Characteristics of Mobile Web Content

Paul J. Timmins, Sean McCormick, Emmanuel Agu, Craig E. Wills
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA U.S.A.
Email: {ptimmins|mccorms}@wpi.edu, {emmanuel|cew}@cs.wpi.edu

***Abstract*— The World Wide Web is no longer tethered to our desktops and laptops. The Web has gone mobile, providing instant access to information anywhere and anytime. The mobile Web can be considered a shadow of the World Wide Web, implemented using specialized markup languages and design techniques adapted for comparatively limited mobile phones and PDAs. Despite the growing importance and usage of the mobile Web, surprising little is known about it.**

This paper presents the results of a study of mobile Web content conducted in May and June of 2006. The study examines the content of over one-million mobile Web pages from around the world using a search-assisted crawling methodology to locate and study pages for three of the most popular mobile Web formats—WML 1.0, WML 2.0/XHTML Mobile Profile (XHTML-MP) and Compact HTML (C-HTML). The objective is to study the relative characteristics of these mobile Web content formats, as well as compare them with a similar sampling of non-mobile (HTML) content.

We found that WML is the dominant mobile Web content type, although regional differences do exist. We found that all three mobile content types studied were on the same order of magnitude for average page characteristics such as number of links (under 10) and number of images (around 1), but pages in the newest format, XHTML-MP, are 50% larger on average than those in WML. Not surprisingly, all of these characteristics are much smaller than for HTML content pages gathered with the same methodology. In terms of specific features, only 7% of pages used WML cards, but 50% of XHTML-MP servers dynamically adapted the content served based on the user agent. Finally, we found less than 4% of mobile pages contained ad objects, which is much less than for HTML pages.

I. Introduction

The World Wide Web is no longer tethered to our desktops and laptops. Web content is increasingly available in mobile Web formats that facilitate information access by cell phones, PDAs, Internet connected watches and other portable computing devices. Mobile users can now access sports, news, stock charts and other Web content while on the move.

Content targeted at mobile devices is typically designed to mitigate the lower bandwidths of wireless networks as well as the reduced CPU and storage limitations of mobile devices. In order to maintain reasonable download times, Web designers reduce the size of mobile Web pages and the number of images per page.

A number of early Web characterization studies informed and influenced the development of the wired Web [1], [2], [3]. As the mobile Web develops, it is important to understand the characteristics of how it is being used. This characterization shall inform optimal choices for configuring network equipment and optimizing Web server parameters, such as buffer sizes and maximum number of incomplete TCP connections. PDA and cell phone manufacturers can also estimate minimum hardware specifications for future mobile devices. Understanding the nature of mobile Web content shall also drive more accurate simulations of Web content in the research community.

In spite of these benefits, surprisingly few attempts have been made to measure mobile Web content and quantify typical page and site characteristics. In this paper, we present the results of a large-scale study to characterize mobile Web content. The study examines the content of over one-million mobile Web pages collected during search-assisted crawls in May and June 2006. The study located and studied pages for three of the most popular mobile Web formats—WML, XHTML Mobile Profile (XHTML-MP) and C-HTML. We also use this same methodology to retrieve and measure wired Web content as a baseline of comparison.

We found that WML is the dominant mobile Web content type, although regional differences do exist—WML (WAP 1.0) is most popular in Europe and C-HTML (i-mode) is most popular in Japan. C-HTML crawling, and therefore C-HTML results, were significantly limited by the fact that many such sites are accessible only through NTT DoCoMo's "i-mode menu" service. We found that all three mobile content types studied were on the same order of magnitude for average page characteristics such as number of links (under 10) and number of images (around 1), but pages in the newest format, XHTML-MP, are 50% larger on average than those in WML. Not surprisingly, all of these characteristics are much smaller than for HTML content pages gathered with the same methodology. In terms of specific features, only 7% of pages used WML cards, but 50% of XHTML-MP servers dynamically adapted the content served based on the user agent. Finally, we found less than 4% of mobile pages contained ad objects, which is much less than for HTML pages.

This paper is organized as follows. Section II provides background on mobile Web technologies. Section III outlines the questions we intend to answer with this study and Section IV details the methodology used. The results of our study are presented in Section V with a summary of these results in Section VI. Section VII describes related work and Section VIII outlines potential future work as well as summarizes the findings of this work.

II. Mobile Web Background

This section reviews technical details of three of the most popular mobile Web technologies (WAP 1.0, i-mode and WAP 2.0) and compares them to the wired Web and HTML. These technologies can be distinguished by the devices that support them, their protocol stacks, and their markup language.

