A Visualization Framework for Self-monitoring of Web-based Information Disclosure

Kulsoom Abdullah
Communications Systems Center
Georgia Institute of Technology
Atlanta, Georgia
Email: kulsoom@gatech.edu

Gregory Conti
Electrical Engineering and Computer Science
United States Military Academy
West Point, NY
Email: gregory-conti@usma.edu

Raheem Beyah
Computer Science
Georgia State University
Atlanta, GA
Email: rbeyah@cs.gsu.edu

***Abstract*—Free online tools such as search, email and mapping come with a hidden cost. Web users obtain such services by making micropayments of personal and organizational information to the web service providers. Web companies use this information to create customized advertising and tailored user experiences. Individually, each transaction appears innocuous, but when aggregated, the result is often highly sensitive. The impact of AOL's inadvertent disclosure of 20 million nominally anonymized search queries underscores the pressing need for increasing web privacy and raising user awareness of the problem. Rather than advocate extreme legal and policy measures to address the dilemma, this paper proposes an equitable self-monitoring solution. Self-monitoring allows individual users and large enterprises to regulate their web-based interactions intelligently and still allow online companies to innovate and flourish. The primary contributions of our work include exploration of visualization techniques that support self-monitoring based on a user requirements survey, a human-centric evaluation, and a Firefox extension based on one of the visual monitoring solutions developed.**

***Index Terms*—googling, information disclosure, privacy, query visualization, search**

I. Introduction

Users of the World Wide Web have made billions of web requests since its inception. Use of electronic commerce grew, as did the use of targeted advertising coupled with free web services. Afterwards, advertising became personalized and more effective by using the information that users provided. The number of free web services has grown but most web users are unaware of the information they disclose. Previously, we found that 81% of college undergraduates had conducted searches for information they would not want disclosed to their current or future employer [1].

The World Wide Web has brought about profound change in how we seek, acquire and communicate information. With a global user base approaching one billion [2], web users are aggressive consumers of free online services provided by companies such as AOL, Google, Microsoft and Yahoo, among others [3]. While exact usage statistics are rarely disclosed by these companies, industry analysts estimate that Googles web search queries alone exceed 100 million per day. Other free tools, such as web-based communication are similarly popular. Yahoos email service and MSNs Messenger have an estimated 260 million and 240 million users, respectively [4].

These companies use data collection and mining in order to provide effective targeted advertising and customized user experiences. Due to the sensitivity of corporate data retention and data mining programs, online companies are reluctant to publicly discuss specifics. In one of the rare instances where the subject has been addressed, Yahoos Chief Data Officer, Usama Fayad, stated that Yahoo collects 10 terabytes of user data per day, not including content, email or images. He further stated that Yahoos first and largest data mining challenge is the ability to capture all of this data reliably, process it, reduce it, and use it to feed the many, many reports and applications. [5] While the exact extent of data retention by online companies is not publicly known, anecdotal evidence suggests that every user interaction is scrupulously logged and stored indefinitely. However, some limited progress has been made. In March 2007, Google announced it would remove IP address information from older query logs after 18-24 months [6]. While promising, we do not believe this negates the need for informed self-monitoring; accidents can and do occur. According to the Privacy Clearinghouse, at least 104,137,499 records containing sensitive personal information have been involved in security breaches [7]. Furthermore, the information these stockpiles contain offers unprecedented power that will be coveted by many who will seek to acquire access by legal and illegal means.

In August 2006, AOL brought media attention to web information disclosure by releasing a search query dataset containing over 20 million searches by 657,426 AOL users. Many queries included sensitive information such as medical conditions, addresses, business dealings and social security numbers. Despite the backdrop of ubiquitous data retention and the AOL disclosure, we do not believe the world would be a better place without the tools provided by these companies. We believe that users must make an informed decision when choosing to use these tools. To this end we have developed a number of visualization techniques that will allow individual users and larger organizations to self-monitor their activity. This approach will empower end users, and the companies they work for, to make better-informed decisions regarding their online activities and, we anticipate, create an environment where online companies can continue to innovate and flourish.

The contributions of this paper are the following: the design and evaluation of four self-monitoring visualization techniques based on the results of a user requirements survey, a visualization centric analysis of the AOL dataset and an initial Firefox extension based on one of the visualization techniques designed. The visualizations we present take into account scaling and usage statistics from the AOL dataset as well as user requirements for the design. Our evaluation shows that the visualizations increase user awareness with minimal negative impact on their relationships with online companies.

