The Influence of Client Platform On Web Page Content: Measurements, Analysis, and Implications

Sean Sanders, Gautam Sanka, Jay Aikat, and Jasleen Kaur

University of North Carolina at Chapel Hill

Abstract. Modern web users have access to a wide and diverse range of client platforms to browse the web. While it is anecdotally believed that the *same* URL may result in a different web page across different client platforms, the extent to which this occurs is not known. In this work, we systematically study the impact of different client platforms (browsers, operating systems, devices, and vantage points) on the content of base HTML pages. We collect and analyze the base HTML page downloaded for 3876 web pages composed of the top 250 web sites using 32 different client platforms for a period of 30 days — our dataset includes over 3.5 million web page downloads. We find that client platforms have a statistically significant influence on web page downloads in both expected and unexpected ways. We discuss the impact that these results will have in several application domains including web archiving, user experience, social interactions and information sharing, and web content sentiment analysis.

Keywords: Web Page Measurement, Mobile Web, Content Analysis

1 Introduction

Users have many choices of *client platforms* — browsers, operating systems, devices, and vantage points — that can be used to request web-based data. Although, it is known that certain client platforms such as device type (e.g., smartphones or laptops) and vantage point can have an influence on the base HTML page that is downloaded [14,26], the extent to which this occurs has not been studied before. Any difference in base HTML pages that is due to client platform can result in data that is incomplete or view-specific. This can present issues for several web-related applications, such as web archival [13,21], document summarization [22,9], and information sharing, because additional care must be taken when (i) designing experiments that yield complete and/or unbiased data and (ii) developing processing scripts that are robust to different web-page designs— the need for understanding these differences has also been recently discussed in [2].

In this paper, we ask the question — *to what extent do different client platforms influence the content of a base HTML web page for the* **same** *URL request?* We perform the *first* measurement study that aims to understand this influence. Our methodology includes collecting measurements across different browsers (Opera, Internet Explorer, Google Chrome, Firefox, and Safari), operating systems (Mac OSX, Windows, Linux, iOS, and Android), devices (laptops, tablets, and smartphones), and vantage points (13 planetLab nodes located in 8

different countries) — this includes over 3.5 million measurements obtained from 3876 unique URLs composed of the top 250 web sites collected over a period of 30 days. We extract both HTML tag-based and content-based features from this data and find differences in web pages across different client types that are both practically and statistically significant. Our key findings are:

- *Expected and Unexpected Results:*
  1. As expected, device type (smartphones, tablets, and laptops) has a significant impact on web page content, with smaller devices being returned leaner pages. However, there is no consensus among current web designers and content providers on which type of page should be designed for tablets (i.e., should tablets simply return default laptop pages, mobile optimized pages, or have a special type of page altogether). An unexpected result is that, surprisingly, the manufacturer of a device, say an iPad Tablet or a Galaxy Tablet, may impact the type of page that is downloaded, say a default laptop or mobile optimized page.
  2. The differences that we find across different browsers are largely unexpected. For example, we find that different browsers may provide different default number of comments to be shown in a comment section for news articles and social media sites. We also find that content providers handle outdated browsers in multiple ways including: 1) fail to fulfill the web page request; and 2) fulfill the web page request by sending a similar, but different, web page that is likely more compatible with the user's browser version (e.g., sending a mobile-optimized page to an outdated laptop browser).
  3. As expected, we find that vantage point has a modest influence on web pages. For example, some content providers provide international versions of web pages that is dependent on the country of a user's vantage point while others provide the same content irrespective of vantage point. An unexpected result is that search results are highly influenced by vantage point, even for search queries where vantage point is not an obvious contributing factor to the result set.
- *Implications of Results on Web-related Applications:* Differences in web pages across different client type have implications in several web-related application domains including web archival [23], document summarization [22,9], sentiment analysis [16,22], and web browsing/systems design [14]. Some examples include: 1) the number of default comments and/or product reviews provided on a page is influenced by client platform— this may impact document summarization and sentiment analysis techniques that leverage this information; 2) web page designs may be client platform-specific which influences the type of content that is available and the effectiveness of parsing scripts that is targeted for a specific page design. This can also be problematic for sharing information on social media because hyperlinks may be client platform specific (hence users may be referring to different content and/or formatting context).

The remainder of this paper is organized as follows. We present our methodology in Section 2. The results and implications of our analysis is provided in

Section 3. Related work is presented in Section 4 and a summary of our study along with intended future work is provided in Section 5.

2 Methodology

Our methodology consists of two components: 1) Data collection and 2) Statistical analysis. We describe these two aspects in this section.

