First Steps in Archiving the Mobile Web: Automated Discovery of Mobile Websites

Richard Schneider
Harding University
Computer Science Dept
Searcy, Arkansas, USA 72149
rschneid@harding.edu

Frank McCown
Harding University
Computer Science Dept
Searcy, Arkansas, USA 72149
fmccown@harding.edu

ABSTRACT
Smartphones and tablets are increasingly used to access the Web, and many websites now provide alternative sites tailored specifically for these mobile devices. Web archivists are in need of tools to aid in archiving this equally ephemeral Mobile Web. We present Findmobile, a tool for automating the discovery of mobile websites. We tested our tool in an experiment examining 10K popular websites and found that the most frequently used technique used by popular websites to direct mobile users to mobile sites was by automated client and server-side redirection. We found that nearly half of mobile web pages differ dramatically from their stationary web counterparts and that the most popular websites are those most likely to have mobile-specific pages.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection.*

General Terms
Design, Experimentation, Measurement.

Keywords
Mobile web, web crawling, web archiving.

1. INTRODUCTION
Individuals are increasingly using mobile devices like smartphones and tablets to access the Web. A recent report shows that 69% of respondents have used mobile devices to access the Web in the past 12 months [11], and StatCounter shows an upward trend in mobile web surfing with 14% of all web traffic in 2013 coming from mobile devices [17]. Due to the smaller screen size and limited bandwidth of mobile devices, many websites provide web pages designed specifically for these devices. These *mobile pages* make up the Mobile Web. Search engines like Google have recently started crawling the Mobile Web in order to provide better search results for mobile device users [9].

Web archivists are also turning their attention to the Mobile Web. Mobile pages are often significantly different than their stationary counterparts, containing smaller images, constrained text, fewer links, and interfaces designed for finger input. To preserve the Mobile Web for posterity, web archivists need tools to identify mobile web pages, crawl them, and present them to users.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

We present a tool called Findmobile which automates the discovery of mobile pages. We share an experiment using Findmobile to examine the top 10K websites ranked by Alexa. We determined the most popular methods websites commonly use to expose their mobile sites, compared the content of mobile pages of these sites with their stationary-page counterparts, and found a clear correlation between a website's Alexa rank and its use of mobile web pages. We hope our efforts will kick-start the efforts of web archivists who are interested in preserving the Mobile Web for future generations.

2. BACKGROUND & RELATED WORK
National libraries, archives, and other memory organizations have worked for more than a decade to archive the Web. However, fundamental changes in web technology have created numerous problems for web archivists, including the growth of the Mobile Web. David Rosenthal summarized a recent workshop at the IIPC General Assembly 2012 which focused on problems of preserving the "future" Web:

> "But the clear message from the workshop is that the old goal of preserving *the* user experience of the Web is no longer possible. The best we can aim for is to preserve *a* user experience, and even that may in many cases be out of reach [15]."

Despite its many challenges, it is important that archivists preserve some semblance of the growing cultural artifact that is the Mobile Web.

In the pre-smartphone era, web pages designed for mobile devices (what we call *mobile pages*) were created with a variety of markup languages like C-HTML, WML, and XHTML-MP [18]. But as smartphones armed with higher bandwidth, more powerful processors and web browsers have been widely adopted in recent years, mobile websites have begun using the same markup languages used by the *stationary* (or traditional) *web*, namely XHTML and HTML5.

Today's smartphones and tablets have web browsers that are nearly as functional as desktop browsers, but they still suffer from latency issues, limited screen size and memory, and slower JavaScript engines which make viewing the stationary web problematic at times [19]. Usability experts also suggest altering the website experience for smaller mobile devices [11]. This is sometimes done by creating pages with completely different content or by using methods like responsive web design [10] which use media queries to format the page to best fit the targeted device's screen size.

