Mining Usage Data and Development Artifacts

Olga Baysal, Reid Holmes, and Michael W. Godfrey
Software Architecture Group (SWAG)
David R. Cheriton School of Computer Science, University of Waterloo
{*obaysal, rtholmes, migod*}*@uwaterloo.ca*

***Abstract*—Software repository mining techniques generally focus on analyzing, unifying, and querying different kinds of development artifacts, such as source code, version control meta-data, defect tracking data, and electronic communication. In this work, we demonstrate how adding real-world usage data enables addressing broader questions of how software systems are actually used in practice, and by inference how development characteristics ultimately affect deployment, adoption, and usage. In particular, we explore how usage data that has been extracted from web server logs can be unified with product release history to study questions that concern both users' detailed dynamic behaviour as well as broad adoption trends across different deployment environments. To validate our approach, we performed a study of two open source web browsers: Firefox and Chrome. We found that while Chrome is being adopted at a consistent rate across platforms, Linux users have an order of magnitude higher rate of Firefox adoption. Also, Firefox adoption has been concentrated mainly in North America, while Chrome users appear to be more evenly distributed across the globe. Finally, we detected no evidence in age-specific differences in navigation behaviour among Chrome and Firefox users; however, we hypothesize that younger users are more likely to have more up-to-date versions than more mature users.**

***Keywords*-Usage mining, dynamic behaviour, user adoption, release history**

I. Introduction

With the continued growth of web services, the volume of user data collected by organizations has grown enormously. Analyzing such data can help software projects determine user values, evaluate product success, design marketing strategies, etc. Such analyses involve searching for meaningful patterns from a large collection of web server access logs.

The mining software repositories (MSR) research is generally focused on analyzing, unifying, and querying different kinds of development artifacts, such as source code, version control meta-data, defect tracking data, and electronic communication. Augmenting MSR-like activities with with real-world usage data can give insights into deployment, adoption, and user behaviour of the systems. Our previous work suggested that by studying dynamic web usage data, we can infer knowledge on user adoption trends [1]. Therefore, we decided to further study user community and adoption practices. Understanding how users adopt and use systems such as web browsers is important for measuring the success of a product and performing target market analysis. Analysis of adoption trends can also help to identify "hot spots" and "black holes" in user adoption at a global scale. By studying users' dynamic behaviour, we are able to gain knowledge on users' deployment environment to compare and contrast the use of a software system on different hardware configurations. This knowledge can then be used to assess software acceptance, user experience, or sustainability of a software system.

Web usage data, an artifact that has yet received little attention, stores valuable information about users and their behaviour related to adoption and use of software systems. Thus, in this paper, we demonstrate how we can employ available usage data and combine it with more traditional MSR-like data and analysis to study broader questions of how software systems are actually used.

We took a statistical approach to extract dynamic behaviour of the users from web traffic logs consisting of over 143 million entires. We analyzed each entry, representing a single page view by a single user, to determine the browser's name, version, and its host operating system, to map each host IP to a geographical location, and to track user browsing behaviour.

This paper addresses three research questions:

Q1: *Are there differences in platform preferences between browser end-users*?

Q2: *Is there a difference in geographic distribution between user populations*?

Q3: *Is there a difference in navigation behaviour between two user groups*?

Our study reveals several notable differences in the Firefox and Chrome user populations. Chrome undergoes continual and regular updates and has short release cycles, while Firefox is more traditional in delivering major updates, yet providing support for older platforms. Our data suggests that Firefox users are primarily centred in North America, while Chrome users are better distributed across the globe. We detected no evidence in age-specific differences in navigation behaviour among Chrome and Firefox users. However, we hypothesize that a younger population of users are more likely to have more up-to-date versions of a web browser than more mature users.

Our work makes several contributions. First, by mining web usage data we define several characteristics of the user population for empirical evaluation. Second, we analyze the usage patterns and highlight the main differences in how the browsers provide operating system (OS) support to the end-users, appeal to the users across the globe, and emphasize age-specific differences among its users in the adoption of new releases. Third, we discuss how characteristics of user population and adoption can provide insights into the sustainability of a product. Our findings may also help improve user experience. First, development team members may consider our findings to target a wider user population. Second, our findings have implications for better support of software applications that appeal to a wider population across the globe, support older and a variety of platforms, and reduce age-specific usability issues. And finally, our work might facilitate further research on user adoption and acceptance of software products.

The rest of the paper is organized as follows. Section II summarizes prior work. Section III describes how to mine dynamic usage data from web logs and the setup of our study. Section IV presents results of the empirical study and Section V discusses our findings on adoption trends, behavioural characteristics of users, and also addresses threats to validity. And finally, in Section VI we summarize our main findings.

II. Related Work

The most relevant related work is research on mining usage data. Web usage mining applies data mining techniques to discover usage patterns on web data. Web usage mining research provides a number of taxonomies summarizing existing research efforts in the areas, as well as various commercial offerings [2], [3].

Mobasher et al. [4] discussed web usage mining including sources and types of data such as web server application logs. They indicated that there are four primary groups of data sources: usage, content, structure, and user data. They discussed the key elements of web usage data pre-processing that required high-level tasks in usage data pre-processing that includes the integration of click stream data with other sources such as content or semantic information, as well as user and product information from operational databases. In this work, we unified usage data with product release history to study user dynamic behaviour and adoption trends.