2.1 Data Collection

Selection of Web Pages to Study: In this study, we target web pages that are comprised from the top 250 web sites of the world according to Alexa [1] — a recent study shows that 99% of web requests comprise the top 250 web domains [6]. We manually browse each of these 250 web sites to obtain a diverse sample of URLs from each. Our web page sample includes landing pages, video streaming pages, search result pages (e.g., web, image, and news search), mobile web pages, clickable content, audio streaming pages, and social networking pages. We do this manual browsing for URL collection instead of leveraging a web crawler in order to better control the diversity and representativeness of our dataset. In total, we collect a list of 3876 unique URLs that are used to drive our data collection procedure.

Client Platforms Used: We next select a diverse set of client platforms, that are used to download the web pages we previously identified. As noted before, we intend to study the impact that browsers, operating systems, devices, and vantage points have on base HTML pages. We control for these different client platforms by requesting web pages using an User-Agent string that corresponds to the appropriate client platform of interest. User-Agent strings encapsulate the operating system, browser type, browser version, and even hardware information about client platforms — content providers use this information when responding to web requests [14]. User-Agent strings can be easily set by using scripts (we use python for this) to download base HTML pages. Table 1 lists the 32 User-Agents used for our study[1].

Our definition of "client platform" also includes location (vantage point). Thus, we also download web pages from different vantage points around the world — we use the PlanetLab network for this [8]. The 13 planetLab nodes that we use are located in Australia, China, Japan, Brazil, Poland, Canada, and the United States (7 nodes — Oregon, Rhode Island, California, Florida, New Mexico, Kentucky, and Ohio).

Repeated Measurements: Modern web content is highly dynamic and may change multiple times a day [11]. We take repeated measurements of each web page across each client platform to eliminate differences in web page content observed across client platform, that are likely simply due to variation over time.

[1] Please note that each of the User-agents we use in this study were obtained from deep packet inspection of web traffic as generated using known client platforms.

Table 1: Overview of User-Agents Used for Web Page Requests

Operating System	Browser(s)	Device
Windows 7	Chrome 38.0.2125.122 - Chrome 33.0.1750.154	Laptop
Windows 7	Firefox 33.0 - Firefox 26.0	Laptop
Windows 7	Internet Explorer 11.0 - Internet Explorer 9.0	Laptop
Windows 7	Opera 25.0.1614.68 - Opera 12.16	Laptop
Windows 7	Safari 5.1.7	Laptop
Windows 8	Chrome 39.0.2171.95 - Firefox 32.0	Laptop
Windows 8	Internet Explorer 11.0 - Opera 24.0.1558.61	Laptop
MacOSX 10.6.8	Chrome 39.0.2171.65 -Firefox 33.0	Laptop
MacOSX 10.6.8	Safari 5.1.9-Opera 25.0.1614.71	Laptop
MacOSX 10.9.4	Chrome 38.0.2125.122-Firefox 33.0	Laptop
MacOSX 10.9.4	Safari 7.0.5-Opera 25.0.1614.68	Laptop
Ubuntu	Firefox 34.0	Laptop
Solaris	Firefox 17.0	Laptop
Fedora	Firefox 2.0.0.19	Laptop
Android 4.4.4	Chrome 37.0.2062.117	Motorola Smartphone
Android 4.4.2	Samsung SM-T230NU-Chrome 35.0.1916.141	Samsung GalaxyTablet
Android 4.4.2	Amazon Silk 3.37	Fire Tablet
iOS 7	Safari 8.0 Mobile/12B41	iPhone Smartphone
iOS 7	Safari 7.0 Mobile/11A501	iPad Tablet
iOS 7	Safari 8.0 Mobile/12A405	iPod Touch
iOS 3	Safari 4.0 Mobile/7D11	iPod Touch

Specifically we take 30 repeated measurements over a period between December 18, 2014 and January 18, 2015. Thus, our dataset includes 3,771,348 page downloads.

2.2 Statistical Analysis:

Overview of Features: We extract different types of quantitative features from our HTML data to describe the properties of the downloaded web pages. A brief overview of the types of features that we extract are provided below:

HTML Tag-based features are used primarily for the analysis of page formatting — these have commonly been used in other HTML-based analysis [3,7]. In particular, we count the occurrence of several HTML tags/attributes that are present on a given web page. These tags represent different established categories of HTML information [3][2]. Our feature set includes:

1. *Flow content:* Used within the body of HTML documents (e.g., "table", "form", "option", "text area", and "menu" tags)
2. *Sectioning content:* Used to partition HTML documents (e.g., "area", "article", "body", "div", and "section" tags)
3. *Heading content:* Used for header-level markup (e.g., "header", "title", and "meta" tags)
4. *Phrasing content:* Used for text-level markup (e.g.,"abbr","b", "p", "strong", and "span" tags)

[2] Please refer to [3] for a complete list of these features.

5. *Embedded content:* Used for elements that load external resources into the HTML document (e.g., "script", "image", "audio", "embed", "param", and "iframe" tags).

