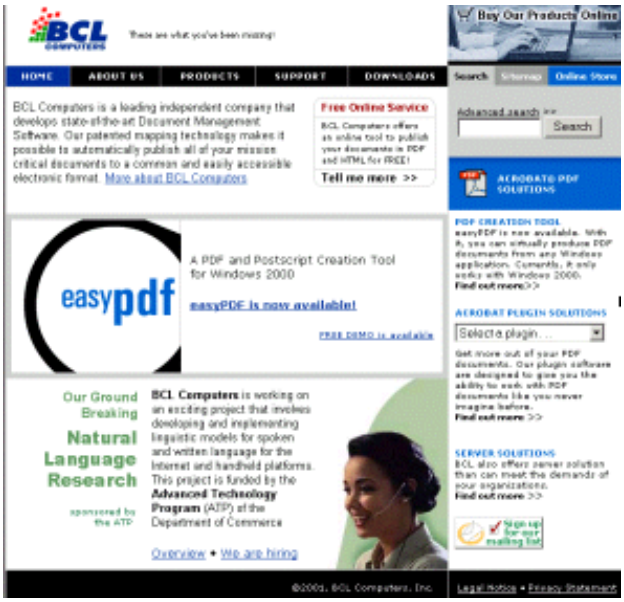

Figure-3 A sample web page: **www.bcl-computers.com**

The system analyses the layout and segments within the page and produces a summarized output (Figure-4). This is the total table of content (TOC). Each member of the TOC represents several segments within the page. Selecting any of these links will enable the user to go to the more detailed content associated with that TOC. For example, selecting the link "BCL Computers" will lead the user to the display shown in Figure-5. Clearly, the idea here is to keep the content intact, but the emphasis is on identifying which segments of the page should be put

together as a related segment that can be adequately described by a single label.

- BCL Computers
- Beta Tester wanted
- Natural Language Research
- PDF SOLUTIONS
- Acrobat Plugins
- Server Solutions
- Free Online Service
- Press Releases, Jobs, and Misc.

Figure-4: Summarized output.

BCL Computers is a leading independent company that develops state-of-the-art Document Management Software. Our patented mapping technology makes it possible to automatically publish all of your mission critical documents to a common and easily accessible electronic format. More about BCL Computers

Figure-5: More detailed content

A PDF and Postscript Creation Tool

for Windows 2000

easyPDF is now available!

FREE DEMO is available

Figure 6: Detailed content in the second level

Also there has to be some way to navigate between the various levels of abstraction. Since the content is in two levels, making the first level labels (TOCs) links solves this problem gracefully. For example, if the user selects the link “Beta Tester wanted” now from this display, the system will show the output of Figure-6.

sponsored by the ATP

BCL Computers is working on an exciting project that involves developing and implementing linguistic models for spoken and written language for the Internet and handheld platforms. This project is funded by the **Advanced Technology Program** (ATP) of the Department of Commerce

Overview • We are hiring

Figure-7: Second level abstraction of the summarized output: Following a single TOC only.

In the same fashion, it is possible to select the link “Natural Language Research” from Figure 4 and arrive at the display presented in Figure 7. Figure 8 displays similar results. In the same way story contents are summarized, sidebars and navigation links are also summarized. For example, Figures 9, 10, 11 and 12 shows the summarized bars from the page www.bcl-computers.co.uk.

ACROBAT® PDF SOLUTIONS

PDF CREATION TOOL

easyPDF is now available. With it, you can virtually produce PDF documents from any Windows application. Currently, it only works with Windows 2000. Find out more>>

Figure 8: More second level

Acrobat Plugins

Magellan (PDF to HTML)

Jade (PDF extraction)

Drake (PDF to RTF)

Freebird (PDF to graphics)

Figure 9: The summarized top bar.

Figure 10: A side bar.

Figure 11: More details

Figure 12: Top Bar

6 Supported Devices

The proposed system works in automatically summarizing live web content on the fly to fit smaller screen devices, such as PDAs and cellular phones with web capability. At the present time, the system supports all PDAs using an HTML 3.2 browser and also cellular phones using WAP, iMode (NTT DoCoMo), J-Sky (J-Phone) and EZweb (KDDI) formats. Live demonstration will be organized for more elaborate understanding of the system during the presentation of the paper.

7 Conclusion

This paper has presented a concept to extract content from HTML documents based on their structural analysis. Based on this extraction, a classification of the content can allow a more efficient representation of the content in context with the importance and logical relationship between various zones of the document. This document analysis approach should therefore be able to organize the content into a meaningful, understandable, manageable and useful representation.

1.8 References

1. H. Alam, A. F. R. Rahman, P. Lawrence, R. Hartono, K. Ariyoshi. Viewing Web pages on small form factor devices, U.S. Patent Application pending, 60/191,329.
2. H. Alam, A. F. R. Rahman, P. Lawrence, R. Hartono, K. Ariyoshi. Automatic summarization and display of web content in various display devices, U.S. Patent Application pending, 60/232,648.
3. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, Addison-Wesley, 1999.
4. H. Alam. Spoken language generic user interface (SLGUI). Technical Report, AFRL-IF-RS-TR-2000-58, Air Force Research Laboratory, Rome, NY, 2000.
5. A. F. R. Rahman and M. C. Fairhurst. Introducing new multiple expert decision combination topologies: A case study using recognition of handwritten characters. In Proc. 4th Int. Conf. On Document Analysis and Recognition, ICDAR97, vol. 2, pages 886-891, Ulm, Germany, 1997.
6. A. F. R. Rahman and M. C. Fairhurst, “Multiple expert classification: A new methodology for parallel decision fusion”. Int. Jour. Of Document Analysis and Recognition, 3(1):40-55, 2000.
7. A. F. R. Rahman and M. C. Fairhurst, “Enhancing consensus in multiple expert decision fusion”. IEE Proc. on Vision, Image and Signal Processing, 147(1):39-46, 2000.