Empirical software engineering research has focused on mining software development data (source code, electronic communication, defect data, requirements documentation, etc.). Relatively little work has been done on mining usage data. El-Ramly and Stroulia [5] mined software usage data to support re-engineering and program comprehension. They studied system-user interaction data that contained temporal sequences of user-generated events. They developed a process for mining interaction patterns and applied it to legacy and web-based systems. The discovered patterns were used for user interface re-engineering and personalization. While also mining web logs, they examined only user navigation activities on a web site to provide recommendations on other potential places to visit, “a user who visited link A also visited link B”. In our work, we explored a number of questions related to users’ detailed dynamic behaviour and adoption trends across various deployment environments. Li et al. [6] investigated how usage characteristics relate to field quality and how usage characteristics differ between beta and post-releases. They analyzed anonymous failure and usage data from millions of pre-release and post-release Windows machines. We study user characteristics across entire product release history and their relation to adoption and actual usage.

In our previous work we examined development artifacts – release histories, bug reporting and fixing data, as well as usage data of the Firefox and Chrome web browsers. In this study, two distinct profiles emerged: Firefox, as the older and established system, and Chrome, as the new and fast evolving system. When analyzing the usage data, we focused on only the difference in adoption trends and whether the volume of defects affects popularity of a browser. Figure 1 depicts observed trends in user adoption of the two browsers. In this paper, we take a more detailed look at the usage data by studying characteristics of the user populations of the browsers.

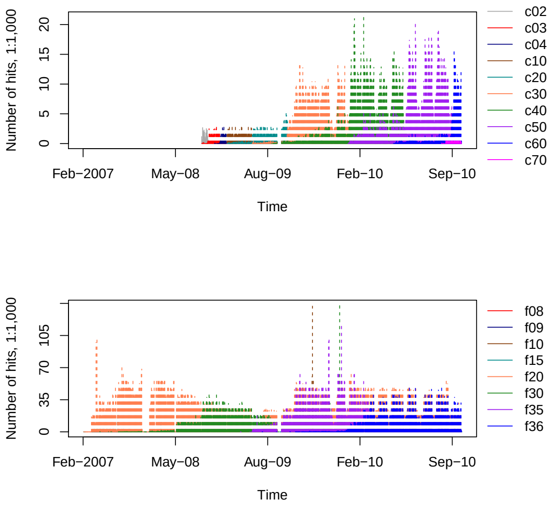

Figure 1. User Adoption Trends for Chrome (top) and Firefox (bottom).

Google Research performed a study on comparing update mechanisms of web browsers [7]. Their work investigates the effectiveness of web browser update mechanisms in securing end-users from various vulnerabilities. They performed a global scale measurement of update effectiveness comparing update strategies of four different web browsers – Google Chrome, Mozilla Firefox, Opera, Apple Safari, and MS

Internet Explorer. By tracking the usage shares over three weeks after a new release, they determined how fast users update to the latest version and compared the update performance between different releases of the same and other browsers. They applied similar approach of parsing user-agent string to determine the browser's name and version number. They evaluated the approach on the data obtained from Google web servers distributed all over the world. Unlike the Google study that investigates updates within the same major version of various web browsers, we studied major releases of the web browsers. We realize that our data is several orders of magnitude smaller than the Google data. However, we address different research questions related to the characteristics of user population and looked at broader analyses than just update speed.

III. Setup of the Study

This section describes browser release histories, usage log data, provides a sample of web server logs, and explains the process of mining dynamic behavioural data from web logs.

A. Release History

Mozilla Firefox is an older web browser, originally released in November 2004 as the successor of the Mozilla project. Google Chrome is a younger web browser that was first released in December 2008. Chrome is based on the Webkit layout engine, which is also used by Apple's Safari browser. Strictly speaking, Chrome is not open source but its core base — a project called Chromium — is.

As of November 2010, Firefox had released 8 major versions [8]; by this time there were 10 major releases for Chrome starting with version 0.2 [9]. In this paper, we define labels (see Table I) and use them when comparing releases of the browsers. On average, a new release of the Firefox browser is launched every 10 months, while a new version of the Chrome browser is released every 2.5 months (see Figure 2). The difference in release delivery of the browsers is statistically significant (p<0.005).

Table I
Browser release labels.

Label	Chrome Release	Firefox Release
b3	0.2	–
b2	0.3	0.8
b1	0.4	0.9
r1	1.0	1.0
r2	2.0	1.5
r3	3.0	2.0
r4	4.0	3.0
r5	5.0	3.5
r6	6.0	3.6
r7	7.0	–

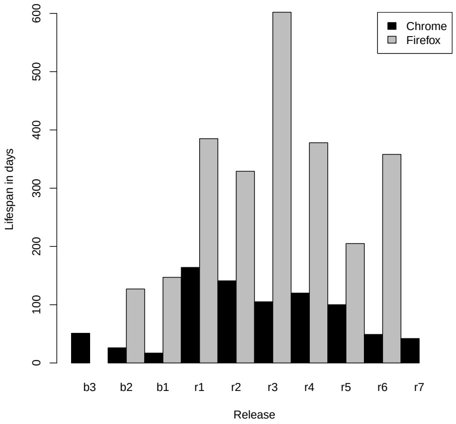

Figure 2. Lifespan of major releases for Chrome and Firefox. The difference in release delivery is statistically significant (p<0.005).