Count statistics that are derived from tags that represent (i) hyperlink-level information (e.g., "a" and "link" tags) and (ii) the *extensions* of embedded objects that is referenced by a page (e.g., .jpeg, .gif, and .png extensions for embedded image objects) are used for *Object-based features* and analysis. We also derive *Content-based features* from our HTML data. We use a simple bag-of-words model to count the frequency of all the words that are present in a document — bag-of-words models are commonly used in natural language processing, machine learning, and computer vision [25]. A *word* in this model is defined as any sequence of characters that is present in an HTML document that is delimited by >, <, ", newline, or whitespace characters. This model allows us to derive features that can measure the overall text-related differences between two documents. We derive features such as (i) the number of words that are shared between two documents (i.e., a baseline document and a test document), and (ii) the number of words that are different between two documents to compactly represent these content-related differences. We use these features simply as a measure to flag significant differences in text for further analysis.

While we are able to obtain a lot of information from base HTML files we are unable to collect all of the information that is referenced by a particular web page. This is because modern web pages make significant use of AJAX and scripting technology. It is nontrivial to extract features that are derived from this information using base HTML pages alone. An analysis of the network traffic generated by web page downloads is needed to obtain this data. Such traffic analysis is beyond the scope of this paper.

Statistical Analysis Procedure: In order to determine which of our 134 features differ significantly across web pages downloaded using different client platforms, we use a standard non-parametric statistical test. The use of a non-parametric test allows us to make minimal assumptions about the distribution of these features. In particular, we use the *Kruskal-Wallis* test to determine whether there is a statistically significant difference between the measured web page samples across multiple appropriate groups of client platforms for each feature. The Kruskal-Wallis test yields p-values that represent the statistical significance of each feature for different client platforms. Here, lower p-values correspond to results that have greater statistical significance. We then use these results to dig deeper into our dataset to (i) determine the source of any significant difference and (ii) discuss the practical significance of our findings.

3 Results

Impact of Browser Platform We first investigate the impact that different browser platforms have on web page content. We initially focus on the influence of different browser platforms installed on the *same* operating system. In particular, we compare the latest versions of the Internet Explorer, Chrome, Firefox, and Opera browsers that correspond to the Windows 7 operating system — refer to

Table 1 for more details about these browsers. The Kruskal-Wallis test for this feature group yields 8 features that have p-value $< .05$ across browser platforms — in fact, these p-values are generally less than 10^{-3}. These 8 statistically significant features are: the number of "label" tags, the number of "tr" tags, the number of "table" tags, the number of "td" tags, the number of "style" tags, the number of "legend" tags, javascript length (i.e., the number of characters present in script tags), and the number of different words present. Upon further analysis of our data, we find that these statistically significant features correspond to the following trends:

- *Differences in javascript:* We find that many content providers such as soundcloud.com and bing.com (particularly image search results) use different javascript code that is suited for particular browsers — these javascript related differences were identified by the number of different words feature. We find that different javascript methods are implemented differently across browser platforms and/or have conditional statements that branch for different client browser platforms. For example, soundcloud.com uses conditional statements that takes the client platform into account during javascript execution to determine whether HLS (HTTP Live Streaming) is supported by the client platform. Alternatively, Figure 1 shows an example where a Youtube.com page has javascript that is browser-specific — here the Chrome javascript for loading a video appears to be HTML5-based while the Firefox javascript appears to be flash-based (This is identified by the "swf" references in Figure 1). It is known that if different client platforms are not taken into account, rendering differences across browsers can occur when the same source HTML is processed — for example, target.com has differences in rendered tables across browsers despite having rendering the *same source code* that renders that portion of the page.
- *Ads:* We also observe "ads" that attempt to get a user to download a particular browser or app that is browser dependent. For example, yahoo.com recommends that users update to the latest version of firefox for non-firefox client platforms, whereas target.com recommends that users on the Chrome browser to download their custom app. These ads seem to be attempts to get users to utilize software that is fully supported by the content provider.
- *Reduced comment and recommendation sections:* Our data also shows that cbssports.com and yelp.com do not provide the same number of comments, recommendations, news feeds, or search results for each browser. The limited information provided by certain browser platforms provides inconsistent data for document summarization and sentiment analysis applications [16,22,9] which can yield misleading and/or incomplete results, depending on the specific choice of browser platform. This limited information also impacts user experience because it may require users to take additional actions, such as a click, to view additional content that may be more readily available (already loaded) on a different browser.