Mobile pages are commonly served to mobile devices by examining the web browser's User-Agent in an HTTP request. Some websites will serve different content using the same URL to mobile user agents. Figure 1 illustrates how `cnn.com` will serve

a very dense web page to a Chrome browser running on a desktop machine, but it will serve a smaller and easier-to-navigate mobile page to the iPod's web browser. Other websites will redirect mobile browsers to different URLs that serve mobile pages. For example, a request to `http://yahoo.com/` on a mobile device will redirect the browser to `http://m.yahoo.com/`.

Figure 1. http://www.cnn.com/ on a desktop browser (back) and iPod browser (front)

Search engines have recently taken an interest in identifying mobile pages. Google reported in 2011 that they are crawling the Mobile Web using a variety of user agents from feature phones and smartphones in order to optimize the search experience for mobile users [9]. Yahoo! obtained a patent for identifying mobile pages using a number of methods including content and link analysis [14]. Our method for identifying mobile pages takes a somewhat simplified approach, focusing on URL analysis and content analysis with pages obtained using different user agents.

So far there has been little attention given to archiving the Mobile Web in the literature. Previous studies on the Mobile Web have examined the link structure of the Mobile Web [8], finding significant differences when compared to the stationary web's link structure. A pre-smartphone era study [18] characterized mobile pages in terms of characteristics like markup languages, page sizes, and image content. Other studies (e.g., [4][12][16]) examine usability issues of the Mobile Web. To the best of our knowledge, this is the first work that addresses the Mobile Web from a web archiving perspective.

3. FINDMOBILE TOOL

In order to help web archivists discover mobile websites, we created the Findmobile tool to automatically discover mobile pages for a website. The tool can be used by a web crawler like Heritrix to discover websites that have a mobile site in order to perform a crawl using a mobile user agent. Findmobile is initially fed a set of seed URLs that point to the root pages of websites. It uses three processors (useragent, urldiff, and mediaqueries) which use different methods to determine if a website is serving up mobile web pages or just standard web pages.

3.1 useragent

This processor initially makes two HTTP requests using user agents representing two of the most popular desktop and mobile browsers: Chrome for the desktop and Safari on the iPhone. We will refer to Chrome's user agent as *stationary* and Safari's as *mobile*. We realize that some websites may handle other user agents differently, but to reduce the total number of requests issued, we use just these two. PhantomJS [13], a headless WebKit browser stack, is used to make the HTTP requests. PhantomJS downloads all the resources making up the page (images, CSS, JavaScript, etc.) and executes any JavaScript. This is important because many websites use JavaScript to transform the page's structure, to download style sheets, or to redirect mobile web browsers to mobile pages. PhantomJS will also follow any 3xx HTTP redirects. Unfortunately, a known bug in PhantomJS stops some client-side redirects from working properly.

If the server or client redirects the mobile agent to a URL that is different than the URL the stationary agent is directed to, we assume the server is redirecting the browser to a mobile page. This page may not be significantly different than the page retrieved via the stationary agent, but we assume the behavior indicates the *intension* of the website to serve something different to mobile agents. If no redirect occurs, the web server might serve different style sheets to mobile user agents. So if no redirect is detected, the URLs for the web pages' style sheets are examined. If they are identical, the processor then examines the tag structure of the two pages since the structures are likely to vary significantly between regular web pages and mobile pages. The processor performs a tag frequency distribution analysis (TFDA) which computes a numerical score indicating the magnitude of the difference; a value of 0 indicates the structures are identical [3]. When non-zero values are calculated, it could be because the website was caught changing its content as some sites frequently do. Therefore the processor will make seven more requests using the stationary user agent in order to calculate additional TFDA scores. The tool can be configured to determine that a mobile page is detected for values over a particular threshold. In our experiments (next section), we occasionally detected extremely low TFDA values because websites would make very small and seemingly insignificant changes, like introducing a single tag, when requested with a mobile user agent.