B. Web Server Logs

Web server logs are automatically generated by web servers whenever a user navigates through the web pages the server hosts. These logs contain detailed information about the browsing behaviour of visitors to a website. Each HTTP request to the server, called a *hit*, is recorded in the server access log. Each log record may contain the following fields: the client IP address, the time and date of the request, the requested resource, the status of the request, the HTTP method used, the size of the object returned to the client, the referring web resource, and the user-agent of the client. An example of a combined log format obtained from `www.cs.uwaterloo.ca` server is given in Figure 3. IP addresses of the visitors have been changed to protect their privacy. The user-agent field identifies the browser's type and version, as well as information about its host operating system.

```
10.0.0.1 - - [20/Oct/2008:23:05:24 -0400] "GET /undergrad/handbook/courses/
waitlist/index.shtml HTTP/1.1" 301 368 "http://www.cs.uwaterloo.ca/
current/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_5; en-us)
AppleWebKit/525.18 (KHTML, like Gecko) Version/3.1.2 Safari/525.20.1"

10.0.0.2 - - [26/Oct/2008:16:47:49 -0400] "GET /~fwtompa/.papers/xmldb-
desiderata.pdf HTTP/1.1" 301 365 "-" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; SV1)"
```

Figure 3. An example of server access log.

We obtained web server logs from the University of Waterloo, School of Computer Science (`http://www.cs.uwaterloo.ca`). The web server logs spanned from February 2007 until November 2010; there was 36GB of uncompressed raw textual data, comprising over 174 million entries.

C. Dynamic Behaviour Mining

Web-based services collect and store large amounts of user data. Analyzing such data can help organizations better understand users and their behaviour, develop marketing strategies, optimize the structure of their sites, etc. Such analysis involves searching for interesting patterns in large volumes of user data. A web server log is one of the primary sources for performing web usage mining since it explicitly records the browsing behaviour of the site visitors [2]. Web usage mining is a process of extracting useful information from web server logs by analyzing the behavioural patterns and profiles of users visiting a web site [2]–[4]. Mining users' browsing history provides insights into how users seek information and interact with the web site. Typical web usage mining consists of three stages: pre-processing, pattern discovery, and pattern analysis [2]. Figure 4 presents an architecture of a web usage mining process (adapted from [2]).

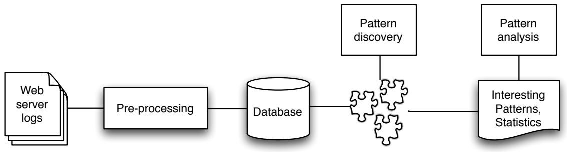

Figure 4. High level architecture of web usage mining.

In the **pre-processing** stage, web logs are cleaned and divided into transactions representing user activities during each visit. Depending on the analysis, the usage data is then transformed and aggregated at different levels of abstraction (users, sessions, click-streams, or page views). In this paper, we evaluate accesses to the web resources by considering page accesses or "hits" (Section V-A explains our decision). Our data cleaning process eliminates all the log entries generated by the web agents such as web crawlers, spiders, robots, indexers, or other intelligent agents that pre-fetch pages for caching purposes. This removed 31,255,963 entries (17%). We also restructured the date and time fields of the log entry to `[year-month-day hour:minute:second]` format. After the cleaning and transformation stages, the traffic data contained 143,613,905 entries, each corresponds to a single page view by a single user, and loaded then into a relational database.

During the **pattern discovery** stage, various operational methods can be applied to uncover patterns capturing user behaviour. The most commonly-used methods are descriptive statistical analysis, association rule mining, clustering, classification, sequential pattern analysis and dependency modelling [2], and prediction. These techniques are typically used for personalization, marketing intelligence, system improvements, and site modification. Statistical analysis provides statistical measures to organize and summarize information. Association rule mining concerns discovery of relationships between the items in transaction. Clustering is an unsupervised grouping of objects, while classification is supervised grouping. In web mining, the objects can be users, pages, sessions, events, etc. Sequential pattern analysis is similar to association rules but it also considers the sequence of events. The fact that page A is requested before page B is an example of a discovered pattern. All these techniques were designed for knowledge discovery from very large databases of numerical data and were adapted for web mining with relative success. In our work, we performed a descriptive statistical analysis when discovering patterns of user behaviour.

In this paper, we took a statistical approach for studying user behavioural dynamics. We examined the web log data to determine the browser type of our visitors. We analyzed the HTTP user agent strings that web browsers report when requesting a web page. We extracted the name of the browser from the user agent strings and calculated the number of accesses to our web site for each release of a browser. The proportion of accesses for each web browser is shown in Figure 5. The left pie chart shows the percentages of the total volume each that browser makes up. To our surprise, Firefox dominated the web traffic (31%), followed by Internet Explorer (24%), while Chrome users contribute only 3%. "Other" represents web traffic from other browsers including Safari, Netscape Navigator, Opera, mobile browsers, etc. To have a more fair picture on the web usage share, we eliminated the "Other" slice and normalized the cumulative access count per browser by its average market share: Chrome (10.70%), Firefox (20.16%), Microsoft's Internet Explorer or MSIE (52.37%) [10]. As can be observed from the right-hand side pie chart in Figure 5, Firefox is an obvious preferred choice among the visitors to our web site.