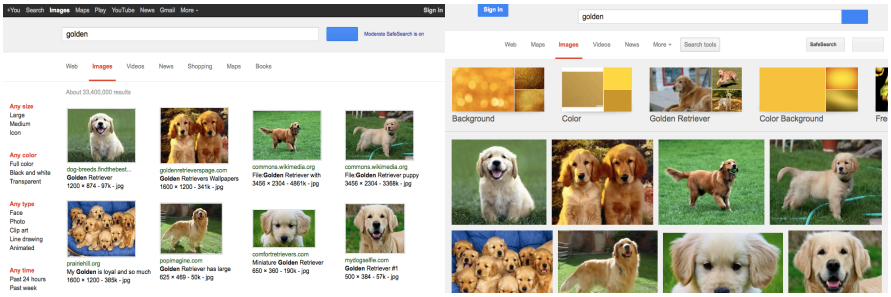
Impact of Browser Version We next compare the impact that browser *version* may have on base HTML pages. Our statistical test yields 13 statistically

```
<!--Chrome Version-->
ytplayer.load = function() {
yt.player.Application.create("player-api",
ytplayer.config);ytplayer.config.loaded = true;};

(function() {if (!!window.yt && yt.player &&
yt.player.Application)
{ytplayer.load();}})();</script>
<div id="watch-queue-mole" class="video-mole mole-
collapsed hid">

<!--Firefox Version-->
swf = swf.replace('__flashvars__',
encoded.join('&'));document.getElementById("player-
api").innerHTML = swf;ytplayer.config.loaded =
true}());</script>
<div id="watch-queue-mole" class="video-mole mole-
collapsed hid">
```

Fig. 1: Example where javascript is different for different browsers (Chrome vs Firefox).

significant features. The most notable features that are not also influenced by browser platform, say Safari vs Firefox, are the number of script tags and the number of HTML5 tags. With respect to the number of script tags, we observe similar differences in scripting behavior as we did with the differences in browser platform. With respect to the number of HTML5 tags, we observe that there tends to be more HTML5-related tags for the latest browser versions as compared to the older versions — we believe this to be a compatibility-related issue.

We also observe cases where content providers treat outdated or unsupported browsers in the following 2 ways. First, the content provider can fulfill the web request, but provide a warning to the user that their browser needs to be updated (zillow.com, soundcloud.com) — this may also result in failed web requests. Second, the content provider can fulfill the web request by responding with a web page that is compatible with the user's browser. This is explained in detail next.

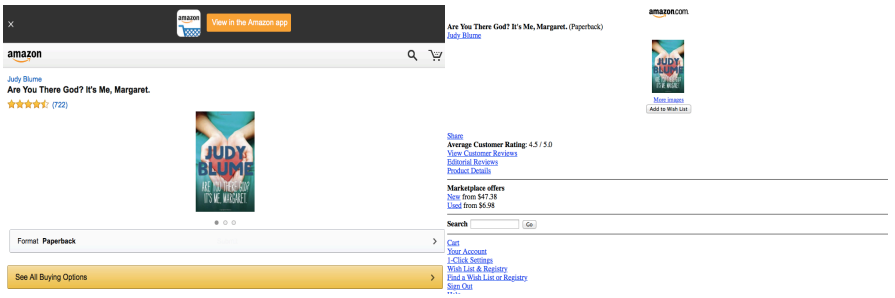


(a) **(b)**

Fig. 2: Different HTML pages are returned when an old version of Opera is used (a) in place of a current version of Opera (b).

We find multiple instances when browser version has an impact on page content. For example, Figure 2 shows that a Google search result that is rendered using an outdated Opera browser (Figure 2 (a)), and an up-to-date Chromium-based Opera browser can be displayed differently (Figure 2 (b))— though these observed differences are almost purely stylistic with respect to image size and visibility of URLs on images. Figure 3 shows a different example of when a web server responds with a web page for an outdated browser. Here, the web request is for a mobile web page of a product on Amazon.com. Figure 3 (a) shows that when a mobile web page is requested using an *up to date mobile device and*

browser (an iPhone in particular), the request is satisfied as expected. When we make the same request for a mobile web page using an *outdated Firefox browser on a laptop* we also get the *same* mobile web page — though we do not observe an ad for downloading an app. Figure 3 (b) shows that when the same request is made to Amazon using an *up-to-date Firefox browser on a laptop* we get a *different* mobile web page that is clearly representing the same product shown in Figure 3 (a). It is clear that these downloaded web pages are both (i) mobile-optimized web pages and (ii) different, where the version of the page shown in Figure 3 (b) appears to be an older mobile web page design than the page shown in Figure 3 (a). We conclude two things from these observations: 1) mobile web pages may sometimes be used to fulfill web requests to outdated browsers (we observe similar behavior for yahoo.com and att.com); and 2) interesting and unexpected quirks exist for some web page requests that are influenced by browsers[3]. The impact that browser version has on web page downloads is important for web crawling tools because (i) web crawlers may be used for years without receiving any significant upgrades and (ii) content providers may respond to known web-crawler User-Agents in a manner that results in errors or downloading data that is limited (in a manner similar to mobile web pages) [21].

