User Identification and Characterization From Web Browsing Behavior

Kevin Kwok
Science and Engineering Apprenticeship Program
Annandale High School
Mentor: Myriam Abramson
Code 5584
Naval Research Laboratory
4555 Overlook Ave. SW
Washington, DC 20375

Abstract

Ever since True Names by Vernor Vinge, identity has been recognized as our most valued possession in cyberspace. Attribution is a key concept in enabling trusted identities and deterring malicious activity. This paper attempts to identify users in a non-adversarial setting based on behavior related to browsing by extracting navigational features which can be derived from a user's clickstream.

Introduction

The problem with establishing a user's identity is one of the fundamental and still largely unresolved problems with the web. In recent history there have been multiple approaches to solving that very problem through schemes like OpenID[3], OAuth[8], BrowserID[6] and smartphone-based two-factor authentication. People find it useful that rather than registering a new account for every new service, they can authenticate using an identity which was established for another provider they have already invested time and trust into. OAuth and OpenID allow users to authenticate themselves to third party services by using their Facebook, Twitter or Google+ IDs without divulging their passwords or any other sensitive information to other parties. In addition to the service oriented authentication system is Mozilla's BrowserID initiative, which is based on the concept that the browser which the user implicitly entrusts with all of his or her activity should serve the role of providing identity to other services.

A trend which has occurred largely in parallel to the trend toward centralized authentication is the large amount of data and usage behavior collected by those services. Facebook, Google and Twitter each have in their databases logs of the activity of every single user. And while this trend has been around for a long time, to the great chagrin of privacy advocates, it is only recently that people have begun taking advantage of this data for personal use. For example, there are people who call themselves "lifeloggers" who wear cameras to stream everything happening in the world around them[10]. More interesting is the kind of analysis that can be done when those large quantities of data are analyzed in aggregate to expose the nuances of human behavior that escape conscious recognition [9]. A recent feature of the Android 4.0 Jelly Bean release is a feature called Google Now which allows Google to use the immense amount of data Google has already collected, through GPS tracking via mobile phones, search and email history as well as machine learning to recognize patterns in the daily lives of users and to provide contextually relevant information[4].

However, the data which has been collected about users by these services and by web browsers can also be used for providing identity services. Users are voluntarily giving services and applications access to their histories in order to gain some kind of benefit. One of these possible realms is for a better notion of identity, beyond the granularity which traditional passwords or even two-factor authentication can bring[7]. As new types of web sites and new features in web browsers arise, such as tabbed browsing and single-page applications, this increases the ability for different users to develop different habits which stay with them wherever they browse. These habits which may manifest in various features such as the number of page views per tab or the time between subsequent clicks, and while each of them may not be particularly identifying, the sum of all these attributes may yield a unique identity among all the users of the web. This paper attempts to categorize users based on certain navigation related browsing behaviors in a non-adversarial setting.

Data Collection

The data for this project was collected in three ways. The first two were browser extension which logged data which comprised a user's clickstream for Firefox and Chrome.

Clickstream Recording

Both the Firefox extension and the Chrome extension collected data in a CSV format, with the fields `timestamp, url, browser user agent` recorded on every line. The Firefox extension was built by Myriam.

The chrome extension used the WebNavigation API in order to retrieve navigation activity and the HTML5 FileSystem API to save the log onto a disk location. Since Chrome's application and extension sandbox disallows direct writing to the user's filesystem, the solution was to write to that managed virtual file system which internally held a folder with non-human-readable file names[2]. When the user would press a "Download" button situtated in the Browser Action popup, it would retrieve a Blob reference to the file and save it to the Downloads folder by converting it into an Object URL and simulating a click event on a link with the `download` attribute set to a filename derived from the subject ID.

Late in the process it was discovered that the chrome extension didn't behave in quite the same exact way as the original Firefox extension. The Firefox extension recorded the data whenever a tab was visited but not while it was unvisited or in the background, in essence, it recorded whatever page was visible to the user at any given moment of time. In contrast, the Chrome extension recorded the initial page navigation and did not record additional rows for when the page was again in view. When there are no tabs which are explored out of order, the behavior would be the same. However the issue was rectified soon after its discovery by observing tab events in addition to the navigation data.

However, throughout the whole process there was not much data collected and this poses a significant problem in the way of actually determining the viability of identifiying users amongst a large pool based on browsing habits.