We are exploring self-monitoring in situations where there is an implicit assumption of a private interaction between the user and an online company. We focus on search queries, as it is a ubiquitous application with millions of users generating approximately 5.7 billion search queries per month [8], but believe our work extends well to other similar applications, such as email, mapping and news.

This paper expands on topics discussed in our previous work [9] and discusses a Firefox extension we developed based on one of the designed visual mockups and the resulting feedback in that work. Section 2 of this paper places our research in the body of related work. Section 3 is an analysis of candidate information sources. Section 4 analyzes the AOL dataset for key visualization parameters. Section 5 presents our visualization design. Sections 6 and 7 contain our evaluation results and analysis. Section 8 discusses the Firefox extension developed to keep track of searches and to visualize them as file tree view. Section 9 outlines our conclusions and directions for future work.

II. Related Work

Self-monitoring of web-based information disclosure is largely an unexplored area. SearchClock is the only search query specific visualization tool we have found [10]. It is an interesting initial prototype that focuses on the entire 657,000 user AOL dataset, not on individual or business scale requirements.

The most readily available tool capable of self-monitoring is the history function included in modern browsers but the history function shows previously visited websites not the information users disclose.

Relevant Firefox extensions include Page Addict and Packet Garden. Page Addict shows the user how much time he or she has spent on different websites; reports are available in text list and simple chart formats [11]. Packet Garden plots Internet activity on a globe [12]. It uses a garden metaphor to grow plants based on online activity. While aesthetically pleasing, Packet Garden is not designed for efficient self-monitoring.

In addition to browser extensions, researchers have developed several related techniques for analyzing online activity based on network monitoring. Etherpeg [13] and Driftnet [14] monitor wireless hot spots and display collages of images they captured off the network. Most recently Maynor and Graham developed the Ferret tool that also captures activity from wireless hotspots, but uses a far more comprehensive approach [15]. Ferret understands 25 protocols and collects a wide variety of online activity including network addresses, email, passwords and search queries. Currently Ferret provides limited text-based reports and a simple tree visualization.

Websense [16], a commercial tool, is designed to monitor web activity but focuses on preventing access to websites with undesired content.

There are text-based visualization techniques related to our work but do not directly address query visualizations. They do provide useful insight into text visualization. PaperLens helped show the interplay between research topics, researchers and research sources [17]. Lins Visualization for the Document Space provides useful insight into how to visualize and create category groupings [18]. Themeriver is useful to consider because of its approach to visualizing themes over time [19].

Our approach is unique because we focus on efficient and effective visual self-monitoring of web activity.

III. Data Source Analysis

To design, and later drive, our self-monitoring system, we considered two primary sources: data collected by applications at the host and data collected by network monitoring devices. In addition we also considered the efficacy of using the AOL dataset.

A. Host-based Data Collection

Host-based data collection depends on instrumenting individual workstations to capture data disclosed via web-based interactions. We envision two likely approaches in this category: the browser extension and the browser form field cache. The advantage of the host-based approach is that it empowers individual users to self-monitor their activity without encountering the privacy concerns one might face when collecting and aggregating multiple web-interaction data flows at the multiple user network level.

The first approach, browser extensions, is a powerful one. Extensions are closely integrated into the browser and have full access to the interaction data, including URLs visited, form data entered and time stamps. We believe it is a natural next step to create plug-ins that allow users to monitor their activities. The Firefox browser, in particular, has a vibrant plug-in development community and bears great promise for future work [20]. Additionally, if a given browser based technique proved popular, the code could be moved from an optional plug-in and integrated directly into the browsers code base.

The browser form field cache is a second potential source of data for visualization. This cache is used by browsers to auto complete form field entries such as search terms and addresses with previously typed information. As part of our work, we investigated the Firefox form field caching mechanism and found significant weaknesses. Firefox stores all form field entries in a single file, using the Mork file format [21]. We developed a tool to extract form field entries from this file, however, the file only includes raw field entries and not time stamps or associated websites [22]. As a result, all web activity is lumped into one large cluster of all form field data from all destination websites. This browser cache contains a great