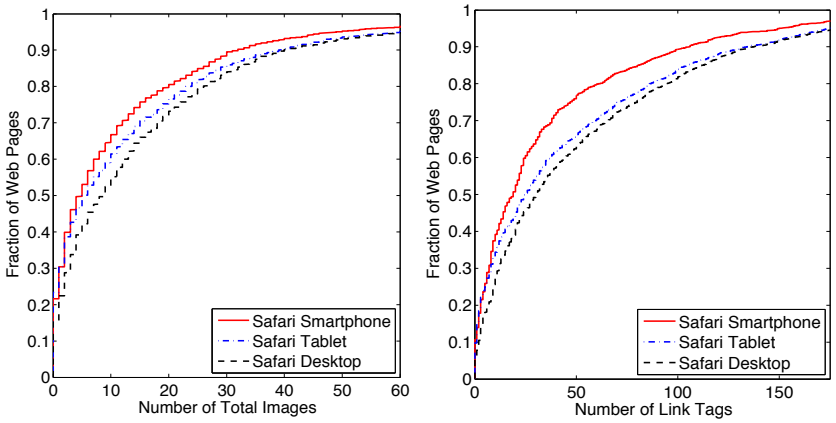


(a) **(b)**

Fig. 3: Content providers can return different pages to account for different browser versions — (a) current mobile browser and (b) current desktop browser.

Impact of Operating System For the purposes of analyzing any implications that operating systems may have on base HTML pages, we compare all laptop-based browsers across each operating system that also has the *same* version of that particular browser. For example, Firefox 33.0 is compared across MacOSX 10.9.4 and Windows 7. We do this for all combinations of browsers where this is valid according to the User-Agents we tested in Table 1. We do not find any statistically significant features that occur for the *same* browser across *different* operating systems. We conclude that browser version and browser type has a much bigger impact on web page downloads than operating systems.

[3] Please note that the significant differences discussed here are primarily true for browser version analysis for Opera and Firefox. This is because we have the largest range in release dates for these two browsers.



(a) **(b)**

Fig. 4: Device type has a significant impact on HTML features.

Impact of Device Type We next study the impact that different devices have on web page downloads. We start by focusing on comparing the iOS 7 iPhone smartphone, iOS 7 iPad tablet, and the MacOSX 10.9.4 laptop where each device runs a version of Safari. We find that:

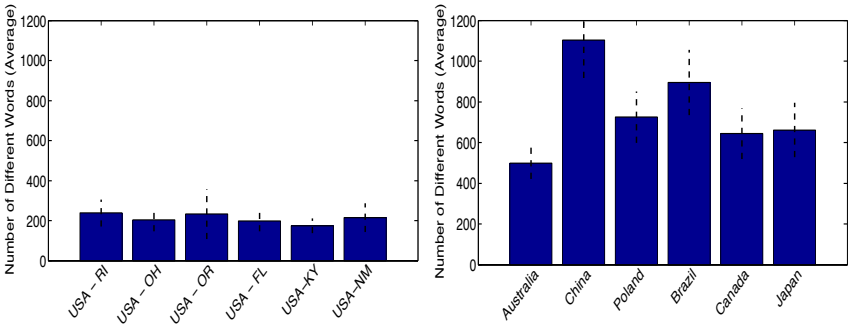
1. *Device type has a statistically significant impact on web page downloads:* As can be expected, devices have a statistically significant impact on many features (67 total) by design intent — pages designed for the small screens of mobile devices are likely to have simpler and smaller content. The most prominent features that differ across phones, tablets, and laptops are embedded object-related features such as the number of images, scripts, and CSS references found in an HTML source, content-related features such as the total number of words present on a page, and the total number of links — all of these features have p-values that are on the order of 10^{-10} or less.
2. *Lack of consensus on the design of tablet-specific web pages:* Figure 4(a) shows the cumulative distributions of the number of images and Figure 4(b) shows the number of link tags stratified by device type. The smartphone and laptop devices tend to exhibit the fewest and largest number of features respectively. Tablet devices behaves in the middle, where it is similar to a mobile device in some cases, and then slowly transitions to be similar to the laptop device in other cases. This behavior of tablet devices is attributed to the lack of a consensus among content providers on the design of web pages for tablets. Content providers tend to either (i) have a unique web page design for each device type (e.g., 163.com) (ii) leverage the similar web page design for both laptop and tablet devices (e.g., imdb.com), or (iii) leverage the similar web page design for both tablet and smartphone devices (e.g., twitter.com). We also find that different tablet manufacturers may receive different web pages. For example, android devices may receive ads to download android apps where iOS devices will receive ads to downloads apps on the Apple store. More interestingly, we find that the Amazon Fire

Tablet will receive a smartphone version of a web page (espn.com) where the iPad Tablet will receive the desktop version of the page — this suggests that screen size is a more important factor in the page that is downloaded than simply referring to the device as a Tablet or smartphone.