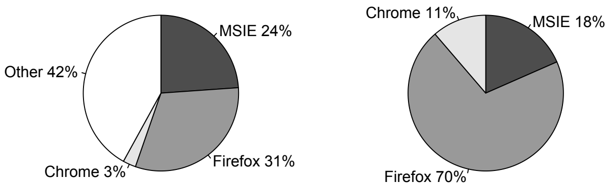

Figure 5. Pie charts representing volumes of accesses for a web browser. The pie chart on the left depicts percentages of the total hits per browser, while right-hand side shows "normalized" traffic shares among three leading browsers.

In the final stage, **pattern analysis**, the discovered patterns and statistics are further processed and used as input to applications such as recommender systems, report generation tools, visualization tools, etc. Since we performed only a statistical analysis during the pattern discovery stage, in this step we used GNU R tool [11] and presented graphs to report and explain our findings.

IV. Empirical Study

This section addresses each research question by describing how we approach the problem and reporting our findings. When applicable, we report results of statistical analysis of the data.

Q1: *Are there differences in platform preferences between end-users of the browsers?*

Since Chrome and Firefox are developed to run on multiple operating systems, we were interested to compare the browser's adoption and support for different deployment environments. We examined the choice of computing platform of the visitors to our website. For each release of the browser, we extracted the number of accesses from three operating systems: Windows, Linux, and OS X. This data was then normalized by the operating system market share obtained from StatOwl.com, which predominately measures United States web sites [12], [13]. Since our server is located in Canada, we consider our choice of market share statistics from StatOwl.com as reasonable to use for our analysis. The OS market share data represents "real" web site browsing community (excludes automated systems like search robots) and excludes mobile usage. For each release of a browser, we calculated an average OS market share percentage (reported every month) for the period of the release's lifespan. Since the market share numbers were reported starting from 2008, we applied the total average market share (Windows – 88.21%, OS X – 11.10%, Linux – 0.54%) to the releases deployed prior September 2008. Table II provides the percentages we used to normalize our usage data with respect to the users' choice of the platform.

Table II
Operating systems market share.

Rel.	Chrome			Firefox		
	Win	**OS X**	**Linux**	**Win**	**OS X**	**Linux**
b3	90.67%	8.87%	0.43%	–	–	–
b2	90.39%	9.13%	0.45%	88.21%	11.10%	0.54%
b1	90.34%	9.00%	0.63%	88.21%	11.10%	0.54%
r1	91.12%	8.21%	0.61%	88.21%	11.10%	0.54%
r2	89.74%	9.59%	0.57%	88.21%	11.10%	0.54%
r3	88.43%	11.06%	0.41%	88.21%	11.10%	0.54%
r4	88.35%	11.07%	0.44%	90.89%	8.49%	0.56%
r5	87.93%	11.40%	0.51%	89.02%	10.37%	0.51%
r6	87.65%	11.66%	0.53%	87.68%	11.65%	0.51%
r7	87.43%	11.84%	0.59%	–	–	–

Results The distribution of the number of user accesses from a platform is presented in Figure 6. A beanplot consists of a one-dimensional scatter plot (aka boxplot), its distribution as a density shape and an average line for the distribution [14]. The left side of a beanplot represents the density of the distribution for Chrome, while the right side of a beanplot shows the distribution for the Firefox browser. Applying Mann-Whitney statistical test, we compared the distributions of each platform across two browsers. The results show that the difference in density for both OS X and Linux platforms between two browsers is statistically significant (p<0.05), while distributions of Windows users across Chrome and Firefox are fairly similar (p=0.40). **Users do not adopt browsers equally across operating systems.** Users on a Linux or OS X choose Firefox over Chrome. On Windows, users equally opt for either one of the two browsers.

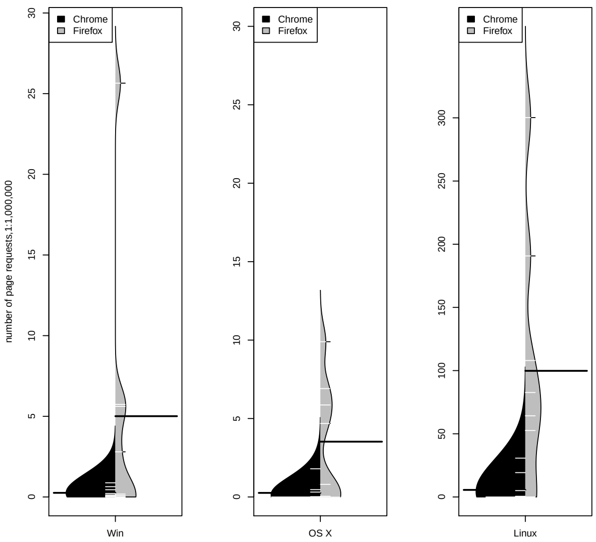

Figure 6. Asymmetric beanplots showing the density of the page requests by user's platform. The left side of each bean consists of hits for the Chrome browser, whereas the right side of a bean contains hits for Firefox. The horizontal lines represent the average. The difference in distributions for both OS X and Linux platforms between two browsers is statistically significant (p<0.05).

