Verifying Genre-based Clustering Approach to Content Extraction

Suhit Gupta
Columbia University
500 W. 120th Street
New York, NY 10027
001-212-939-7184
suhit@cs.columbia.edu

Hila Becker
Columbia University
500 W. 120th Street
New York, NY 10027
001-212-939-7100
hila@cs.columbia.edu

Gail Kaiser
Columbia University
500 W. 120th Street
New York, NY 10027
001-212-939-7081
kaiser@cs.columbia.edu

Salvatore Stolfo
Columbia University
500 W. 120th Street
New York, NY 10027
001-212-939-7080
sal@cs.columbia.edu

ABSTRACT

The *content* of a webpage is usually contained within a small body of text and images, or perhaps several articles on the same page; however, the content may be lost in the clutter, particularly hurting users browsing on small cell phone and PDA screens and visually impaired users relying on speed rendering of web pages. Using the genre of a web page, we have created a solution, Crunch that automatically identifies clutter and removes it, thus leaving a clean content-full page. In order to evaluate the improvement in the applications for this technology, we identified a number of experiments. In this paper, we have those experiments, the associated results and their evaluation.

Categories and Subject Descriptors

I.7.4 [**Document and Text Processing**]: Electronic Publishing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based Services*

General Terms

Human Factors, Algorithms, Standardization.

Keywords

Website classification, clustering, content extraction, reformatting, HTML, context, accessibility, speech rendering.

1. INTRODUCTION

Web pages are often cluttered with extraneous materials, perhaps attempting to attract the user's attention or improve the user's efficiency, but they may end up *distracting* the user from the actual content. These "features" may include script and flash-driven animations, other kinds of images not directly associated with a main text body, menus and guides, links scattered around the screen, etc. The automatic extraction of heuristically-defined "content" from webpages has many applications, including enabling end-users to access the web more easily over constrained devices and providing better access to the web for the blind or otherwise disabled. We have developed a framework, Crunch [1] [2] [3], as a web proxy that employs various heuristics in the form of filters and filter "settings" that are applied to the DOM (www.w3.org/DOM) of the web page to achieve content extraction via clutter reduction.

In order to reduce human involvement in selecting the heuristic filter settings, we consider utilizing a website's *genre* classification. Crunch can then obtain some previously (manually) adjusted settings for a newly visited website by automatically classifying it as sufficiently similar to a genre-cluster of known websites, at least one of which has "known good settings" – which, we found empirically, produces better content extraction results than *any* possible one-size-fits-all default settings. The clustering algorithm and genre-based adaptive application of filters have been described in our prior publication. [3]

To evaluate the effectiveness of our clustering approach and the subsequent clutter free web pages, we conducted experiments to test Crunch's benefits in the two areas mentioned above. In this paper, we present our experiments and associated results.

2. EXPERIMENTS & RESULTS

2.1 Speech rendering experiment

The goal of our first experiment was to consider the "readability", by conventional screen reader software, of the extractions produced by Crunch compared to the raw webpages. We measured the length of time, in seconds, that it took the screen reader to render (speak) each variant of a given page, considering a variety of webpages. We ran trials with both the "free demo" version of JAWS (www.freedomscientific.com) and a licensed copy of Home Page Reader (www.ibm.com), which vary slightly in how they preprocess the raw HTML – although the measured time differences were under 3 seconds in all cases, so we average the results in our reporting below. The idea was to determine the amount of time a visually disabled user might save by using our content extraction technology. The notion of "content" is inherently subjective and our determination of what is the content vs. non-content was performed by visual and auditory inspection.

For this experiment, we chose 11 websites that represent a variety of layout formats. We included websites from all the major genres that appeared in our corpus (e.g., news, shopping, tech news, astronomy), but were also careful to cover different structures (columns, single-body articles, portal-based and blog-style sites), as well as W3C-compliant "accessible" sites vs. non-compliant sites. We passed the original webpages vs. the Crunch outputs through the screen readers. We then measured the time it took, in minutes, to read (speech render) the entire webpages.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

Table 1 – Speech rendering results

Site (accessed on May 18th, 2005)	Read original webpage (minutes)	Read page produced by Crunch (minutes)
CNN.com front page	10 :09	1 :08
CNN.com subsidiary page	7:35	2:44
Slashdot.org front page	25:20	17:53
Slashdot.org article page	14 :13	6 :15
MSNBC front page	10 :47	2 :40
MSNBC article page	11:12	3:43
Yahoo News front page	25:15	16:39
Yahoo News article page	14:08	5:13
NASA Ames front page	2 :18	1 :48
NASA Ames Research page	1:57	1:17
Amazon.com front page	13:28	7:42