3. *Inconsistent redirect behavior that is based on device type across content providers:* We also find that there is a lack of consistency in the device-triggered redirect behavior across content providers. For example, some content providers will redirect mobile web page requests made by laptop clients to its corresponding laptop-based web page, while other content providers will not redirect requests in such a manner. This observed redirect behavior for devices is similar to the redirect behavior we observed for browser versions. This behavior can be problematic for a number of web-related applications. For instance, web crawlers may be redirected from the mobile view of a web page to the laptop view of a web page (in an undesired manner). This has an impact for web page archival because undesirable or even less informative views of a page (mobile or desktop) may be archived instead of the desired page. This also raises concerns for information sharing across social media (e.g., search engines and social networking) because users can be referring to *different* views of information, or, at times, entirely different information altogether, via the *same* hyperlink. For example, if a user shares a link on a social media site, say Facebook.com, and a friend uses a different client platform to view it, the two users could be observing different content (especially comments and recommendations listed on a page). This can be particularly difficult if one user is referring to a particular comment or review on a page that is not immediately viewable by another user.
4. *Different search result sets for web search queries:* Device type is taken into account by web search engines such as bing.com and google.com when returning search results. We find that generally, smartphones tend to have more mobile optimized web pages included in a search result set than tablet and laptop devices — this is because search providers take into account the mobile-friendliness of a web page when providing search results [4]. We also find that the search result set may have different meaning on different devices — this is mainly because search engines are increasingly providing web content to users instead of simply links to pages. For example, the search result set for the "nba standings" search query yields a different order of the basketball team rankings for a smartphone and a laptop (division rankings vs conference standings). This further underscores the impact that device type can have on information sharing and other applications because a user may refer to portions of a page, say the rank of a basketball team, where a friend does not immediately see the same ranking that is being referenced.

Impact of Vantage Point We next discuss the impact of vantage point on base HTML pages. We find that:

1. Our statistical analysis shows that *none* of the HTML tag-based features are significantly impacted by vantage point. This result shows that web page design and formatting is not significantly influenced by location. This includes



(a) **(b)**

Fig. 5: Impact of vantage point on number of different words (U.S.(a) and World(b)). The baseline for comparison located in California.

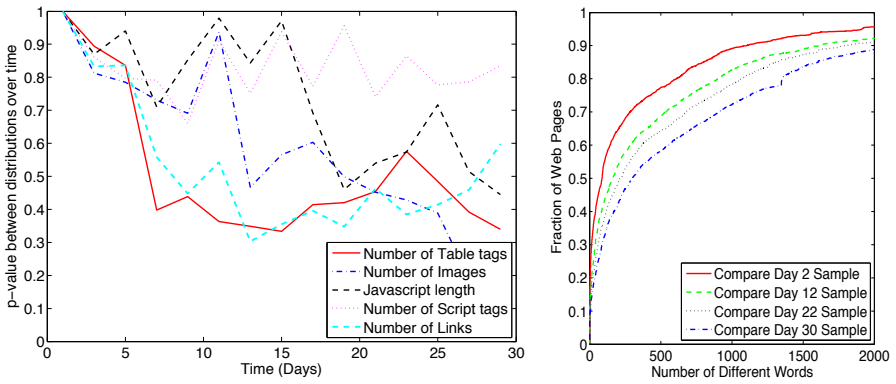
locations across different continents, which is surprising given cultural preferences in content layout and appearance.

2. Vantage point has a significant impact on content-related features. Figure 5 (a) shows that the average number of *different* words for each vantage point in the U.S is roughly 200-250 words, while Figure 5 (b) shows that the average number of different words for each vantage point that are outside of the U.S is over 500[4] — Figure 5 (a) and Figure 5 (b) both include 95% confidence interval bars around the average. Most of these differences across all vantage points (both U.S-based and world) correspond to (i) differences in topics of local interests, (ii) differences in search result sets, and (iii) temporal changes (discussed later). We observe a larger difference for vantage points around the world mainly because content providers have international versions of content that is likely to be of interest to the local population (cnn.com and yahoo.com does this). We also find that international web pages may include notes concerning (i) privacy awareness about the use of cookies on web sites and (ii) options to view the U.S version of web pages.
3. Bing search results, whether it is web, news, or image search, may yield different links, ads, and images across different vantage points — please note that we verified that this is not primarily a consequence of time[5]. Some of these differences are obvious due to location-based searches, say when a user is searching for McDonalds, and the search engine returns the address of the nearest McDonalds. Other differences are more complex, such as when more generic and random search queries such as "a hello berry" and "golden" yield different search results. The impact of vantage point on search results is important to note because search engines are a primary tool for various applications including web page scraping [17] and web security [19]. Vantage point driven search results also impact users because location can be

[4] Please note that while we study the top 250 web sites in the world, many of these sites are served by content providers that are in the U.S.