We then performed Kruskal-Wallis statistical test to compare distributions between the three different platforms for each browser (see Figure 7). Unlike Chrome (p=0.23), the difference in distributions of operating systems between each other for the Firefox browser is statistically significant (p=0.05). This suggests that **Linux users have an order of magnitude higher rate of Firefox adoption than OS X or Windows users. While Chrome is being adopted fairly consistently across platforms**.

By analyzing historical trends of Chrome and Firefox support for different operating systems (see Figure 8), we noticed that Firefox offers outstanding OS compatibility from the very beginning, while Chrome begins to reach for Linux and OS X users only starting from release `r4`, i.e., Chrome 4.0 (Google officially started to offer OS X and Linux OS support with the release of Chrome 5.0). Firefox

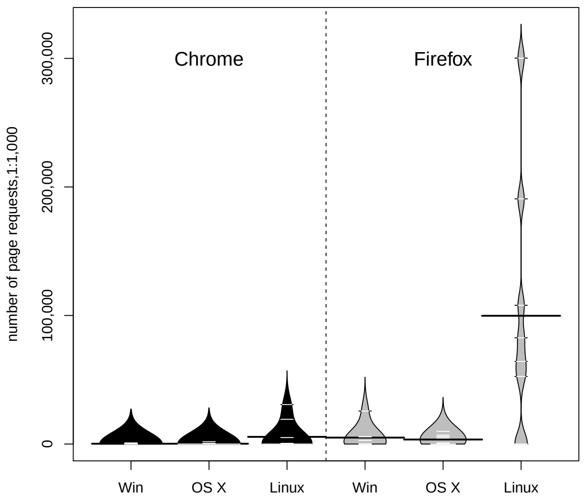

Figure 7. Beanplots showing the density of the page requests by user's platform within a browser. Black beans represent Chrome, and grey beans represent Firefox. The difference between OS platforms for Firefox is statistically significant (p=0.05).

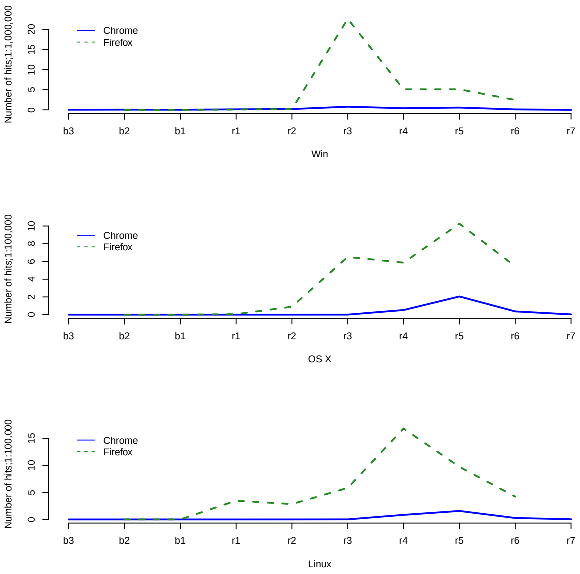

Figure 8. Graphs showing support for Windows, OS X and Linux platforms across releases of Chrome and Firefox.

reaches the peak of its adoption among Windows users in release `r3` (Firefox 2.0), among OS X population releases `r3` (Firefox 2.0) and `r5` (Firefox 3.5) are more well adopted than others, while Linux users seem to favour release `r4` (Firefox 3.0). We should also mention that Firefox can run not only on Windows, OS X and Linux, but also on BSD and other Unix platforms [15]. Therefore, we can say that Firefox provides early and better OS compatibility.

Q2: *Is there a difference in geographic distribution between user populations?*

The previous question has shown that there are clear differences between how two browsers are being adopted across operating systems. We now study geographical location of the users and whether there is a difference in adoption of the browsers across the globe. We used a geolocation service to track the geographic distribution of visitors to our website. We used Geo::IPfree, a Perl module, to look up a country of an IP address. We used six continents from WorldAtlas.com [16] to map a user's IP address to a geographical location. The list of continents (we call them regions) includes Africa (AF), Asia (AS), Europe (EU), North America (NA), South America (SA) and Australia/Oceania (OC). During the process of mapping IP address to the country and region, we detected a number of private IP addresses (local network), which we excluded from the analysis.

Results Figure 9 illustrates the differences in the distribution of the user populations by world regions. The Geo::IP database contains information from various registry sources. In some cases, a country is only indicated as Europe, which means that the requests from such hosts may come from anywhere in the European Union. To bridge the global digital divide – the disparities in the opportunities to access the Internet between developed and developing countries [17], we normalized our user accesses by the world's Internet usage data. The statistics on the distribution of the Internet users by world regions report the following numbers: NA 13.0%, AS 44.0%, EU 22.7%, SA 10.3%, AF 5.7% and OC 1.0% [18]. In our sample, we found that 85% of Firefox users and only 72% of Chrome users are located in North America. While overall, Chrome adoption is better distributed across the globe.

We then compared user adoption of the browsers with respect to the country coverage. From our sample, we found that Chrome is adopted by the users in 187 countries, while the Firefox user population covers 207 countries. Table III presents statistics on the user accesses by the top 5 countries. We were not surprised to see China and India in the top 5 countries list as these two countries contribute to the majority of the international students in our school's undergraduate program.

The results suggest that **Firefox adoption has been more heavily concentrated in North America, while Chrome users are better distributed across the globe**. Thus, it is safe to say that Chrome has a more culturally diverse user population.