Clicktree Recording

The third source of data was another Chrome extension which logged additional data, such as tab behavior. Rather than storing data in a CSV format, it stored every log as a single JSON-formatted line. That chrome extension logged additional information such as when tabs are created and how many tabs and windows are currently open in the browser.

The logs which are more frequent and contain more information can be used to reconstruct a "clicktree"[11]. The range of browsing behavior in a tabbed environment can be expressed degrees of different types of graph search methods. In theory it may be possible to, with certain degrees of ambiguity, construct a clicktree from a clickstream data source based on various assumptions. For instance, every page which is visited can be analyzed and scraped (assuming the content is not exclusively dynamic and does not demand authentication) to generate a list of links which may spawn new tabs. HTML attributes such as `target` may be used to determine if the link will default to the creation of a new tab. Each subsequent link can be looked up in a list which is sorted by the last time accessed to determine the most likely parent of the following URL. Repeating this process would eventually yield a clicktree, albeit it is impossible to consider it a perfect reproduction as certain circumstances may generate ambiguity.

The single tab or traditional browsing dynamic can be thought of as an implmentation of a depth first search algorithm. The user keeps adding web pages to a history stack and hits back when the user finds that he or she has deviated too far from the target. At that point the user backtracks and proceeds back up the chain to branch off at some other point which appears promising.

The tab based interaction mode may be interpreted as some loose form of a breadth first search, which rather than descending linearly and backtracking, might open several tabs from a single page and explore them. This is not a model which can describe all types of tab behavior, but serves as a basic interpretation of the “parallel browsing” paradigm[5].

While it may be useful to think of tabbed browsing behavior as some linear combination of these two modes, the reality is much more nuanced as many other behaviors play into characterizing tabbed navigation behavior. For example, one might be able to browse with tabs in a manner which equates an interpretation of the depth first search by opening several tabs for each page and exploring tabs based on the time opened.

This model also fails to incorporate the conditions of which browser tabs are closed, as this can vary wildly between individuals. However, the rate at which tabs are closed depend possibly as much on the actual design of web browsers. Some browsers scale better with large numbers of tabs, while others, like Internet Explorer 9 hold additional interface chrome elements on the same row, reducing the maximum number of tabs. That also depends on the size of the window and the size of the monitor.

Results

The first thing which was analyzed was patterns in the distribution between navigational events. Each record on the clickstream includes (in the first column), the UNIX timestamp representation of the time (according to the local machine) within millisecond accuracy.

Since the data is in CSV format, it is fairly easy to parse in most situations. However, the parser in the R statistical software package finds difficulty parsing the raw clickstream because commas in URLs and User Agent strings aren’t properly escaped. So the data is first piped into a short python script to extract the timestamps.

One interesting thing which could be done with this sort of data is analysis over patterns which recur over durations of days, weeks, months or years. However, as noted before the amount of data is quite limited and as such it would be unlikely to yield any useful results. However, there are some very basic daily patterns which are discussed in Figure 1.

Since there isn’t much data, the patterns which can be observed must be microscopic rather than macroscopic, showing up in the range of milliseconds rather than hours. One of the more interesting hypotheses which came up in terms of this idea was Burstiness.

