

# Assuming Accurate Layout Information for Web Documents is Available, What Now?

Hassan Alam, Rachmat Hartono, Aman Kumar, Fuad Rahman<sup>§</sup>, Yuliya Tarnikova and Che Wilcox  
*BCL Technologies Inc.*  
*fuad@bcltechnologies.com*

## Abstract

*Re-authoring of multi-modal documents, such as web pages, has proved itself to be a difficult problem. With the widespread recognition and subsequent adoption of XML, expectations were high that the WWW will see a transformation in terms of web page authoring and sharing. Unfortunately, that has not happened. Today, as was last year, and the year before that, the majority of the web pages are still written using HTML, with little or no content descriptions and explicit relationship between data and structure. That has made the task of web page re-authoring difficult. But what if we had accurate layout information for these very troublesome web pages? This paper discusses some interesting ideas about the future of web page re-authoring with this assumption.*

## 1. Introduction

Web page re-authoring is a problem that has been studied for a number of years. This has attracted renewed attention recently as various handheld devices such as Personal Digital Assistants (PDAs) and cell phones have become capable of browsing web pages. The limited form factor of these devices in terms of display area is a major obstacle to mass adoption of this technology. The majority of the web pages are still designed to be viewed by full resolution desktop monitors and any attempt to effectively browse and locate information using tiny viewing displays becomes very difficult. Keeping this in mind, researchers have turned their attention to web page re-authoring to make the rendering most appropriate given local display area restrictions.

Since most of the web pages in the world are still written using HTML, any re-authoring requires transformations into various different formats including HDML, HTML 3.2, WAP, iMode (NTT DoCoMo), J-Sky (J-Phone) and EZweb (KDDI) to name a few. On top of that, format translation is only part of the problem. The content can be often encoded in tabular (HTML table

construct) format, and different multi-media objects such as graphics, hyperlinks, flash, java codes etc. can be embedded into it. This makes the task of accurate transformation very difficult.

In [1], to be presented in ICDAR 2003, we have discussed how it is possible to use a combination of natural and non-natural language techniques to produce high quality re-authoring solutions for web pages written in HTML. These so-called 'legacy' documents are all we have before a universal high level formatting is adopted by everyone (and the dream is on!). Now if we can *assume* that accurate layout information is available, information such as "this is a table, rectangle coordinates xx, yy, ...", "this is a paragraph of texts, rectangle coordinates xx, yy, ...", "this is a graphic, rectangle coordinates xx, yy, ...", etc., what do we do then? How do we use this information? How do that information help us in getting better re-authoring solutions? This paper discusses some of the pathways to this future.

## 2. Related Work

Before we start, here is a very brief outline of past work. Over the years, researchers have proposed different solutions to the problem of web page re-authoring. The possible solutions can be categorized as:

- **Handcrafting:** Handcrafting involves typically crafting web pages by hand by a set of content experts for device specific output. This process is labor intensive and expensive. BBC ([www.bbc.co.uk](http://www.bbc.co.uk)) web site is an example, where handcrafted web content is served in text only format.
- **Transcoding:** Transcoding [2] replaces HTML tags with suitable device specific tags, such as HDML, WML and others.
- **Adaptive Re-authoring:** The research on web page re-authoring can be broadly separated into two parts: approaches that explicitly use natural

---

<sup>§</sup> Corresponding author. The authors gratefully acknowledge the support of a Small Business Innovation Research (SBIR) award under the supervision of US Army Communications-Electronics Command (CECOM).

language processing (NLP) techniques based on computational linguistics [3,4], and the approaches that use non-NLP techniques [5,6].

### 3. Web page summarization for handheld devices

Web browsing on a handheld device can be a very annoying task. The principal reason for this is the display area available on these devices. Typically, a PDA (such as a Palm V) can have 160X160 pixels of resolution, whereas a cell phone might be as low as 120X80. PocketPCs have much higher resolution, but the price point is still prohibitive. Trying to use these devices to browse or try to find information on the web is a tiring task. On top of that, since a page is downloaded as a whole before viewing, this can take up significant time. Summarization is a very attractive solution in these cases.

In [1], we propose a web page re-authoring approach based on a combination of NLP and non-NLP techniques. The following is a brief outline of the proposed approach:

- **Web page data structure:** If HTML is used to compose a web page; the data ("content") is arranged using an HTML data structure.
- **Content analysis:** Content analysis aims to decompose a web document based on the extracted structure from the tree hierarchy.
- **Content processing for re-authoring:** The aim is to re-author the content in a format that is most suited for displaying in a target device.
  - **Verbatim.** The content is not processed any further and is displayed in the target device 'as-is'.
  - **Transcode.** The content is re-flowed by replacing HTML tags with suitable device specific tags (HDML, WML etc.).
  - **Summarize.** The content is summarized using NLP and non-NLP techniques.
- **Node merging and segmenting:** In cases where the node is a neighbor of other nodes with similar type of content, the nodes are merged. In cases, where the content is too large, it is split into smaller coherent segments.
- **Representing the complete web page:** Once merging and segmentation is completed, it is possible to recreate the original web page by combining these merged and segmented nodes.
  - **"To summarize or not to summarize":** It so happens that not all the segments of

a web page are good candidates for NLP summarization.