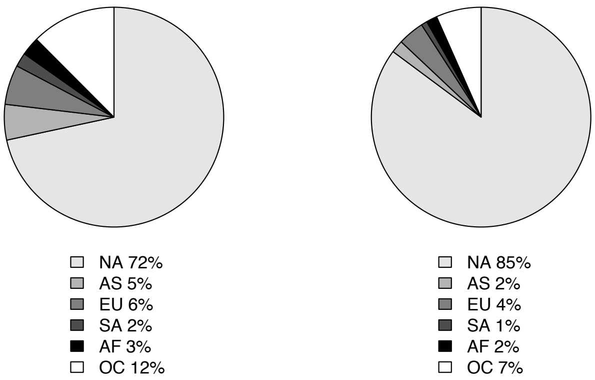

Figure 9. Pie charts showing the density of the page requests by region for Chrome (left) and Firefox (right).

Table III
TOP 5 COUNTRIES OF USER'S ACCESSES

Chrome	Firefox
Canada (1,968,421)	Canada (32,236,878)
USA (603,149)	USA (3,424,990)
India (193,805)	India (670,807)
China (87,058)	Europe (557,854)
UK (73,986)	UK (386,408)

Q3: *Is there a difference in navigation behaviour between two user groups?*

By looking at the content of the pages requested by the visitors, we wanted to identify whether the user populations of two browsers have different browsing goals and behaviour. We were interested in classifying users according to their navigation behaviour on the CS website. To investigate patterns in the browsing behaviour of the users, we first determined the types of the web content our web site offers to the visitors. Our school's web site is mainly designated to the following visitors:

1) students – offering information to current and prospective students about the courses, their description, schedules, lectures, assignments, exams, etc.
2) researchers/industry partners – offering information on faculty's and grad students' research interests, current projects, publications, potential collaboration opportunities, etc.

Based on the content of the page requests, we defined two profiles of the user accesses related to *undergrad* teaching or *research*. We note that not every access is related to either of the two profiles. Table IV defines our rules for classifying visitors' requests into two profiles.

Requests to the publications are defined as ones to any .pdf document located under `/pubs/`, `/publications/` or `/papers/` directories.

Table IV
PATTERN MATCHING RULES TO CLASSIFY USER ACCESS TYPE

Undergrad	Research
• requests to any CS 100–600-level undergraduate course	• requests to any CS 700–800-level graduate course
• requests to course descriptions and course schedules for undergrads	• requests to publications
• requests to information for future undergrads and prospective students	• requests to anything under `/research`

Results Figure 10 illustrates the differences in the browsing behaviour between Chrome and Firefox users. As we expected, undergrad pages are accessed more often than research-related ones by both Chrome and Firefox users. While **we detected no statistical evidence for age-specific differences in browsing behaviour among Firefox and Chrome users**, Figure 10 suggests that the browsing habits of Chrome users follow approximately normal distribution, while for Firefox the distributions of the accesses of both research and undergrad pages are more spread out. Since we did not detect any statistical difference between the distributions, we performed the Kolmogorov-Smirnov test to test for the equality of two distributions. The results suggested that for the undergrad profile, Chrome and Firefox samples come from the same distribution. This tells us that **Chrome and Firefox users behave the same way when navigating to undergrad content**.

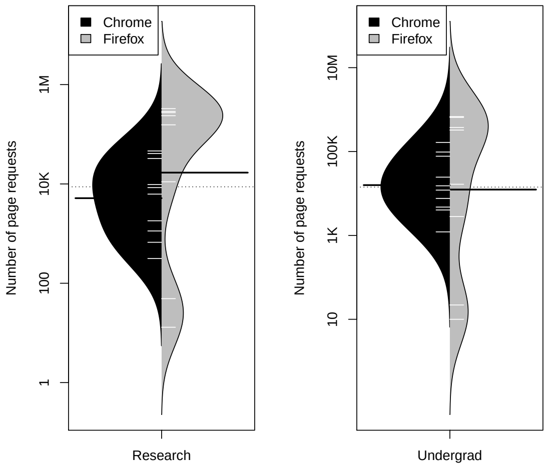

Figure 10. Beanplots showing the differences in navigation behaviour between Chrome and Firefox users.

Historical trends of the user accesses for each release of a browser are demonstrated in Figure 11. Chrome users have similar navigation patterns when viewing both types of the web content: both research- and undergrad-related pages were accessed from more recent releases of a browser

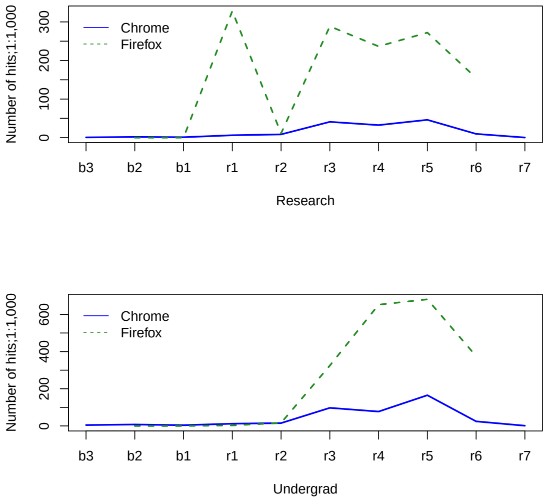

Figure 11. Plots showing the distributions of user accesses to research and undergrad web content per each release of a browser.

starting from Chrome 3.0. Unlike Chrome, we found quite different patterns in viewing web content among Firefox users. Most hits to the research-related pages came from Firefox 1.0 (shown as release `r1` in Figure 11). We were surprised to see no accesses from the earlier releases of Firefox to the undergrad content. Firefox 2.0 is the oldest browser used to navigate to the undergrad information, while the largest volume of the page views to this content originated from the Firefox releases 3.0 and 3.5. These findings suggest that undergraduate students, a younger population of users, have more up-to-date versions of a web browser (true for both Chrome and Firefox), while researchers, a more mature population of users, do not update their browsers as quick as their younger counterparts.