From these tests and from the anecdotal accounts of visually impaired users (e.g., attendees at the 2005 W4A meeting), it is clear that blind web users typically spend tens of minutes listening to nearly any single webpage using a commercial screen reader alone - and this is absolutely unacceptable! We found that using Crunch together with such a screen reader reduces by 10-80% the time spent in reading the page while the content on the webpage remains qualitatively accurate. The least significant improvement (< 10% speedup in reading) using Crunch was on the main page of a given site, where the settings preserved a larger percentage of navigation links (Crunch's heuristics distinguish between front and auxiliary pages since front pages are often intended to operate as portals). The greatest improvement noticed was on subsidiary pages of websites, usually containing contentful articles.

2.2 Constrained screen testing

We also evaluated how well Crunch compared to other content extraction and webpage reformatting technologies designed for devices with limited screen real-estate. We used the same samples as for the speech rendering tests above, and displayed both the original webpages and the pages output by Crunch on various combinations of handheld devices and browsers. We tested the system on the Toshiba e805 and HP iPaq 2215 PDAs running Microsoft's PocketPC OS, with Pocket Internet Explorer and BitStream's Thunderbird browsers, respectively. We measured the amount of content on the first screenful at both 320x240 and 640x480 resolutions. We also used a Blackberry 7100t running a proprietary Blackberry browser and a Microsoft Smartphone i600 running Internet Explorer and Opera Mobile Browser.

The purpose of these tests was to demonstrate the increase in "relevant and useful" content displayed on a small screen when using Crunch vs. not using Crunch. We would like to reiterate that, in the general-purpose case absent any model of the author's or reader's intents, content is subjective. For this experiment, we define content as the number of relevant words shown on the screen, measured by visual inspection.

From the data presented in Table 2, we see that Crunch with genre-based automatic selection of filter settings is very useful towards maximizing the amount of content displayed on constrained devices. The most significant difference was on a 320x240 resolution PDA screen, where there was on average a 215% increase in the amount of content displayed on the screen.

Table 2 - Constrained device testing results

Number of words (PDA 320x240)		Number of words (PDA 640x480)		Number of words (Blackberry)		Number of words (Opera on Smartphone)	
I	II	I	II	I	II	I	II
29	38	102	217	17	30	10	32
29	185	158	338	17	68	13	59
49	154	134	270	48	68	43	63
45	80	215	215	48	68	43	63
56	56	111	111	27	27	27	27
123	123	370	370	14	53	13	48
20	34	20	75	27	33	25	29
20	93	20	93	27	51	25	45
7	34	185	185	3	19	3	19
15	112	112	112	12	30	12	28
28	35	247	247	12	34	43	43

I – without Crunch, i.e., original webpage.
II – with Crunch, i.e., the page is passed through Crunch, with automatic genre-based settings.

This increase jumped dramatically up to a 750% when considering only news articles. With 640x480 resolution, we found an average increase of 133% of content on the first screenful. Several of the pages tested were able to fully render within that screenful. When testing with the cell phone browsers, we found the results to be almost identical, almost 185% improvement in both cases, presumably because of the very similar screen sizes. The main difference was due to the Opera browser's default behavior of jumping to the "middle" of the page where it found the largest concentration of text attempting to skip over anticipated non-content, which was lacking on the Blackberry. In none of our trials did a webpage rendered using Crunch display less content on the first screenful than the original page rendered on the same constrained device. However, the Opera comparison is somewhat problematic, not always counting the same words, due to Opera's skip-to-the-middle heuristic.

3. CONCLUSION

In this paper we have presented experiments that demonstrate the substantial improvement achieved by the application of content extraction via clutter removal to web pages, especially in the domains tested. We also hope that the experimental process shown here is used towards future content extraction applications so as to evaluate the "goodness" rate achieved.

4. REFERENCES

[1] Suhit Gupta, Gail Kaiser, David Neistadt, Peter Grimm, "DOM-based Content Extraction of HTML Documents", 12th International World Wide Web Conference, May 2003

[2] Suhit Gupta, Gail Kaiser, "CRUNCH - Web-based Collaboration for Persons with Disabilities", W3C WAI, Teleconference on Making Collaboration Technologies Accessible for Persons with Disabilities, Apr 2003

[3] Suhit Gupta, "Context-Based Content Extraction of HTML Document", PhD Dissertation, New York, NY, June 2005.