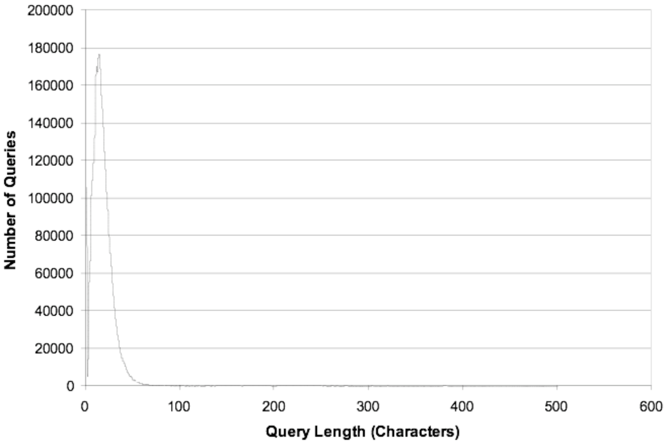


Fig. 1. Query Length (Character)

deal of sensitive information that users disclose through web interactions [23], but due to its lack of precision we believe it is only of limited value for our self-monitoring task. We leave investigation of other browser caching mechanisms, such as those found in Microsofts Internet Explorer, for future work.

B. Network-based Data Collection

Network based approaches to gathering web interaction data do so by capturing network traffic. By inspecting network traffic destined for desired websites, individuals, but more likely organizations, can easily gather interaction data for a large number of machines. For example, an organization which desires to self-monitor its employees disclosures, could capture network packets at a centralized point, such as a corporate firewall, using an application capable of network sniffing. The contents of the network packets could be inspected for HTML form data as well as HTTP GET and POST data to identify desired content. By combining this data with packet IP addressing information as well as Domain Name System (DNS) data, the organization could acquire disclosures from its own internal machines sent to desired external web service companies.

An important variant of the network sniffing approach is to use a proxy server through which all organizational web traffic flows. A properly configured proxy server, such as Squid [24], could then generate web interaction datasets on behalf of the organization. One issue for both the network sniffing and proxy server approaches to dataset generation is encryption. If the users browser and the destination web server uses SSL then the contents of the interaction will not be accessible, unless the organization places appropriate certificates on its individual workstations or has access to encryption keys. In todays environment, most interactions with web services, particularly search, are not encrypted. For example, Google redirects HTTPS requests (https://www.google.com/) to its primary, unencrypted, page at http://www.google.com/.

TABLE I
SUMMARY OF REPRESENTATIVE USER CHARACTERISTICS

ID	Total Queries	Unique Queries	Average Query Length	Average Queries Per Day
574625	124	69	22.86	28
671641	4598	1162	21.95	86
3059644	36	23	21.36	11

C. AOL Dataset

In many respects the AOL Dataset is a cornucopia of web search data. It consists of 21,011,340 search queries of which 10,154,742 are unique. Each record consists of a randomized anonymous user id, a time stamp, the query as well as the rank and URL of the search result clicked (if any). Released on 4 August 2006, the dataset contains unfiltered search queries for 657,426 people over a 93 day period ending in 31 May 2006. While the complete dataset is 2.2GB, the set was distributed in ten equal size files of 222MB each. It is important to note that while the dataset contains a very large number of users, anecdotal evidence suggests they may represent less experienced users. Similarly, AOL search, although ultimately provided by Google, is listed as the 4th most popular search service by Nielsen-Netratings. While the AOL dataset is a unique source of data, because it was collected by a search company and inadvertently released to the general public, we use it due to its high quality and relevance. We believe that it meets our intent of determining the correct visualization parameters, such as query length and frequency, so as to intelligently design search visualizations.

The AOL dataset contains extremely valuable data, but also contains highly sensitive, sometimes personally identifying, information. Each researcher must consider the ethics of using the dataset and to what extent. The information is now publicly available on many sites across the Internet [25]. We hand picked users that could not be readily identified based on their queries. We will discuss more specific details in the next section.

IV. DATA ANALYSIS

A. Scaling Analysis

In order to create realistically scaled visualizations, we analyzed the AOL dataset to determine several key characteristics of the search queries and user interaction timing. In particular we determined the average number of queries per day per user (0.34), the average number of unique queries per day per user (0.17) and the average number of characters in a query (17.5). Figure 1 illustrates the distribution of query lengths of 3.5 million queries from the AOL dataset. Note that the majority of queries were relatively short, but there is a tail of longer queries.

B. Selecting Representative Users

Ideally we would like to design visualizations that support the widest possible range of user types, but for this work we hand selected sample users who exhibited significant variations in search volume and frequency. In addition, as we searched for these users, we manually examined each of their queries and did not select users with personally identifying or offensive content. We believe our omission of individuals with these forms of sensitive content did not impact our design goals, which depended upon volume of queries and frequency of search, not on sensitivity of the search. Based on our selection criteria we chose the following users: User 574625 (sporadic use), User 671641 (heavy and frequent use) and User 3059644 (very light use).