WAP 1.0 was introduced in 1998 as the first Mobile Web standard [4]. Several companies including Nokia and Motorola teamed up to develop the initial Wireless Application Protocol (WAP) protocol stack. It was envisioned that WAP 1.0 would enable a wide range of devices including mobile phones, laptops and PDAs, to send email and access the Web. Due to the limited resources of mobile devices, a lightweight, XML-based standard called the Wireless Markup Language (WML), was developed. WML also supports a "deck of cards" feature that allows the Web programmer to aggregate multiple related pages (cards) into a batch (deck). WAP 1.0 is connection-oriented and a mobile user has to make a telephone call to the web server while web pages are being downloaded.

The i-mode (information mode) system was created in Japan at about the same time as WAP 1.0. It was deployed by NTT DoCoMo, the Japanese mobile network operator, in early 1999 and is loosely based on the WWW protocols. The i-mode system allows its users to email, surf the Web, and exchange images but requires specialized mobile handsets. Pages in i-mode are programmed using compact HTML (C-HTML) [5], a markup language that is similar to HTML 1.1. Most i-mode sites are accessible through NTT DoCoMo's "i-mode menu" service, which limits access to most i-mode sites to paying customers.

The second generation WAP 2.0 was introduced in 2001. It was designed by the WAP forum to be backwards compatible with the WAP 1.X protocols and WML. It includes a lot of the features of i-mode and Internet protocols, as well as new features. Compared with WAP 1.0, which has a maximum speed of 9.6 kbps, WAP 2.0 operates at speeds of up to 384 kbps. It supports XHTML, a new markup language that was developed for a variety of low computing power devices such as televisions, vending machines, mobile phones, PDAs, and watches. XHTML-MP [6], a mobile profile, extends XHTML basic by adding features to enhance the Web experience on resource constrained mobile devices. WAP 2.0 is designed to run over packet-switched networks and supports both push and pull models of content access.

III. Study

With this background, the broad goal of the study is to understand the characteristics of mobile Web content as it is currently being used. This broad goal encompasses a number of specific questions that form the basis for the methodology used in this work. These questions and their rationale are:

1) *Format usage:* What is the relative usage of the three markup languages—WML vs. C-HTML vs. XHTML-MP? The answer to this question establishes the relative use of these three formats by content providers.
2) *Geographic distribution:* What is the geographic distribution of content in the three markup languages? It is important to understand not only how much, but where the different content types are being used.
3) *Page sizes:* What is the distribution of markup (base page) and total page size for each of the three formats as well as baseline HTML content? A standard characterization for Web content is to understand the size of pages in terms of the number of objects they contain, the number of servers these objects come from and the total number of bytes contained.
4) *Page connectivity:* What is the degree of connectivity in terms of the number of links for mobile pages and are these links internal to the same domain or to different domains? This question examines how the level of connectedness compares amongst the three content formats and with HTML content.
5) *Image content:* What are the characteristics of image objects in mobile Web content in terms of number on a page and size? Images are commonly used in wired Web content. It is important to understand how much they are used in mobile content.
6) *WML cards:* Are unique schema features, such as WML cards, used? WML content can be served as bundles of pages or "decks." An interesting question is to understand how much this feature is used.
7) *User agent adaptation:* To what degree do servers adapt the markup type of the content based on the User-Agent field of the HTTP request header? This question affords better understanding on whether users need to explicitly identify the needed content type or whether servers can and do make the appropriate transformation.
8) *Advertisement content:* What is the presence of advertisement content in the mobile Web world? Previous work found that ad content exists on the majority of popular pages [7] and we are interested to understand its use in mobile pages.

IV. Methodology

This study was conducted in two phases. In the first phase, mobile Web servers were crawled to find and retrieve mobile Web content. An open source Web crawler was modified to classify content as mobile, non-mobile HTML, image or other. Mobile and non-mobile content was retrieved in its entirety, whereas only the size and URL of images were obtained.

The second phase of this study analyzed the retrieved pages to measure page size distributions, connectedness, and design features. Key features of each page were summarized in a MySQL database, allowing detailed analysis through SQL queries.

A. Mobile Content Crawler

To find and retrieve mobile Web pages, we modified Larbin, an open source Web crawler `http://larbin.sourceforge.net/index-eng.html`, to create a "Mobile Content Crawler." Larbin provides a configurable and

extensible framework for Web crawling, with many options to control the crawler's behavior. The Mobile Content Crawler extends Larbin to identify mobile content, record page and image metadata to disk, (including HTTP response headers and content size), and retrieve the individual pages. The crawler was configured to use a 30-second delay between consecutive retrievals from a single server, with no delay for links to new servers. The effect of this is a preference on discovering new servers, but continued discovery of new pages within a site. Additionally, modifications were made to the Crawler so that it filtered non-mobile HTML content, recorded the image size/URL then discarded the image file, and ignored "robots.txt" (used by servers to prevent crawler access) so it could crawl search engine results.