3.2 urldiff

If the useragent processor was unable to discover a mobile page for a given website, the urldiff processor will attempt to discover other URLs that a website might be using for their mobile page. This processor uses the redirects discovered by the useragent processor to infer what the mobile page's URL might be. The algorithm used to infer mobile URLs uses the Diff, Match and Patch library [6]. Here are some examples of URLs that were created by urldiff when given `www.example.com`:

```
http://m.www.example.com/
http://www.example.com/m/
http://mobile.example.com/
http://www.example.com/?m=1
http://www.example.com/mobile/
```

After applying the transformation rules to the list of URLs, the processor requests the inferred URLs to see if valid pages are returned and uses the same technique as the useragent processor to

confirm that a mobile page was detected. Because this method often results in soft 404s [2], redirection to the site's home page or a search page, the processor uses Ben Hoyt's soft 404 detection script [7] to detect soft 404s. The script is not foolproof, and it sometimes produces false positives.

3.3 mediaqueries

This processor handles any URLs that passed through the previous two processors without detecting a mobile page. It looks specifically for CSS media queries which target mobile devices via their small screen size to indicate that the web page is targeted to a mobile device. A web page using media queries presents itself different on a small screen instead of a large one, but the HTML, CSS, etc. is identical. Therefore a web archive would not need to archive anything in addition to the stationary web page, but it might want to provide emulators for showing users how the page would render on a mobile device.

4. EXPERIMENT

We devised an experiment to test the Findmobile tool on well-known websites. We obtained a list of Alexa's top ranked 1,000,000 websites [1]. These are websites that Alexa Toolbar users frequently access and are therefore a good representation of popular websites. From this list we created two data sets: a random selection of 10K URLs (RAND) and the top 10K URLs (TOP). We fed the URLs from RAND and TOP into Findmobile in July 2012 and recorded *if* Findmobile found a mobile page and *how* it was able to make that determination. The experiment was performed again a few weeks later, and the findings differed little, so we present only the results from the first run here.

As mentioned in Section 3.1, Findmobile initially makes two HTTP requests for each URL, one with a stationary user agent (Chrome desktop) and one with a mobile agent (Safari/iPhone). When the experiment was executed, requests made by Findmobile that produced responses with error codes (4xx, 5xx, etc.), timed-out, or were garbled for either of the HTTP requests were ignored. So of the 10K URLs in the RAND data set, 9,342 of them produced a valid HTTP response for both requests, and 8,970 of the TOP URLs produced valid responses for both requests.

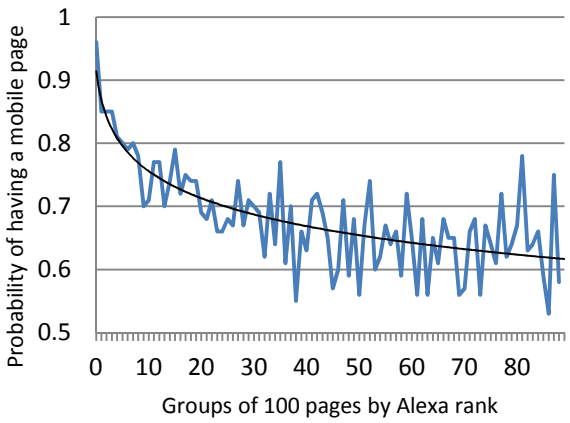

Figure 2. Relationship between Alexa rank and probability of having a mobile page

Findmobile found 50.2% of the 9,342 RAND sites produced a mobile page, and 68.5% of the TOP sites did. Since TOP has websites with higher Alexa popularity, this finding suggests that more popular sites are more likely to have mobile pages. To determine if there was a relationship between a website's Alexa rank and the likelihood of it having a mobile site, we grouped websites from TOP by Alexa rank in bins of 100 and then plotted the probability of each bin having a mobile site. The result, shown in Figure 2, shows a clear relationship between Alexa rank and the probability of having a mobile website. The top 100 sites have more than a 95% chance of having mobile pages, whereas the least popular sites in TOP have little better than a 60% chance of having a mobile site. It is perhaps not surprising that more popular websites have greater resources to produce mobile web pages.