Surprisingly, the first five releases of the Chrome browser have no or comparatively fewer number of hits to both types of the web content on our web site. Since our web traffic data is dated to February 2007 and Google Chrome was release later in 2008, we expected to see our visitors having early releases of Chrome installed on their computers. This observation suggests that Chrome adoption started slowly, with the first wave a year later with Chrome 3.0.

Firefox 1.0 was released in November 2004, yet at the time of the first records in our logs, this release was more than two years old. Thus, Firefox was well adopted from the very first release (mainly due to earlier deployment of the browser under different names - *m/b* (mozilla/browser) under the Mozilla Suite, Phoenix, and Mozilla Firebird) and users stayed quite loyal to the initial release of the browser, hesitating to update it to Firefox 1.5 up until Firefox 2.0 became available in October 2006.

V. Discussion

This section discusses our findings and lessons learned about the differences in user populations and adoption of two open source web browsers. It also addresses several threats to validity.

Software projects collect and store enormous archives of web usage data that are often disregarded or unused. Such archives contain data on user characteristics including user environment, locality, browsing behaviour. This paper shows that applying mining techniques we can extract such user characteristics from web logs to study user dynamic behaviour and adoption and combine them with traditional MSR data. By analyzing user behavioural and adoption patterns, we can, for example, assess several properties of sustainability of a software project. A sustainable software system can be defined as being "socially and environmentally bearable, viable economically without introducing impacts to the environment, and socially and economically equitable and accessible to everyone" [19]. Therefore, our empirical findings could provide the following insights on sustainability:

- Development process and practices, in particular release history of a product, can account for the maintenance quality. For example, shorter release cycles underlay better maintenance and delivery of more reliable and defect-free software.
- Firefox's support for older operating systems and platform compatibility fosters hardware sustainability and reduces e-waste.
- Since Chrome users appear to be better distributed through the different regions in the world, Chrome supports larger diversity and ethnicity among its user population. Tracking cultural trends of the user adoption can provide information on the globalization of a project.
- User navigation behaviour can offer insights on user population and age diversity, as well as the success of a software release.

From these findings, we propose to assess sustainability of a system by measuring not only environmental aspects (e.g., development of software solutions that require lower energy consumption), but also social ones (user behaviour, adoption and interactions).

For open source products, it can be quite challenging to construct dynamic usage trends based on the number of product downloads due to the lack of a central repository to track such downloads. However, analysis of web usage data can provide valuable information on how users adopt and use software projects, analysis of historical trends can justify the popularity of certain releases of a software product. It is important to know the end-users of a product not only from the statistics collected by the marketing surveys but also by analyzing real usage data such as web logs to infer knowledge on user population, their technical environment, locality and navigation behaviour. The knowledge of these usage characteristics can lead to better understanding of how software systems are actually being used in practice.

A. Threats To Validity

External validity. Our findings are limited by the obtained data set: web server logs. Since we perform a comparative study, our school's web traffic is representative of the worldwide user population of two browsers. Usage data sets are typically not publicly available due to privacy and business concerns. In our analysis we tried to balance the data representativeness and to avoid being biased by normalizing the number of the page requests by the system's usage share. Further studies may be necessary to confirm our findings.

Internal validity. Our web usage data has a few gaps due to the specifics of the university's backup routine, making the quality of the logs an important threat. A small number of CS undergraduate courses are offered through UW-ACE — a web-based course management system. Our web server logs do not include user accesses to such courses. We also need to mention that CS graduate courses normally reside on the faculty's web space. However, some faculty members have their web sites hosted by the web servers belonging to the Faculty of Mathematics. In such cases, we were not able to track accesses to these courses. For example, CS846 course has been taught by several professors through the years and its web site is located on both `plg.uwaterloo.ca` and `se.uwaterloo.ca` sub-domains. Neither `plg.` nor `se.` sub-domains are hosted under `cs.uwaterloo.ca`.

Our choice of the granularity in analyzing web logs is determined by the existing challenges to identify users. Accurate tracking of the individual users by IP address is not always possible. A user who accesses the web from different machines (e.g., work vs. home computer) might have different IP addresses. A user that uses multiple browsers on the same machine will appear as multiple users (user agents will differ). ISPs can assign multiple IP addresses to a user for each request or several users might share same IP address.

Unlike Google who uses cookies to track individual users and their navigation behaviour over the web, we are limited with the data captured in typical web traffic logs.

Since March 2011, Mozilla has accelerated the Firefox release cycle, and has provided several new releases with shorter timespans between them. Since our web logs are spanned from February 2007 until November 2010, new Firefox update policy is not reflected in our study. Adoption of Firefox beyond release 3.6 is not considered in this paper. Newer web logs would be needed to reflect the influence of Chrome-like rapid release deployment practices of Firefox on its user adoption rates.