In trial runs, Larbin was only modified to support mobile content, but not filter out non-mobile content, and seeded with a set of 14 starting mobile Web URLs, such as `mobile.espn.com` and `wap.yahoo.com`. These starting URLs were manually selected to represent a diverse population of content from a variety of mobile markup languages. During these trial runs, it was observed that a disproportionate number of HTML (non-mobile) Web pages were retrieved. As later results will show, this result is probably due to the higher connectivity (in terms of hyperlinks), of HTML content. The resulting effect was that retrieving HTML content reduced the number of mobile pages retrieved in that run.

To improve the crawler's capability to retrieve the desired type of content, Larbin was modified to filter non-mobile content and retrieve only pages that could be explicitly identified as being mobile. This filtering was based on the content-type response header and the document type, if present. To reduce the volume of data stored, Larbin was also modified to first download images and store the size of the image in the header, using a preprocessor directive.

In addition to encountering problems in filtering out non-mobile content, the trial runs also showed problems in using a small fixed set of starting URLs. The result were not as diverse as desired either in the content type or the subject matter. As a consequence a search-assisted strategy was employed.

B. Search-Assisted Crawling

To address problems encountered in the trial runs, a search-assisted crawling strategy was employed for this work. It was noted that Google's Mobile Web search engine (`mobile.google.com`) provides access to a large index of mobile web sites, and thus the search results as crawling starting points. This is similar to the strategy used in a previous study of Spyware [8]. Rather than directly select a set of starting URLs, the results of a Google Mobile Web search were used as crawling starting points, passing in specific keywords to the search engine. Google Mobile Web search allows searching of content by markup type (WML, XHTML-MP, or C-HTML). As a comparison, we also issued a search for HTML-based content using Google's standard search engine.

A number of keywords were used to obtain search results for each of the four types of content to seed our crawler. These keywords are shown in Table I, chosen to ensure diversity of search results. The upper portion of the table shows category-based keywords while the lower portion shows that we focused some searches on servers from specific country-based Top Level Domains (TLDs). These keywords are intended to obtain a broad set of pages for seeding.

TABLE I
SEARCH KEYWORDS USED FOR CRAWL SEEDING

news	sports	weather
games	portal	science
health	business	finance
arts	shopping	world
site:.jp	site:.uk	site:.ja
site:.au	site:.ve	site:.cn
site:.kp	site:.ca	

This study is based on four crawl runs done in May/June 2006, focusing on collecting HTML, WML, XHTML-MP/C-HTML, and C-HTML-only, all using the same keywords but different Google Mobile search restriction options. We found that, regardless of the search restriction option (ie: mrestrict=wml), crawl results are dominated by the most popular markup. Therefore, multiple runs were used, with later runs filtering out HTML and WML results. Additionally, the crawler was configured to report an appropriate User-Agent string in the HTTP request header, depending on the type of content we were attempting to crawl.

As with any crawling, the choice of starting points can bias the results. Search-engine assisted crawling, as used in [8], is biased by the search-engine's results. Fewer than 15% of servers were directly linked from the search results, with the remaining servers being crawled indirectly by following hyperlinks. This high percentage of indirectly-crawled servers lessens the impact of bias caused by the search results.

An early goal of this research was to additionally contrast content based on subject matter, such as comparing news versus finance content. This goal evolved into the search-assisted crawling strategy, with the assumption that content could be characterize based on search keyword. However it was observed that a high degree of pages associated with multiple search keywords, thus this goal was set aside for future research. Alternative crawler strategies, including shallower searches, might yield results that are distinguishable by keywords.

C. Identifying Mobile Content

Two techniques were employed to identify content as being mobile content. First, the Document Type Declarations (DOCTYPEs) identify the DTD for a particular XML document. DOCTYPEs are optional, however were present in over 75% of servers crawled. Secondly, the HTTP CONTENT-TYPE response header identifies the expected type, such as HTML or image content. WML content is identified with a CONTENT-TYPE of "text/vnd.wap.wml", and all other text content typically identified simply as "text/html".