Returning to the mobile page URLs that were discovered by Findmobile, we recorded *how* these mobile pages were discovered for both data sets and show the results in Table 1. Each of the methods in Table 1 is defined as follows:

- **Redirection** – the client or server redirected the browser to a new URL that contained the mobile page. For TOP sites, this was one of the most popular methods used.
- **Style Sheets** – the server responded with a mobile page that requested style sheets that were different from the stationary page. This was usually because the mobile page differed structurally from the stationary page (just like cnn.com in Figure 1). This technique accounted for many of the discovered mobile pages in RAND and TOP.
- **Adaption** – the server responded with a mobile page at the same URL that was requested by the stationary agent, and both the mobile and stationary pages used the same style sheets, but the pages were *structurally different*. The small numbers in this category in RAND and TOP are because most of the time when the structure differed, so did the style sheets (accounted for in the previous category).
- **URL Guessing** – the requests with the mobile and stationary agents returned identical content, so the mobile page was discovered by guessing its URL by the urldiff processor. Some false positives due to soft 404s likely inflated this category some, but many of the sites included links to these mobile pages from their stationary pages.
- **Media Queries** – the requests with the mobile and stationary agents returned identical content, but the mediaqueries processor discovered media queries were transforming the page for the mobile device. This technique was used very infrequently in both RAND and TOP.

Table 1. How mobile pages were discovered by Findmobile

Method	*RAND*	TOP
Redirection	15.9%	34.8%
Style Sheets	32.3%	25.1%
Adaptation	9.0%	11.7%
URL Guessing	38.2%	24.6%
Media Queries	4.6%	3.8%

If a mobile page was discovered, we wanted to know *how different* it was from its stationary web counterpart. Our analysis focused on the textual content of the web pages. In other words, we wanted to know if the pages said the same thing or something very different. The images, style sheets, and other resources were likely also different, but we limited our analysis to textual content and left further analysis to future work.

To measure the textual content differences, we calculated the Jaccard similarity coefficient between the stationary web page and mobile web page in both the TOP and RAND data sets. The results showed a bimodal distribution in both sets with a large

number of pages having nearly identical content (values close to 1) and a large number having very little in common (values close to 0). Figure 3 shows a histogram of the TOP (mean=0.457, median=0.281), and RAND (mean=0.522, median=0.459) data sets. TOP, which had far more mobile pages than RAND, had a larger percent of dissimilar pages than did RAND. Perhaps this is because more popular sites have more resources to dedicate to creating a unique, tailored mobile site.

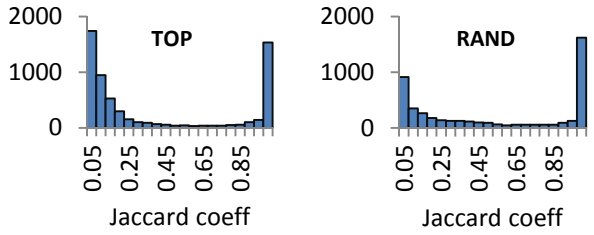

Figure 3. Distribution of TOP and RAND Jaccard coeff comparing the rendered text from stat. & mobile pages

When we manually examined a selection of the pages that fell in the two modes, we discovered pages where only the formatting differed on one extreme and pages that were entirely different on the other. For example, at the date of writing, the desktop version of MIT's web site (http://www.mit.edu/) is very different from the mobile version (http://m.mit.edu/), while the mobile version of Facebook (http://m.facebook.com/) has most of the same textual content as the desktop version (http://www.facebook.com/).

5. FUTURE WORK AND CONCLUSIONS

In this paper we have described our Findmobile tool which could aid web archivists with the task of archiving the Mobile Web. Through experimentation, we have demonstrated 1) how the tool discovered the proportion of techniques websites typically employ to direct mobile users to their mobile pages, 2) that popular websites are more likely to have mobile pages, and 3) that the textual content of a web page and its mobile page counterpart is half the time nearly identical and half the time widely divergent. These findings give archivists an idea of what to expect when charged with the task of archiving a website's mobile presence. It also suggests more work is needed to determine if archivists should invest more resources into archiving widely divergent web content or if mobile content that is almost identical to its stationary counterpart can be safely ignored.