Based on the entire 93-day period, for each user we calculated the total number of queries, total number of unique queries, average query length (in characters) and average number of queries per day. Table I summarizes the results.

We also determined the distribution of searches over time as we believed that this distribution is a key factor when constructing visualizations. While not included in the original dataset, we were also interested in a general categorization of the queries to provide additional semantic information to test potential visualizations. To provide this functionality, we chose to categorize a portion of the dataset by extracting the list of unique queries for each of our representative users and manually adding a textual category field and numeric count field for each entry on the list. We began with the list of categories, provided by Pass, Chowdhury and Torgeson [26], modifying it slightly. We understand that manual categorization is unreasonable for the entire dataset, but believe that some degree of automated categorization will be possible by applying techniques from the data mining community. We leave this exploration for future work.

V. Visualization Design

Our design strategy was to first create mockups of the visualizations and then to perform an evaluation of those mockups. Section 9 discusses the next step we took in creating a functional prototype.

Priority was placed on the mockups so that we could obtain evaluations at the start of the design, rather than design and build the tool(s) without any initial user feedback. Mockups provide more flexibility, and allow us to work out potential problems we would not have realized. Static mockup images were created based on the AOL dataset, which is real-life, but not real-time data.

For our mockups, we used queries and timestamps from the dataset as well as the manually added categorical information we discussed in section IV.B. Some information parameters included in the dataset that we did not use show promise for future analysis. These included the destination the user went to after submitting the query and the rank of the search result they clicked on. In addition, query reformulations should also be considered. A reformulation is when a user submits a query and tries to reformulate the query to correct or refine what they are searching for. This could be due to a misspelling of the query or to narrow down a large search result. In addition, the user may or may not have clicked on a result after any of these query submissions. The AOL paper [26], mentioned that 28% of queries were reformulations and that an average query is formulated 2.6 times.

A maximum resolution of 1024 by 768 was used in the mock-ups as this is a common monitor resolution. The amount of queries that can be seen in one screen is limited by this size, and is most significant with the high query user (#671641). Also the histogram view (Section V.B.1) shows a limited number of days due to the horizontal resolution.

A. User Task Analysis

The primary task we are addressing is that of individual user self-monitoring of web-based information disclosures. To help assess user specific requirements for our visualization mock-ups we conducted a focus group session with 18 undergraduate college students. We deliberately solicited participants from non-technical majors because we believe they are more representative of our projected user base. To help put our work in context, we began the sensing session with a short discussion of the AOL dataset disclosure and our desire to provide the means for users to monitor their web-based information disclosure. After this initial discussion we asked session participants to suggest tasks that our system might facilitate. Suggested tasks include:

1) providing a time-sequence listing of disclosures, preferably including date and time
2) categorizing and grouping information disclosed by content and destination site
3) monitoring most frequently used search terms
4) listing most frequently visited sites
5) helping monitor cookies, including the number sent per site and the expiration date
6) listing the time spent at different websites
7) listing their activities at each site
8) mentioning whether login was required for each site
9) providing a way to highlight disclosure of sensitive information
10) determining if they had shopped on a given site

Based on this session and our own assessment we chose to address the first three (italicized) tasks. In order to further scope the problem we focused on web search activities, but suggest future work across all forms of disclosure, such as online mapping, instant messaging and finance.

To facilitate informed self-monitoring we plan to explore showing users their web search activity over time. In particular we wish to provide users with the ability to rapidly scan their queries over varying times scales in a way that allows them to self-assess the sensitivity of their aggregated disclosures. While, aggregating many user flows into a single enterprise-level visualization is very relevant future work, here we focus on only the single user problem.

The first task is addressed by our histogram and Seesoft [27] based views. The second task is addressed by visualizing query use via categories which is shown with a bubble chart and a file explorer-like hierarchical view. Most frequently used queries are visualized by a bubble chart for the third task.

B. Initial Prototypes

For mockups we used representative AOL search terms, quantities and search timings to provide the data, perl for scripting as well as Omnigraffle for creating the images themselves. The only exception is the bubble chart used with IBMs ManyEyes service [28]. Four visualization types are used to visualize search queries and search query categories: histogram, bubble chart, Seesoft visualization, and file tree display.