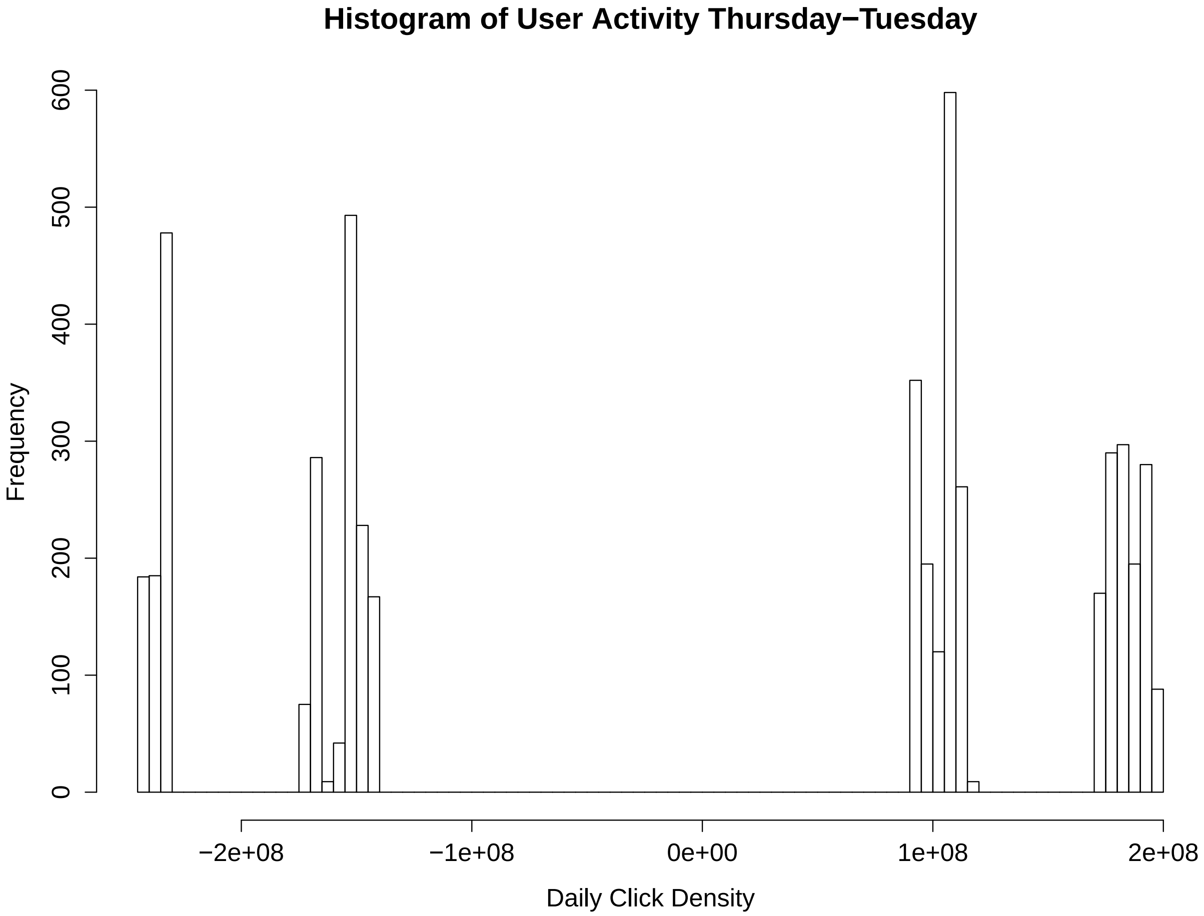

Figure 1: This histogram corresponds to a section of the data which was collected from my own browsing clickstream. The leftmost bars of the graph represent Thursday afternoon, followed by Friday and a gap for the weekends. There is a general daily pattern which appears largely gaussian. Also visible is a depression around the middle of each day corresponding to an afternoon lunch break, however soon afterwards the graph tends to surge to the daily high before declining towards the end of the typical work day.

Time to Click

Since patterns over long durations are hard to find with such a limited quantity of data, one way to change the scale of actions is by looking at the time elapsed between subsequent actions. In mathematical terms this can be thought of as the first-order discrete difference of the set of timestamps recorded. This is calculated by subtracting each element by the term immediately preceeding it.

