

Understanding the Flow of Content in Summarizing 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. Current Needs

A host of personal communication devices has been devised recently to assist in various aspects of modern life. These vary from Personal Digital Assistants (PDAs) to cellular phones to e-books. In most cases, the practical requirement of portability dictates that the physical dimensions of these devices are small, which in turn severely reduces the available viewing area. The current way of viewing documents using these devices, specially web-based documents, can be very restrictive. There are several solutions to this problem. One solution is to make the content flow dictated by the size of the display device. This approach works well when the document to be viewed is relatively simple and specifically, linear. However, if the document has a complicated layout, with multiple rows and columns, has a mixture of different formatting styles and depends on the rendering engine of commercial browsers for the correct visual effect, this approach is hardly adequate. There is another school of

thought, which is to summarize the document into a compact, yet meaningful way. The advantages of this approach lie in better navigability, more efficient use of the available display area and ease of finding information from an otherwise maze of related content. The disadvantage of this approach, however, is that this needs far greater understanding of the flow of content in these documents to be accurate and thereby useful at all. This abstract addresses the second approach and discusses a possible solution scenario for this.

2. Open Problem

It is important to address the type of information available in the document layout in order to use that to produce an intelligent and automated summarization. Part of the problem lies in the fact that a web document (HTML/XML) is multi-layered and multi-directional. The final layouts of these documents depend on their structure hidden in the HTML code. When this structural information is incorporated with context, the reader is directed into *content* or *meaning*. This observation shows that extracting the structure of a web document can help in analyzing such documents successfully. The overall process can be divided into various phases, such as analysis of the structure of a web document (*Structural Analysis*), decomposing a web document based on the extracted structure (*Decomposition*), creation of constituent sub-documents based on its context (*Contextual Analysis*), producing a sentence or sub-sentence (a *label*) indicating the content of this sub-document (*Summarization or Labeling*) and finally these labels can be put together as a summary of the whole document, giving rise to a Table of Content (*TOC*). Each entry from this TOC points to specific sub-documents within the document.

3. Research Direction

As already emphasized, a way of addressing this problem is to analyze the structure of a HTML document. This analysis can be used to create a multi-dimensional representation of the two dimensional layout of the web page. This leads to a multiple level abstraction of the content of a specific page. At the highest level of abstraction, a summary of the whole page, properly labeled, is displayed. As the abstraction levels are traversed, more and more detail about the page is revealed, and at the lowest level, the complete content of the original page is accessible. This reasoning leads to the formation of a layered tree structure, where each level of the tree corresponds to a level of abstraction of the page.

In general, the content of a page is separated into constituent 'objects', the definition of these objects depends on the specific types of classes of elements they encompass [1]. Each such object, therefore, demonstrates specific properties that distinguish them from each other. These properties also provide subtle clues as to their importance within the whole page. Each of these objects is then analyzed in turn and each is given a specific label describing the essence of that object [2]. Part of this analysis can exploit stochastic grammar, which is a context-free grammar defining syntactic structures by means of a system of rules dealing with specific probabilities and associated rankings that correlate with the application probabilities of the rules. Stochastic grammar, initially developed for study of linguistics, provides a very useful tool to analyze and understand content of these objects, which is essential in determining the relative importance of an object within the whole document.

The problem here is not only to summarize and separate the content into objects, but also to reconstruct the document in a way so that the original essence of the document is preserved. In order to do that, web content needs to be analyzed, the document deconstructed and then reconstructed, based on logical and visual cues, producing a representation that enables the user of a cell phone or other PDAs to grasp the total experience of web surfing on their devices. The overall scheme can use multiple rule-based experts in a horizontal combination scheme [3,4,5]. Not only it can be totally device neutral and automatic, it can also guarantee that the total information content of the web is preserved and available

for browsing. This approach ensures that the web content remains totally faithful in all ways to the original document.

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.

3 Results

The way the system works is best described on a real life application. Figure 3 shows the first page of the web site www.bcl-computers.com. The content is presented in multiple segments with an implied relationship between these segments.