- **Creating a label:** For creating a label, non-NLP techniques are adequate. Visual clues are used to detect the most important segments of the content, or a 'label'.
- **Creating a summary:** NLP techniques need to be employed to create short summaries of the content [7,8].

However, although this hybrid of NLP and non-NLP approaches has shown great promise, it still lags the relationship between the structure and the content.

As is evident from the work of researchers in the last few years, HTML is hugely exciting, and easy to adopt, but is not very receptive to data descriptions, sharing, automatic conversion, machine readability and interoperability. Since web page re-authoring needs to transform and adapt existing web pages to the display capabilities of other devices, in general of smaller display area, an explicit relationship (i.e. ontology) between the content and the data structure is extremely important.

Assuming accurate layout information is available, we explore how adoption of linguistic knowledge and semantics with a combination of explicit ontology and XML representation can solve this problem gracefully.

### 4. The Future: Marrying Ontology with XML

We assume that now have information regarding the structure of the web page, but what we do not have is an ontology defined for that domain. Web pages can be very variable, defining ontology applicable to all web pages, therefore, can be a daunting task. Here we attempt to define a generic ontology defining all web pages.

The approach described in [1] does not need to extract very detailed structure of web pages in the re-authoring process, not do we require it. A high level structure is all that is required to apply these ideas. The extracted elements then can be conveniently used to generate an intermediate XML description of the web page. On the other hand, it is entirely possible to map these high level structures and their relationship to the data using a simple ontology.

Ontology is a specification of a conceptualization [9]. Ontology establishes a joint terminology between members of a community of interest. These members can be human or automated agents. In order to define such ontology for the domain of web pages, we need to define four things, a list of the elements, concept hierarchy, concept association and finally rules or axioms [10]. **Table 1** has a list of possible elements of a web page.

web address	Ticker	icon	Navigation	Map
Title	graphic banner	box	bulletlist	Section
Heading	side burst	content	panel	Highlights bannerad (advertisement) (border, tab, height, width)
sub-heading	Image	text	menu	Coordinates
banner ad	graphic	link	author	Table
directions				

**Table 1:** A possible list of elements

The concept hierarchy based on these elements can be described in the following way:

```
Object []
address :: Object
  computerAddress :: address
  WebAddress :: computerAddress
```

```
text :: Object
  title :: text
  line :: text
  headline :: line
  subheading :: heading
  banner :: line
```

```
message :: Object
  promotion :: message
  ad :: promotion
  bannerAd :: promotion
```

```
representation :: Object
  symbol :: representation
  ticker :: symbol
```

```
artifact :: Object
  creation :: artifact
  representation :: creation
  display :: representation
  image :: display
  icon :: display
  graphic :: image
  illustration :: representation
  graph :: illustration
```

```
form :: Object
  figure :: form
  rectangle :: figure
  box :: rectangle
```

```
communication :: Object
  message :: communication
  content :: message
```

```
relation :: Object
```

```
communication :: relation
writing :: communication
coding system :: writing
computerCode :: codingSystem
command :: computerCode
link :: command
```

```
activity :: Object
control :: activity
direction :: control
steering :: direction
navigation :: direction
```

```
relation :: Object
communication :: relation
content :: communication
information :: content
database :: information
list :: database
bulletlist :: list
menu :: database
writtenCommunication :: communication
writing :: writtenCommunication
section :: writing
genre :: communication
prose :: genre
nonfiction :: prose
article :: nonfiction
```

```
artifact :: Object
display :: artifact
window :: display
panel :: window
creation :: display
representation :: creation
map :: representation
```

```
animateThing :: Object
person :: animateThing
author :: person
```

Similarly, the concept association can be conveniently expressed as:

```
WebAddress[name =>> STRING; colon =>> STRING; slash =>>
STRING; domainName => STRING;
newspapername =>> webNewspaper; portal =>> webPage]
text[font =>> NUM; format =>> STRING; size =>> NUM]
banner[color =>> STRING; background; sharp;]
ticker[font =>> NUM; format =>> SRING; size =>> NUM; case;]
graph[color =>> STRING; background; size =>> NUM]
image[color =>> STRING; background; size =>> NUM]
box[text =>> STRING; size =>> NUM; font =>> NUM; format =>>
STRING]
link[color =>> STRING; size =>> NUM]
bulletlist[size =>> NUM; font =>> NUM; format]
article[font =>> NUM; format =>> STRING; size =>> NUM; link;
author =>> STRING; subheading; section]
map[format =>> STRING; size =>> NUM; link; text]
```

