Assessing the printability of web sites

Krishnan Ramanathan, Dan Weeks, David Pitkin, Ajay Gupta, Ashita Sriraman

HP Laboratories
HPL-2008-128

Keyword(s):
printability, webpage modeling, printer friendly, web printing

Abstract:

A significant percentage of home printing happens off the Internet. However, most websites today are designed without much consideration to the printability of the web pages. Parameters used to assess content quality are not a reliable indication of the printability of web sites. We describe a tool that enables an automated assessment of the printability of web sites. Our tool allows identification of the poorly printable pages on a website (and the Internet) and allows HP to measure its progress in the goal of making the Internet more printable.

External Posting Date: October 6, 2008 [Fulltext]   Approved for External Publication
Internal Posting Date: October 6, 2008 [Fulltext]

Assessing the printability of web sites

Krishnan Ramanathan[1], Dan Weeks[2], David Pitkin[2], Ajay Gupta[1], Ashita Sriraman[3]

[1]OST/HP Labs, [2]IPG(Ink supplies marketing)

Email Addresses (firstname.lastname@hp.com)

Abstract

A significant percentage of home printing happens off the Internet. However, most websites today are designed without much consideration to the printability of the web pages. Parameters used to assess content quality are not a reliable indication of the printability of web sites. We describe a tool that enables an automated assessment of the printability of web sites. Our tool allows identification of the poorly printable pages on a website (and the Internet) and allows HP to measure its progress in the goal of making the Internet more printable.

1 Introduction

The World Wide Web is accessed by users for informational and transactional purposes. A crucial part of the web experience is printing. Documents are printed off the Internet for various reasons: as receipt of some transaction, proof of correspondence, for sharing and annotating or because the availability of the document on the website later is uncertain. Unfortunately, print is not a first class citizen of the web today. A lot of care is taken in the design of web pages from a display and navigational perspective but not from a print perspective. In fact, a large number of web pages are not suitable for printing and lead to a bad print experience.

Services such as Tabblo [2] improve the print experience on the Web. Tabblo allows web sites to send content to the Tabblo server which then formats the content for print using professionally designed templates. However, since Tabblo requires active participation of the website owner, a large part of the printable web is still not touched by Tabblo. It is important to be able to measure the printability of the non-Tabblo printable web and create solutions (including Tabblo-fying it) to make it more printable.

The parameters used to measure content quality of web pages (such as the Google PageRank) are not useful for assessing the printability of web pages. This is because websites with high content quality often carry a large number of advertisements (aimed at monetizing traffic to the websites) to the detriment of print quality. Although, some web sites provide a "Print-friendly" button on web pages, we observed that this does not guarantee a good print experience. Also, there is a wide variation in the print quality of individual web pages on the same website. A different set of parameters (than those used to measure content quality) are required to assess whether a web page will print well and a different engine is required for comparing the printability of different websites.

2 Our solution

We have designed and implemented a prototype "printability" engine for the Internet that enables an automated assessment of the print quality of web sites. Our engine scores each individual page on a website on its printability, aggregates the page scores to provide a printability score for the entire website and suggests improvements to the website owner to improve printability.

We first identified web page parameters that are most important from a print perspective. We then devised a scoring scheme (based on these parameters) for ranking a web page for printability. Finally, we implemented a solution using an open source search engine (Nutch) to crawl websites, score each page on the website for the printability and provide an aggregated score for the entire website.

2.1 Parameters that influence printability

We identified a number of parameters that influence and affect printability

[3] Intern at HP Labs when this work was done

- Page Width: The width of a web page is one of the most important qualities when it comes to printing. If the width of the page is too large, the page content is chopped off, mostly at the right end, rendering the print useless.
- Font Size: If the font size is too small, the printed page is hard to read. If the font size is too large, paper and ink is wasted. The ideal font-size for print is 12 point.
- Font Color: A font color that is readable on screen may be unreadable on print. A darker font color is preferable from a print perspective.
- Advertisements: Advertisements must not be present on a printed page since they draw attention away from the main content and also use up paper and ink without compensating the consumer.
- Navigation Bars: Navigation Bars too, like advertisements, serve no purpose on a printed page.
- Search Boxes: Search Boxes are used only online and have no utility on the printed page.

Parameter	Value	Score
Page Width	Retains All Content	+40
	Content Chopped Off	-41
Font Size	> 13pt	+14
	11-13pt	+15
	10pt	+12
	9pt	+5
	8pt	+3
	<8pt	-75
Font Color	Good	+15
	Ok	+7
	Bad	-75
	Not Standard Color	0
Navigation Bars	None	+5
	Top/Bottom OR Side	+3
	Top/Bottom AND Side	0
Advertisements	Absent	+5
	Present	-3
Search Boxes	Absent	0
	Present	-2

Table 1 Scores for print parameters

2.2 The scoring scheme

After identifying the parameters that influence printability, we scored them based on how severely the parameter affects printablity. The score of a web page is the normalized (to 100) sum of the scores for all the parameters. Page width, Font size and font colour are given high weightages because they have a huge influence on the print experience. Font colors are classified into good, ok and bad (based on a pre-decided color table). Table 1 lists the scores for the parameters.