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

As a first stage for summarizing this page, we need to understand the flow of content. Figure 4 shows (partially) how the original page is automatically segmented based on the *type* of content and its flow. Each segment has its own boundary and a sequence number associated with it. This allows separation of content into clearly separable *zones*. For example, a segment having a lot of textual

content might be identified as a “story”, another primarily composed of links might be identified as “links”, others as “navigation”, “forms”, or “images”. Also this provides us clues about the flow of content by noting how segments of specific types are followed by segments of other types.

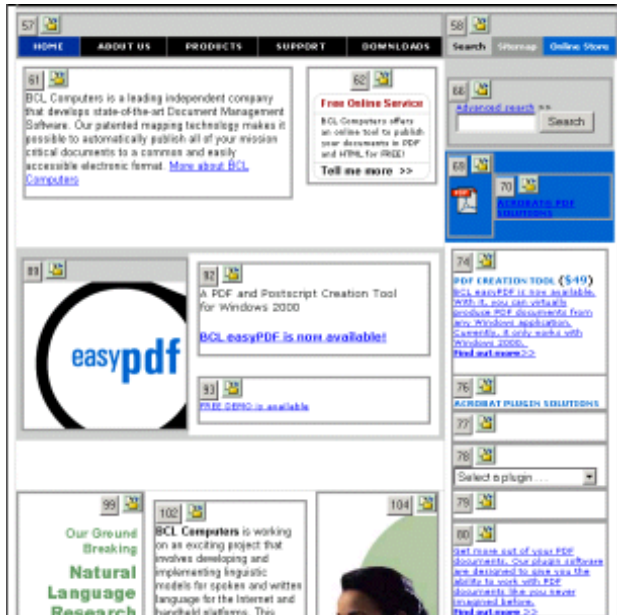


Figure 4: Automatic segmentation based on content flow.

Based on the type of content of the constituent segments, a summarized output as shown in Figure 5 is produced. 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 6. 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 (merging of segments). Similarly Figure 7 shows how the content can be reached if the link “Free online service” is selected.

- [BCL Computers](#)
- [Free Online Service](#)
- [Beta Tester wanted](#)
- [Natural Language Research](#)
- [PDF Creation Tool](#)
- [ACROBAT PLUGIN SOLUTIONS](#)
- [Find out more >>](#)
- [SERVER SOLUTIONS](#)
- [Search](#)
- [Acrobat Plugins](#)
- [Server Solutions](#)
- [Misc. Items](#)

Figure 5: 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 6: More detailed content



Figure 7: Detailed content in the second level

In the same way “story” type contents are summarized, sidebars and navigation links are also summarized. For example, Figure 8 shows a summarized navigation bar from the page www.bcl-computers.co.uk.

Acrobat Plugins

[Maqellan \(PDF to HTML\)](#)

[Jade \(PDF extraction\)](#)

[Drake \(PDF to RTF\)](#)

[Freebird \(PDF to graphics\)](#)

Figure 8: The summarized navigation.

Web pages are live pages, and therefore any conversion has to make sure that active component within such pages are preserved after the summarization process takes place. Figure 9 shows that the “search” mechanism is nicely bunched together in a segment, although the segmentation process produced two separate segments. Understanding flow of content allows us to merge such segments based on the type of content of these segments.

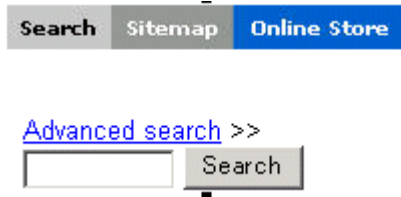


Figure 9: Search mechanism

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 summarize HTML documents based on their structural analysis. It is argued that since this type of summarization takes advantage of the layout of the document, it is thereby able to organize the content into a meaningful, yet compact format. This makes the content more understandable, manageable and useful.

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. 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.
4. 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.
5. 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.