TABLE II
PAGE, SERVER AND DOMAIN CONTENT TYPE STATISTICS

Type	Num. Pages	Num. Servers	Num. Domains	Avg Pages/Server
WML	1,055,589	13,672	5,734	77
XHTML-MP	145,314	842	446	173
C-HTML	14,206	27	26	526
HTML	227,462	47,110	38,143	5

Content was first characterized by the CONTENT-TYPE, then by the DOCTYPE. The DOCTYPE of each page was used to classify content into the following categories,

- WML: DTD WML
- XHTML-MP: "XHTML Mobile Profile" or "XHTML Basic"
- C-HTML: "Compact HTML"

A possible limitation is our method of identifying content based on the DOCTYPE XML tag, which is was only present in pages from 71% of servers. We observed the majority of the remaining pages contained "wap", "imode", "chtml", "wml" or "mobile" in their URLs, but did not contain the necessary DOCTYPE. Manual inspections of these pages indicated that the presence of these keywords in a URL did not necessarily indicate mobile content, and therefore such content was excluded from the study.

D. Analyzing Page Content

Once the pages were retrieved and data about them stored in a MySQL database then the last phase of our study gathered data needed to answer the study questions. To obtain the size of a mobile Web page, queries were written to retrieve image and page sizes. This result indicates the amount of information that is downloaded by a mobile browser. From a resource perspective, large Web pages would consume excessive amounts of memory, CPU, battery power and wireless bandwidth. They would also take too long to download. However, small pages may not contain all the information that a user wants, making it necessary to establish new TCP connections to download additional pages.

We also examine the number of links on each Web page and distinguish whether the target pages are located on the same domain (internal link) or on a different domain (external link). Computing the internal versus external link numbers for both mobile and wired content is a means to compare their degrees of connectivity.

Links and images were counted uniquely per page, therefore multiple links to the same URL counted as one link. As well, the total page size was computed by adding the markup size to the size of each image in the page. Only complete pages, pages where the size of each image was measured, are reported so as to not skew the results, and images were counted only once per page as most browsers will not retrieve the same image more than once per page. In addition, total page size was normalized to compensate for the fact that pages with images were less likely to be completely retrieved by summing the weighted average of the markup size of pages with no images with the markup and image size of pages with images.

V. RESULTS

This section presents results from our search-based crawling approach and subsequent analysis. The results are presented in a parallel order to the study questions posed in Section III.

A. Format Usage

Table II summarizes our crawl results concerning pages, servers and "domains" for each of the four formats. We define the domain of a server to be its 2nd-level domain[1] so servers such as `www.cnn.com` and `images.cnn.com` are each part of the `cnn.com` domain.

A key observation from Table II is that the number of WML pages found was an order of magnitude more than XHTML-MP, and and two orders of magnitude greater than C-HTML. This result could either imply that there are indeed more WML pages in existence, or the results could also be skewed by the chosen search strategy. The low representation of C-HTML content is presumably due to NTT DoCoMo's "i-mode menu" service, which provides paying customers access to pre-approved i-mode sites and thus is not accessible by the general public. Subsequent crawling runs using different search and filtering tactics were used to attempt to increase the number of XHTML-MP and C-HTML pages collected, including using Google's international web servers and filtering out WML results. These subsequent crawling runs did not significantly increase the number of pages.

Table II also provides the average number of pages crawled per server. Recall from Section IV, the crawler was configured to wait at least 30 seconds before page retrievals from the same web site, and would immediately retrieve pages from newly identified servers. The resulting crawler behavior is to prefer breadth over depth. Thus, the HTML results are expected: an average of only 5 pages per server, showing that the crawler was continually identifying and crawling new servers and thus is expected from the highest degree of external links (Table VI). For mobile content, C-HTML and XHTML-MP had much higher numbers of pages per server than WML, a fact not explained by the number of external links per page. This results can be attributed to C-HTML and XHTML-MP pages linking to a less diverse set of servers than the WML pages.

[1]In cases where the Top Level Domain (TLD) is a country code and the TLD is subdivided using recognizable domains such as "com" or "co" then the domain of a server is its 3rd-level domain.

B. Geographic Distribution

Tables III and IV summarize the percentage of unique Web servers by top level domain, which provides some means to understand the geographic distribution of the gathered pages. The international flavor of the results show that the usage of mobile content is global. As expected, WML is more popular in Europe and China where WAP is mostly used. The breakdown of C-HTML sites by top level domain is not reported, due to the significantly fewer sites included in the study.