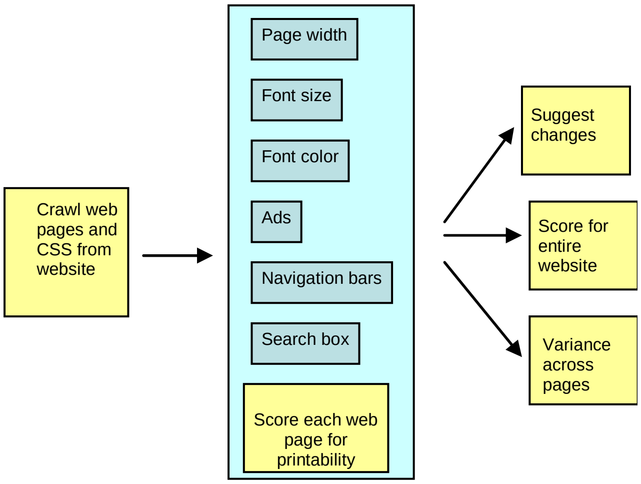

Figure 1 The solution architecture

2.3 Solution architecture

The solution architecture is shown in figure 1. The process consists of two phases: A web page fetch phase and a page scoring phase.

Fetch phase: In the fetch phase, the web site is crawled using the Open Source Nutch crawler [1] and all the (reachable) pages on the website are fetched and stored in a Lucene index. If the web page has an associated CSS (Cascading style sheet) for a print-friendly version of the page, the CSS is parsed and the print friendly version of the page (if it exists) is retrieved and scored. If no print-friendly version exists, then the web page itself is scored.

Page scoring phase: In the page scoring phase, each web page is retrieved from the Lucene index and parsed for print parameters using an off-the shelf HTML parser. The page is then scored for printability as described above. After all the pages have been scored, the average score of the website is computed. We then compute the variance of scores across the web pages and identify the best and worst page (from a print perspective) on the website. We also list recommendations for the web site owner to improve the printability of pages on the website.

3 Evaluation

We have implemented this tool and scored a few websites. We employed a light crawl to avoid stressing the websites. The tool outputs a report for each site with parameter scores for each page that has been crawled. Summarized results are shown in table below. Note the wide variation in best/worst page score on Wikipedia/Yahoo health.

	Wikipedia	Yahoo travel	Yahoo health	Rent.com	News.com
Pages evaluated	2288	73	114	99	44
Pages with print friendly button	1578	18	28	3	38
Best/Worst page scores	92.59(-25.9)	96.2(91.35)	92.59(-9.87)	96.2 (61.72)	90.1 (80.2)
Main problem	Width exceeds printable page, poor font color	Navigation bars on printed page	Width exceeds printable page	Variable font sizes, poor font color	Advertisements, Navigation bars

The tool outputs recommendations to improve the site. These recommendations are based on a handcrafted knowledge base and the parameter scores for web pages on the site. For example, the recommendation for Yahoo travel was:

Please remove the top and side navigation bars on the printable page to get a printability improvement of 7.4%

The tool also outputs the urls on the website with the best and worst scores. The objective is for the web site owner to be able to look at these pages and infer what he got right and where he went wrong.

4 Related work

Individual websites (such as the New York Times website) have data on which are the most printed pages on a particular site. However, this data will include pages printed for their content quality and is not a reliable indicator of the page printability.

There is some academic research on assessing a web page from a display perspective. Ivory and Hearst [3] describe a method for computing the statistical profile of highly rated websites. The authors show that quantitative measures such as: the amount of text on a web page, the number of links, the link, graphic and page formatting and the page loading performance could be used to predict whether human experts would rate the page as good design. Web page analyzers [5] provide an automatic assessment of a web page quality (based on parameters like number of html elements, number of images, page loading time). Fogerty [6] discuss the use of numerical optimization techniques for solving rendering problems on displays. They also cite the linedrive system that introduces distortions to make loops clear on a printed map. Dai [4] present a system for classifying web pages as being of commercial or non-commercial intent. A human labeled corpus of pages with commercial and non-commercial intent was used to train a Support Vector machine for classification. This technique could be used to build a classifier that classifies pages as navigation or content pages.

5 Future work

We would like to extend the current tool to make it more robust (handle deformed HTML) and to be more intelligent (by constructing classifiers to distinguish content pages from navigation pages). We plan to extend the printability parameters based on documents people actually print. For instance, the number of words in a document and the length of the title of the document could be a factor in printability. We will also validate the printability ranking with human experts to ensure that high printability scores correlate with a great print experience. We will also extend these tools for dynamically generated pages by working with websites for generating such pages (e.g. based on form input).

We also wish to develop methods to expose the printability information to web page designers. We have built a simple web service (servlet running in Apache Tomcat) where a website owner can get a report on the printability of his website. We will explore whether search engine results can be annotated with printability scores. Finally, we will also explore embedding the printability advisor "in situ" into web page design tools such as Microsoft FrontPage.

References

1. Nutch Open source crawler, http://en.wikipedia.org/wiki/Nutch

2. http://developer.tabblo.com/

3. Melody Ivory and Marti Hearst, Statistical profiles of highly rated websites, Proc. Of CHI 2002.

4. Honghua Dai et.al, Detecting online commercial intention (OCI), Proceedings of WWW 2006 conference.

5. Web page analyzer, http://www.websiteoptimization.com/services/analyze/

6. James Fogarty and Scott Hudson, GADGET: A Toolkit for Optimization-Based Approaches to Interface and Display Generation, Proceedings of the 16th annual ACM symposium on User Interface Software and technology, Vancouver, Canada, 2003.