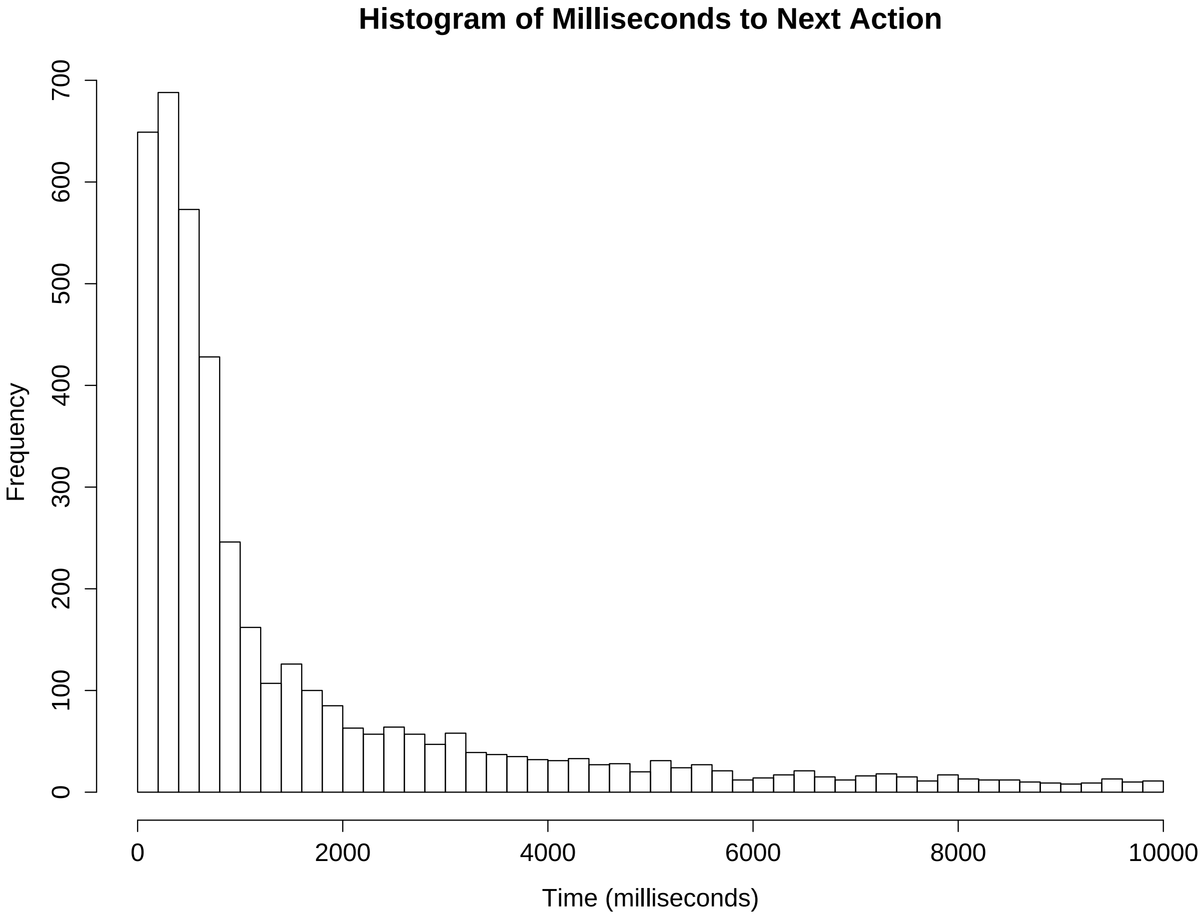

Figure 2: This histogram corresponds to the number elapsed between one action and the one immediately after. This graph depicts gaps between 0 and 10 seconds, however the tail goes on for much longer, as is shown in the logarithmic histogram on Figure 3, which shows the plot from zero to half an hour. The vast majority of the actions take place within a few hundred milliseconds of another. This establishes that at least in the case of this user, actions tend to take place fairly rapidly involving very little actual careful reading or inspection of the page. However, this does not necessarily mean that the main mode of browsing is short clicks because this does not represent the fraction of time used by each mode.