[5] We discuss results pertaining to bing.com because other search engines such as google.com are blocked in some countries.

misleading for users who access the web via 3G or 4G services — thus, the wrong location can be used to target search results.



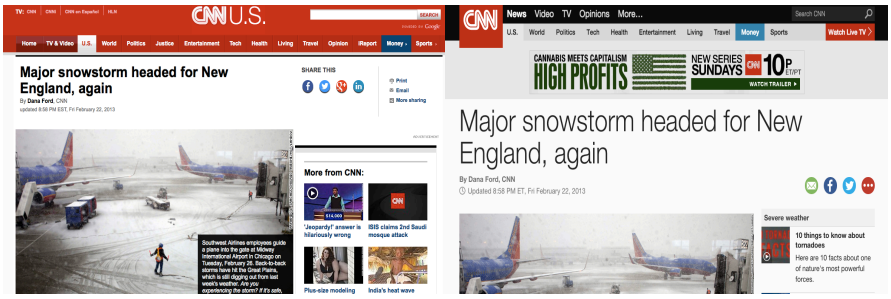
(a) **(b)**

Fig. 6: Impact of time on various HTML features (tag-based features (a) and content-based features (b)).

Impact of Time Lastly, we investigate the impact that time has on base HTML source files. We perform many univariate Kruskal-Wallis tests between our first measurement (i.e., baseline measurement) and each subsequent measurement. Figure 6(a) shows a time series plot of the p-values for these statistical tests for different tag-based HTML features for the Chrome browser. Figure 6(a) shows that the tag-based features that were statistically significant for some of our prior analysis (e.g., number of images and number of script tags) do not vary significantly (i.e., p-value below .05) over time. In fact, all of the tag-based features that we examine do not change significantly over time. These results imply that the differences across browser and devices that we observe are not significantly influenced by time. This further validates our tag-based findings because it shows that our results are not likely to be due to randomness. We do find rare cases where web page design has changed over time. For example, Figure 7 shows that the format for CNN web pages changed during our data collection procedure. We observe the new format for the CNN page (Figure 7 (b)) for all browsers and devices and conclude that CNN made this format change in order to serve a single web page that adapts to various screen sizes instead of serving multiple web pages to different device types. We also find that Overstock.com will display different versions of a page, one that includes product recommendations and another that does not, at different points in time (we find similar results for zillow.com with respect to content recommendations and imdb.com with respect to ads that completely change the layout of a page). We observe these differences over several browsers and believe that product recommendations are missing at certain instances in time for performance reasons — dynamically generating pages with up-to-date recommendations or ads may be costly. It is important to note these dynamic changes in web page design because it will impact the effectiveness of web page parsing tools that are optimized for a particular page

design. This may also impact web crawling procedures because some pages may have links to related/recommended pages while others do not.

Figure 6(b) shows that time has a large influence on content-based features — this is shown by the increasing shift between the CDF plots for the number of different words feature when comparing our day 2, 12, 22, and 30 samples without our initial day 1 sample. This observed difference over time for content-based features is statistically significant, where the pages that are the most heavily influenced tend to correspond news, social networking, homepages in general (e.g., dailymail.com, weather.com, msnmoney.com, and twitter.com) and the pages that are least influenced tend to correspond to business/e-commerce and reference sites (e.g., target.com, dictionary.com, wikipedia.org, and webmd.com).



(a) **(b)**

Fig. 7: Web page layout and design can be changed over time — (a) Day 1 sample and (b) Day 11 sample.

4 Related Work

Past work has, to some degree, studied the influence that different type of client platforms may have on mostly performance-related applications. This includes studies that discuss the usability and design trade-offs between mobile web pages and traditional web pages [18,27] and understanding the energy-efficiency and performance-related impact of using mobile browsers and devices for web browsing [24,15]. There has also been recent work that studies (i) the diversity of web page downloads with respect to a single browser [5] and (ii) the impact that different web browsers have on the accuracy of in-browser load time measurements [12]. Time is a factor that is generally accounted for when performing web page measurement studies to ensure that the results are repeatable [5]. [11,10] are examples of measurement studies that thoroughly investigates the influence that time has on the frequency in which web page content changes — these studies also provide insight on the impact that time-related changes have on web crawling. [20] studied the impact that vantage-point has on web page content with respect to price discrimination and found the vantage point has a large influence on the price of goods on many major e-commerce sites.

Our work is different from this prior work because we explicitly study the impact that client platforms have on web page content. In particular, our work (i) investigates the *general* influence that client platforms have on web page *content* without considering performance and (ii) we explicitly consider client

platforms that are typically not considered in prior studies including different operating systems, browser types and versions, and tablets.