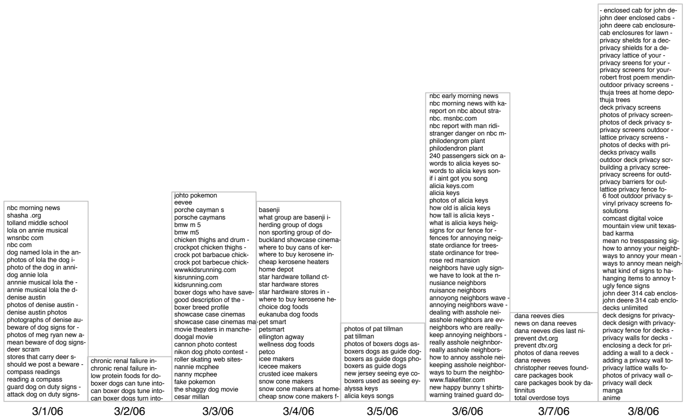


Fig. 2. Search term histogram over time with 24 hour increments (AOL User #671641). This prototype displays search queries on the vertical axis and time on the horizontal axis. The labels on the horizontal axis are month/day/year/.

1) Search Historgram Over Time : Our first mockup was a histogram of search queries with time across the horizontal axis. It used queries extracted from the AOL dataset organized into 24-hour increments (Figure 2).

All mockups (except the query frequency variant of the bubble chart) show queries only the first time they are used. In the evaluation section, we will discuss whether this hides important data from the user.

For time scale, an increment of 24 hours is used for the Seesoft and histogram views. Category based views (bubble chart, hierarchical) used the entire 93 days. The time based views, Seesoft and histogram, were more sensitive to these scaling decisions than the category based views.

To cope with long queries we truncated queries at 25 characters and used a dash to indicate the truncation. Because the average query length from the AOL dataset was 17.5 characters we believe this was a reasonable design trade off. This length proved quite suitable for the width of the histogram and Seesoft bars. Also, 25 characters covers a large majority of the queries without many being truncated (see Figure 11). Standard font sizes are used and only active days are shown to conserve horizontal space. For a size of 1024 by 768 and 25 characters maximum for queries, up to 8 days can be seen in one screen. 480 maximum queries can be viewed, assuming the queries are spread 60 queries/day for 8 days.

2) Bubble Chart: We chose a bubble chart because it is good for presenting a non-time based representation of the data, particularly when values differ by several orders of magnitude or with datasets with tens to hundreds of values.

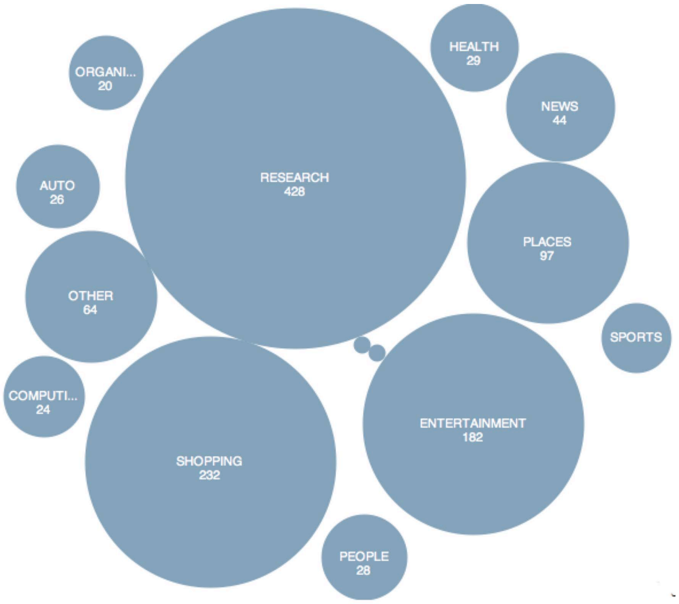


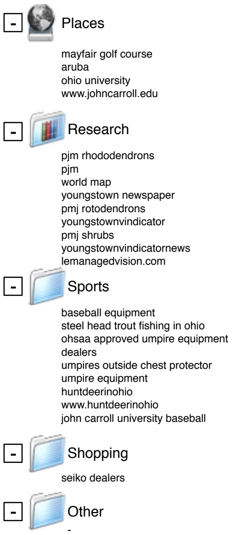
Fig. 3. Using a bubble chart to display categories (user #671641), individual query terms, and query phrases.