Finally, the rules or axioms are expressed as follows:

*FORALL Pers1, Art1*

*Art1: Article[author ->> Pers1] <->  
Pers1: Person[Article ->> Art1]*

*FORALL Rep1, Rep2*

*Rep2: Banner[containsAd ->> Rep1] <->  
Rep1: BannerAd[banner ->> Rep2]*

*FORALL Dis1, Dis2*

*Dis1: Graph[illustration ->> Dis2] <->  
Dis2: Image[illustrates ->> Dis1]*

*FORALL Tex1, Fig1*

*Fig1: Box[Line ->> Tex1]  
Tex1: Text[Line ->> Fig1]*

This shows that the ontology for web pages can be conveniently expressed in terms of the derived XML structure. Since the hybrid method of web page re-authoring, presented in [1], exploits all these structures, it is very easy to exploit this ontology to generate the most appropriate format of a web page for a specific device. So the re-authoring takes a different path than was described earlier: The structure is generated, the processing mode is selected, content of each block is either summarized or re-flowed, output level (either single or double levels) decided, and then the ontology is used to re-format the source web page. This improves the quality of the output in many ways. Not only that it becomes possible to capture the contextual relationship among various components within the document, it also leads to better understanding of the information contained within the document. This additional information can be used in other processes, such as document categorization and contextual search.

Future mobile web browsing will be very much simpler as semantic web becomes more widely used and accepted.

## 5. Conclusion

This paper has presented some ideas about how to produce better web page re-authoring solutions by using linguistic knowledge and ontology assuming accurate layout information for web pages is available. It is shown that such an approach will produce high quality intelligent summary for web pages allowing fast and efficient web browsing on small display handheld devices such as PDAs and cell phones.

One assumption of this approach is the use of explicit classification of document elements in addition to accurate layout information. In a multi-modal web page, specific identification and classification of component structures, such as headings, text blocks, images, graphics, audio, video, flash, image maps etc. leads to more precise re-

authored output. Since each of these components are labeled at early stages of the process, putting them together while maintaining the correct contextual relationships among these components becomes possible. During the presentation of the paper, demonstrations will be arranged to show this functionality.

It is also implicitly assumed that the future of mobile browsing lies in the adoption of semantic web technology. However, before that realizes, the proposed approach offers a workable compromise to generate high fidelity re-authored web pages suitable for viewing on a host of display devices automatically from currently available multi-modal web pages.

As is stated in the beginning, this is an exploratory paper offering a specific pathway to the future of web page re-authoring provided accurate layout information is available. Currently, it is beyond the capability of any algorithm to achieve this level of accuracy. However, approximations to that accuracy are attainable and even practical. It will be interesting to discuss other possibilities in this space during the DLIA workshop.

## References

1. Alam, H., Hartono, R., Kumar, A., Tarnikova, Y., Rahman, F. and Wilcox, C., "Web Page Summarization for Handheld Devices: A Natural Language Approach", 7th Int. Conf. on Document Analysis and Recognition (ICDAR'03). In press.
2. Hori, M., Mohan, R., Maruyama, H. and Singhal, S., "Annotation of Web Content for Transcoding". E3C Note. <http://www.w3.org/TR/annot>.
3. Berger, A. and Mittal, V., "OCELOT: A System for Summarizing Web Pages". Research and Development in Information Retrieval, pages 144-151, 2000.
4. Buyukkokten, O., Garcia-Molina, H., and Paepcke, A., "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices". Proc. of the Tenth Int. World-Wide Web Conference, 2001.
5. Bickmore, T., Girgensohn, A. and Sullivan, J., "Web page filtering and re-authoring for mobile users". The Computer Journal, 42(6):534-546, 1999.
6. Zhang, H., "Adaptive content delivery: A new research in media computing". Proc. Int. Conf. on Multimedia data Storage, retrieval, Integration and Applications, MDSRIA, 2000.
7. McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Schiffman, B., and Teufel, S., "Columbia Multi-Document Summarization: Approach and Evaluation". Proc. of the Workshop on Text Summarization, ACM SIGIR Conference, 2001.
8. McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E., "Towards Multidocument Summarization by Reformulation: Progress and Prospects". AAAI/IAAI, pages 453-460, 1999.
9. Gruber, T., A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993.
10. Erdmann, M. & Studer, R., Ontologies as conceptual models for XML documents. Twelfth Workshop on Knowledge Acquisition, Modeling and Management, 1999.