There is much work to be done in scoping just those pages that make up the Mobile Web. Many sites have mobile-only pages that link to non-mobile pages, blurring the line between Mobile Web vs. Stationary Web, but Findmobile could be modified to help a crawler target mobile-only pages. Also, mobile pages are often tailored to users based on their location and other factors which leave plenty of work for archivists.

Our tool has been made available to the public, and we solicit feedback [5]. There is a need for more thorough evaluation of the methods used to discover mobile pages as some of the assumptions we made in detecting mobile pages do not always hold. We have started developing a new tool called IDmobile to augment Findmobile. It uses neural network classifiers to examine a web page's content and determine if it is designed specifically for a mobile device. Early tests show it as 80-90% accuracy detecting mobile pages in a limited domain.

6. ACKNOWLEDGMENTS

We would like to thank Adam Miller at the Internet Archive for his assistance in devising the Findmobile tool and feedback on our work. This research was supported by the National Science Foundation (IIS 1008492).

7. REFERENCES

[1] Alexa's top 1,000,000 websites. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

[2] Bar-Yossef, Z., Broder, A. Z., Kumar, R., Tomkins, A. 2004. Sic transit gloria telae: towards an understanding of the web's decay. In *Proceedings of the 13th international conference on World Wide Web*. ACM, New York, NY, USA, 328-337.

[3] Cruz, I. F., Borisov, Sl., Marks, M. A., Webb, T. R. 1998. Measuring structural similarity among web documents: preliminary results. *Electronic Publishing, Artistic Imaging, and Digital Typography, Lecture Notes in Computer Science*, 1375, 513-524.

[4] Cui, Y., Roto, V. 2008. How people use the web on mobile devices. In *Proceedings of the 17th international conference on World Wide Web* (WWW '08). ACM, New York, NY, USA, 905-914.

[5] Findmobile, http://www.harding.edu/fmccown/research/findmobile/

[6] Fraser, N. Diff, Match and Patch Library. http://code.google.com/p/google-diff-match-patch/

[7] Hoyt, B. 2008. Link rot, soft 404s, and DecentURL. http://blog.brush.co.nz/2008/01/soft404s/

[8] Jindal, A., Crutchfield, C., Goel, S., Kolluri, R., Jain, R. 2008. The mobile web is structurally different. *IEEE INFOCOM 2008 - IEEE Conf. on Computer Communications Workshops*. (Apr. 2008), 1-6.

[9] Kato, Y. 2011. Introducing smartphone Googlebot-Mobile. http://googlewebmastercentral.blogspot.com/2011/12/introducing-smartphone-googlebot-mobile.html

[10] Marcotte, E. 2010. Responsive web design. *A List Apart Magazine* (online - May 25, 2010). http://www.alistapart.com/articles/responsive-web-design/

[11] Mobile Web Watch 2012, Accenture, http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Mobile-Web-Watch-Internet-Usage-Survey-2012.pdf

[12] Nielsen, J. Mobile Site vs. Full Site. Jakob Nielsen's Alert Box (Apr. 10, 2012). http://www.nngroup.com/articles/mobile-site-vs-full-site/

[13] PhantomJS, http://phantomjs.org/

[14] Rao, S., Prabhakar, B., Seth, S., Murugesan, S., Gupta, A. 2011. US Patent No. 8041703.

[15] Rosenthal, D. 2012. DSHR's Blog. http://blog.dshr.org/2012/05/harvesting-and-preserving-future-web.html

[16] Schmiedl, G., Seidl, M., Temper, K. 2009. Mobile phone web browsing: a study on usage and usability of the mobile web. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI '09). ACM, New York, NY, Article 70.

[17] StatCounter Global Statistics, Mobile vs. Desktop for 2013, http://gs.statcounter.com/#mobile_vs_desktop-ww-monthly-201205-201304

[18] Timmins, P. J., McCormick, S., Agu, E., Wills, C. E. 2006. Characteristics of mobile web content. In 2006 1st IEEE Workshop on Hot Topics in Web Systems and Technologies (Nov. 2006). 1-10.

[19] Zakas, N. C. 2013. The evolution of web development for mobile devices. *Queue*. 11, 2 (Feb. 2013).