TABLE III
TOP LEVEL DOMAIN BREAKDOWN FOR WML CONTENT

Domain	% of WML Servers
.com	30%
.ru (Russia)	22%
.cn (China)	13%
.net	8%
.hu (Hungary)	3%
.de (Germany)	2%
.org	2%
.cz (Czech Republic)	2%
.uk (United Kingdom)	2%
other	18%

TABLE IV
TOP LEVEL DOMAIN BREAKDOWN FOR XHTML-MP CONTENT

Domain	% of XHTML-MP Servers
.com	47%
.ru (Russian Federation)	15%
.net	10%
.jp (Japan)	4%
.de (Germany)	3%
.ch (China)	2%
.no (Norway)	2%
.cn (Canada)	2%
other	14%

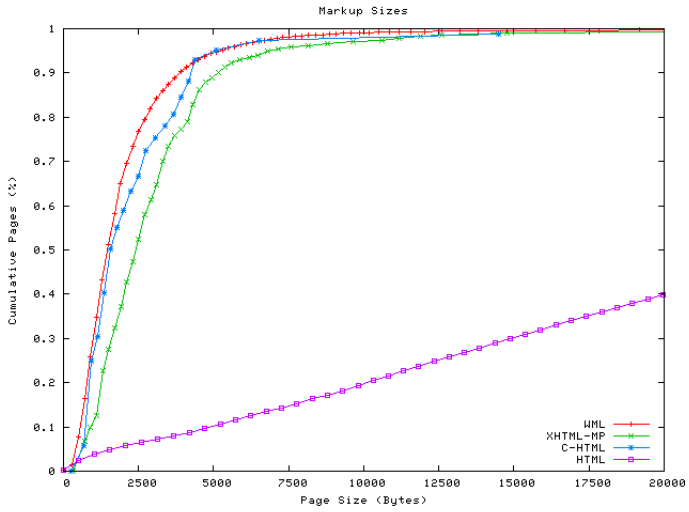

Fig. 1. Distribution of Content (Markup) Sizes

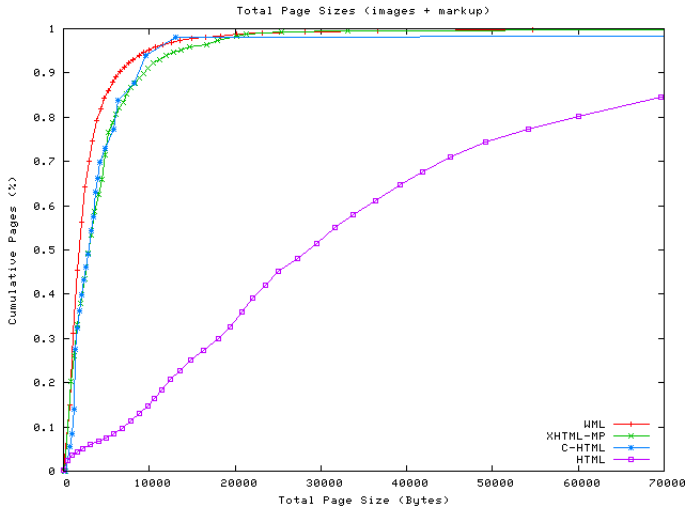

Fig. 2. Distribution of Total Page Size (Images + Markup)

C. Page Sizes

Figures 1 and 2 show the distribution of page sizes, in terms of markup alone and total page size, emphasizing the size difference between HTML and mobile content. Total page size was only reported for pages where all images were collected, so that partial page sizes were not reported.

Table V shows average sizes of mobile markup, as well as the average total (markup + images). Total size counts each unique image in a page exactly once, but background images are not included. WML markup objects are the smallest on average with 2,159 bytes. As a comparison, results published in 2003 report an average of size of 1,230 bytes [9]. XHTML-MP markup objects were larger than WML pages by 40% on average, at 3,018 bytes. C-HTML markup objects were close on average to those of XHTML-MP type at an average size of 2,911 bytes. As is expected, HTML markup objects are the largest, by an order of magnitude at an average size of 35,490 bytes. As shown in the last column of Table V, the relative results for the average total page size are comparable for all four content types, although the sizes of XHTML-MP and C-HTML pages are more than 50% larger on average than WML pages.

Markup and image sizes are important considerations for content providers aiming to provide acceptable performance for their users. By providing content that is at or below average size, content providers can ensure that users will not suffer from abnormally long network transmission delays or memory consumption problems arising from large content. Mobile Web browsers and platforms should also be designed with expected content sizes in mind to ensure adequate memory and resources are provided to allow pages to be retrieved and cached. Results show that page sizes for the newer XHTML-MP format are an order of magnitude less than for HTML, but more than 50% larger than for WML.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.