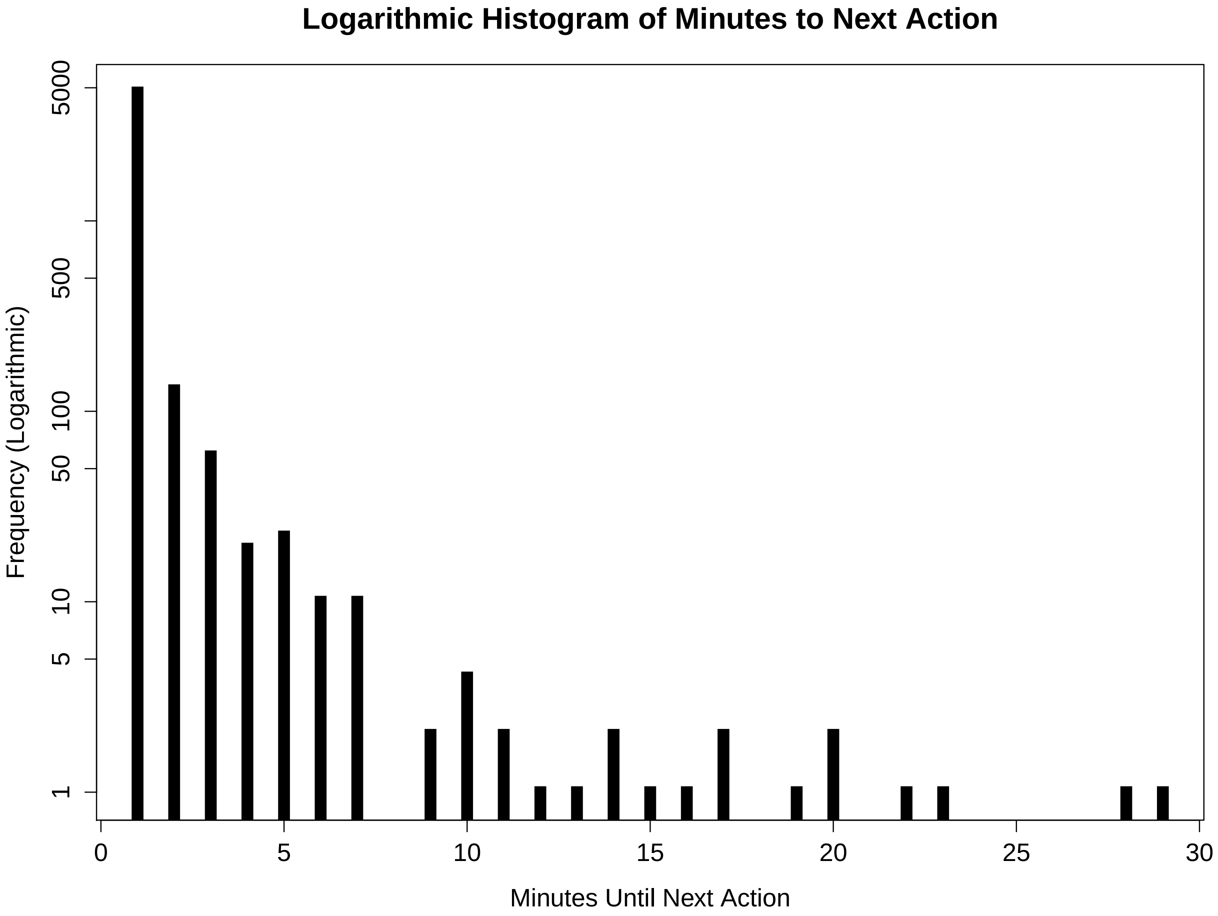

Figure 3: This is a logarithmic histogram extending from 0 minutes to half an hour, showing that there is a very long tail of mostly uniformly distributed elements.

Burstiness

Human activity tends to occur in limited bursts of time[1], and as such browsing navigation actions are expected to exhibit the same properties. One way to demonstrate that is by taking the second order discrete difference between the times recorded, which will give a measure of how actions tend to occur with some notion of intertia, so that subsequent actions are to occur at approximately the same rate. An expected result would be a large cluster around 0, meaning that many of the clicks tend to take approximately the same amount of time as the previous one.