Construct validity. We have chosen a set of metrics to quantify the value of the collected data that captures only a part of its potential meaning. Our choices are a function of our interest in exploring the data and the availability and structure of the data sets.

Conclusion validity. We reported findings based on the statistical significance. We applied statistical analysis when needed, and were able to reject null hypotheses and detect interesting patterns.

VI. Conclusion

This paper demonstrated how analysis of real-world usage data can reveal valuable information on how software systems are actually used in practice. In particular, we showed how usage data can be extracted from web server logs and combined with development information to provide insight into user dynamic behaviour as well as adoption trends across various deployment environment. We took a statistical approach to mine dynamic usage data to determine characteristics of the user population. We analyzed discovered usage patterns and outlined the main differences in user adoption, deployment and usage of the Chrome and Firefox browsers. We also discussed how usage characteristics can help to account for sustainability of a software system.

We found that while Chrome is being adopted at a consistent rate across platforms, Linux users have an order of magnitude higher rate of Firefox adoption. Also, Firefox adoption has been concentrated mainly in North America, while Chrome users appear to be more evenly distributed across the globe. Finally, we detected no evidence in age-specific differences in navigation behaviour among Chrome and Firefox users; however, we hypothesize that younger users are more likely to have more up-to-date versions than more mature users.

Mining usage data is a powerful way to track and assess user dynamic behaviour and adoption trends of a software system.

Acknowledgement

We wish to thank the Computer Science Computing Facility (CSCF) of the University of Waterloo for providing web traffic log files for this study, as well as Dr. Ian Davis for his help on extracting the raw usage data from the web traffic archives.

References

[1] O. Baysal, I. Davis, and M. W. Godfrey, "A Tale of Two Browsers," in *Proceeding of the 8th working conference on Mining software repositories*, ser. MSR '11, 2011, pp. 238–241.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from web data," *SIGKDD Explor. Newsl.*, vol. 1, pp. 12–23, January 2000.

[3] R. Cooley, B. Mobasher, and J. Srivastava, “Web mining: information and pattern discovery on the world wide web,” in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, nov 1997, pp. 558 –567.

[4] B. Mobasher, N. Jain, E.-H. S. Han, and J. Srivastava, “Web mining: Pattern discovery from world wide web transactions,” Tech. Rep., 1996.

[5] M. El-Ramly and E. Stroulia, “Mining software usage data,” in *Proceedings 1st International Workshop on Mining Software Repositories*, ser. MSR’04, 2004, pp. 64–68.

[6] P. L. Li, R. Kivett, Z. Zhan, S.-e. Jeon, N. Nagappan, B. Murphy, and A. J. Ko, “Characterizing the differences between pre- and post- release versions of software,” in *Proceeding of the 33rd international conference on Software engineering*, ser. ICSE ’11. New York, NY, USA: ACM, 2011, pp. 716–725.

[7] T. Duebendorfer and S. Frei, “Why Silent Updates Boost Security,” TIK, ETH Zurich, Tech. Rep. 302, May 2009.

[8] Wikipedia, “Mozilla Firefox — Wikipedia, the free encyclopedia,” http://en.wikipedia.org/wiki/Mozilla_Firefox, [Online; accessed 28-November-2010].

[9] ——, “Google Chrome — Wikipedia, the free encyclopedia,” http://en.wikipedia.org/wiki/Google_Chrome, [Online; accessed 28-November-2010].

[10] StatOwl.com, “Web browser market share,” August 2011. [Online]. Available: http://www.statowl.com/web_browser_market_share.php

[11] R. Development Core Team, “The R Project for Statistical Computing,” http://www.r-project.org/, [Online; accessed 10-July-2011].

[12] StatOwl.com, “Operating systems market share,” July 2011. [Online]. Available: http://statowl.com/operating_system_market_share.php

[13] ——, “About our data,” July 2011. [Online]. Available: http://statowl.com/about_our_data.php

[14] P. Kampstra, “Beanplot: A boxplot alternative for visual comparison of distributions,” *Journal of Statistical Software, Code Snippets*, vol. 28, no. 1, pp. 1–9, 2008. [Online]. Available: http://www.jstatsoft.org/v28/c01/

[15] Wikipedia, “The Comparison of Web Browsers — Wikipedia, the free encyclopedia,” http://en.wikipedia.org/wiki/Comparison_of_web_browsers, [Online; accessed 14-Aug-2011].

[16] WorldAtlas.com, “Countries listed by continent,” July 2011. [Online]. Available: http://www.worldatlas.com/cntycont.htm

[17] M.-T. Lu, “Digital divide in developing countries,” *Journal of Global Information Technology Management*, vol. 4, no. 3, pp. 1–4, 2001.

[18] InternetWorldStats.com, “Internet usage and population statistics,” [Online; accessed 20-September-2011]. [Online]. Available: \url{http://www.internetworldstats.com/stats.htm}

[19] W. Adams, “The future of sustainability: Re-thinking environment and development in the twenty-first century,” IUCN Renowned Thinkers Meeting, Tech. Rep., January (2006. [Online]. Available: http://cmsdata.iucn.org/downloads/iucn_future_of_sustanability.pdf