5 Concluding Remarks

In this paper, we address the question — *to what extent does a client platform influence the content of a base HTML web page for the* **same** *URL request?* We download base HTML-source files in a manner that controls for the influence of over 30 different client platforms. We extract quantitative HTML-based features and perform a comprehensive analysis of the differences that are present across different client platforms. We find differences in web page downloads across client platforms in both expected and unexpected ways. In addition, these observed differences have practical significance in a number of important web-related applications including web archival, mobile web development, document summarization, information sharing, and user experience. While there are many other differences that we find that are due to client platform, such as fonts and colors, we do not discuss them in detail because they have minimal utility in current popular web-related applications. In future work, we intend to (i) study the impact that user personalization (without regard for client platform) has on web page downloads, and (ii) study the influence that client platforms have on the *traffic* generated by web page downloads.

6 Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144081 as well as by NSF under Grant CNS-1526268.

References

1. Alexa. http://www.alexa.com. Accessed: 2013-02-19.
2. Future of the web workshop: Introduction and overview. http://netpreserve.org/sites/default/files/resources/OverviewFutureWebWorkshop.pdf. Accessed: 2015-05-29.
3. Html5 reference: The syntax, vocabulary and apis of html5. http://dev.w3.org/html5/html-author/. Accessed: 2015-04-30.
4. Make sure your site's ready for mobile-friendly google search results. https://support.google.com/adsense/answer/6196932?hl=en. Accessed: 2015-05.
5. M. Butkiewicz, H. V. Madhyastha, and V. Sekar. Understanding website complexity: measurements, metrics, and implications. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011.
6. T. Callahan, M. Allman, and M. Rabinovich. On modern dns behavior and properties. *ACM SIGCOMM Computer Communication Review*, 43(3):7–15, 2013.
7. D. Canali, M. Cova, G. Vigna, and C. Kruegel. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th international conference on World wide web*, pages 197–206. ACM, 2011.
8. B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman. Planetlab: an overlay testbed for broad-coverage services. *ACM SIGCOMM Computer Communication Review*, 33(3):3–12, 2003.

9. N. de Boer, M. van Leeuwen, R. van Luijk, K. Schouten, F. Frasincar, and D. Vandic. Identifying explicit features for sentiment analysis in consumer reviews. In *Web Information Systems Engineering–WISE 2014*. Springer, 2014.
10. F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, volume 119, 1997.
11. D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th international conference on World Wide Web*, pages 669–678. ACM, 2003.
12. E. Gavaletz, D. Hamon, and J. Kaur. Comparing in-browser methods of measuring resource load times. In *W3C Workshop on Web Performance 8*, 2012.
13. S. He and E. Chan. Surfing notes: An integrated web annotation and archiving tool. In *IEEE/WIC/ACM Web Intelligence and Intelligent Agent Technology*, 2012.
14. M. A. Himmel. Customization of web pages based on requester type, Dec. 26 2000. US Patent 6,167,441.
15. J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 165–178. ACM, 2010.
16. H. Iwai, Y. Hijikata, K. Ikeda, and S. Nishida. Sentence-based plot classification for online review comments. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 245–253. IEEE, 2014.
17. G. Jacob, E. Kirda, C. Kruegel, and G. Vigna. Pubcrawl: Protecting users and businesses from crawlers. In *Presented as part of the 21st USENIX Security Symposium*, pages 507–522, Berkeley, CA, 2012. USENIX.
18. T. Johnson and P. Seeling. Desktop and mobile web page comparison: characteristics, trends, and implications. *Communications Magazine, IEEE*, 52(9), 2014.
19. N. Leontiadis, T. Moore, and N. Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *USENIX Security Symposium*, 2011.
20. J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted search for price discrimination in e-commerce: First results. 2013.
21. G. R. Notess. The wayback machine: The web's archive. *ONLINE-WESTON THEN WILTON-*, 26(2):59–61, 2002.
22. M. S. Pera, R. Qumsiyeh, and Y.-K. Ng. An unsupervised sentiment classifier on summarized or full reviews. In *Web Information Systems Engineering–WISE 2010*, pages 142–156. Springer, 2010.
23. X. Roche et al. Httrack: Website copier. *on-line][consulta em 23-12-2008]. Disponível em: http://www. httrack. com*, 2012.
24. Z. Wang, F. X. Lin, L. Zhong, and M. Chishtie. How far can client-only solutions go for mobile browser speed? In *Proceedings of the 21st international conference on World Wide Web*, pages 31–40. ACM, 2012.
25. K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.
26. J. Westfall, R. Augusto, and G. Allen. Handling different browser platforms. In *Beginning Android Web Apps Development*, pages 85–98. Springer, 2012.
27. D. Zhang. Web content adaptation for mobile handheld devices. *Communications of the ACM*, 50(2):75–79, 2007.