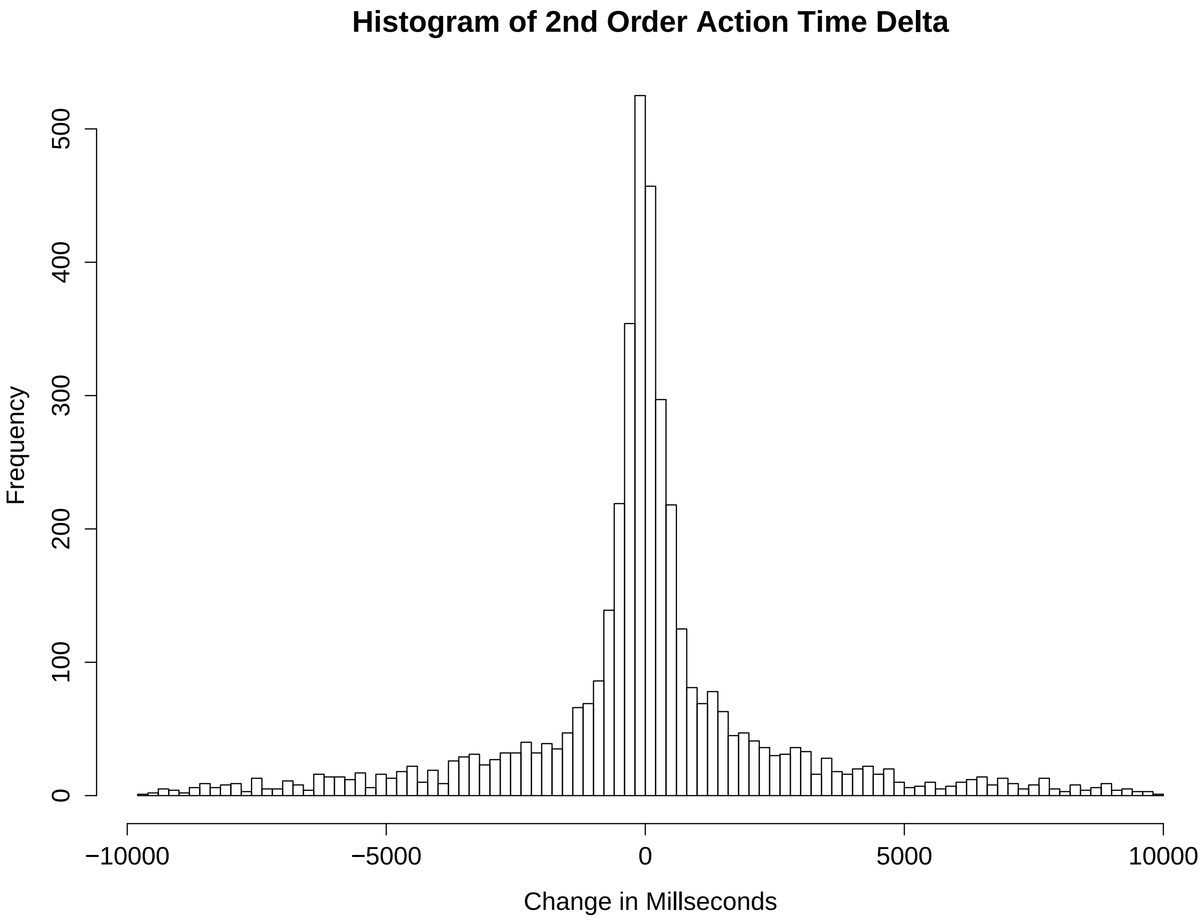

Figure 4: This is a histogram of the change in times between the times between actions. The plot is surprisingly gaussian and centered, however the mean of both this subject's data and another subject's both have a mean of approximately -50 milliseconds, there isn't enough data to prove that this is a trend but it certainly seems interesting.

However, this may not be necessarily indicative of the existence of burstiness because running a difference operation on any normally distributed set of data will yield something which is apparently gaussian. However, the height of the peak at zero seems to suggest that this is not simply a formulation of an effect which is innate of any normally distributed variable and that instead, it is because there actually is some kind of bursty behavior which the user is exhibiting.

Conclusion

There wasn't much progress in the way of the original goal in finding features which may be useful in discriminating between different users online in large part because of the very little data which as been gathered. From the two basically complete sets of data, many attributes seem somewhat distinct, but this might be within normal variation of a single user. However, there were still a few somewhat interesting results which came from the plots of timestamps.

The first thing that strikes as remarkable is the degree to which the second order plot is gaussian. The tails however seem much longer than one would expect for a typical gaussian plot. And in both sets of data, the average value ended up in the same ballpark, -50 milliseconds within 5 milliseconds of each other. This strikes as some pattern which may hold some deeper meaning in terms of human behavior.

Acknowledgements

I would like to thank my mentor, Myriam Abramson, for guiding me through the process of developing this project. Throughout this project I have learned a lot about the field of web analytics and the value of being able to establish identity online. I would also like to thank my alternate mentor, Stephen Russell for guidance and help along the way. I also want to thank the branch head Ranjeev Mittu. I also want to thank the other people working on the project, Charlotte Young and Alex Wu who have been friendly and helpful along the way. I also want to thank David Aha for ideas on how to progress despite the limited amount of data. I also want to thank fellow SEAP students, Sreenath Are, Jay Sebastian and Will Lewitus for making me feel welcome. Last but not least, I want to thank the SEAP coordinators for providing me this opportunity to do work at such an incredible facility with skilled and helpful professionals. My time spent here has served as an incredibly valuable and rewarding experience which is sure to prepare me for the future.

References

[1] A.L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[2] Eric Bidelman. File System API changes for Chrome 13, June 2011.

[3] OpenID Foundation. Openid Authentication 2.0, December 2007.

[4] Google. Introducing Google Now, June 2012.

[5] Jeff Huang and Ryen W. White. Parallel browsing behavior on the web. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 13–18, New York, NY, USA, 2010. ACM.

[6] David Mills. Introducing BrowserID, July 2011.

[7] Andrew Hintz Nishit Shah et al. Advanced sign-in security, February 2011.

[8] D. Recordon. The OAuth 2.0 Authorization Framework, July 2012.

[9] Deb K Roy and Alex P Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113 – 146, 2002.

[10] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R.W. Picard, and A. Pentland. Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments*, 6(4):386–398, 1997.

[11] Maximilian Viermetz, Carsten Stolz, Vassil Gedov, and Michal Skubacz. Relevance and impact of tabbed browsing behavior on web usage mining. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 262–269, Washington, DC, USA, 2006. IEEE Computer Society.