We created two bubble charts, using the ManyEyes service provided by IBM, for each user: one for query categories, Figure 3, and one for top queries. The first mapped the total count of queries belonging to each category to determine the size of the bubble. The second mapped the number of occurrences of each query to the size of the bubble.

As we created the bubble charts we made several important design decisions. We set the maximum number of bubbles to be 50 because more bubbles would have been difficult to read or interpret. Our initial assessment of the bubble chart is that it quickly becomes congested and is of marginal use. We will discuss user feedback in the next section. While color is an important characteristic for future work, we did not utilize it at this time.

3) File Tree Display: This visualization, Figure 4, is similar to the browser history function but is optimized for information disclosure monitoring. We created this categorical view from our sample users data. This view would also be suitable for search terms, phrases, destination websites and time. We manually grouped terms into categories using those mentioned in Passs AOL paper. Unfortunately, the paper did not mention how they categorized queries so we defined category meanings ourselves, modifying the list slightly. In this view, the queries were not truncated as query length was not affected by horizontal space as the other time-based visualizations. Here we are not limited horizontally, but vertically where a maximum of 15 queries can be seen at once. A user would scroll through queries, opening and closing folders as part of the interaction.

4) Seesoft Visualization: A Seesoft based mockup is shown in Figure refseesoft, which depicts 69 queries with 28 active



- Places
  - mayfair golf course
  - aruba
  - ohio university
  - www.johncarroll.edu
- Research
  - pjm rhododendrons
  - pjm
  - world map
  - youngstown newspaper
  - pmj rotodendrons
  - youngstownvindicator
  - pmj shrubs
  - youngstownvindicatornews
  - lemanagedvision.com
- Sports
  - baseball equipment
  - steel head trout fishing in ohio
  - ohsaa approved umpire equipment
  - dealers
  - umpires outside chest protector
  - umpire equipment
  - huntdeerinohio
  - www.huntdeerinohio
  - john carroll university baseball
- Shopping
  - seiko dealers
- Other
  - -

Fig. 4. Hierarchical display of user search terms (AOL user #3059644).

days from 3/1/06-5/31/06. This design is based on [28] which visualized program code. It can show a greater amount of queries and days compared to the histogram view. An empty box is shown to represent a non-active day. A maximum of 5 columns and about 368 queries fit into one screen. However, the number of viewable days depends on the distribution of the queries.

VI. Evaluation

The visualizations were evaluated to determine its strengths and weaknesses by 52 undergraduate students.

A. User Study

These were the questions asked for the user study: Questions on the usefulness of the mock-ups:

- What visualization is best for allowing self-monitoring of your online search activities?
- Was the visualization easy to understand?
- How effectively could you self-monitor your activity?

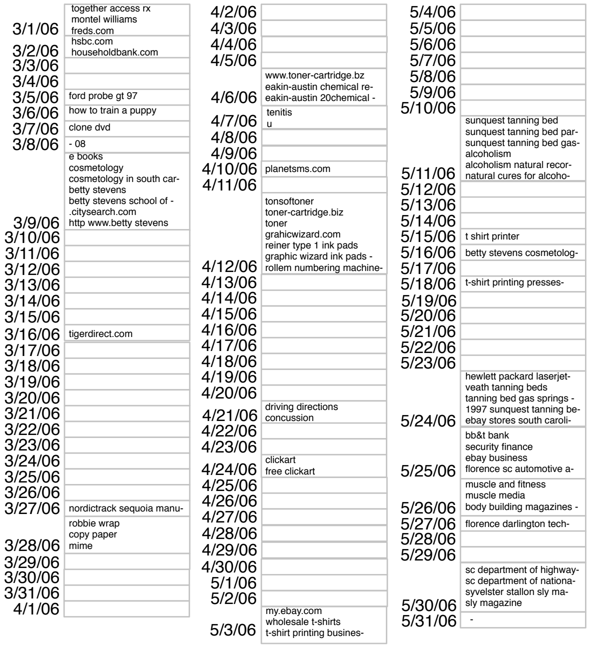
Questions related to how much search queries reveal:

- What percentage of the queries would you consider sensitive?
- What can you tell about the person based on this query visualization?

Other evaluation questions:

- What is the maximum number of queries this technique can handle before it becomes too crowded or otherwise unusable?
- How does it fare with various realistic time scales?
- How reasonable were our text size and time scale decisions?
- Did truncation matter to users (what length is best)?
- How would the user like to interact with the visualization?

The user study result summary is given in the next section.



3/1/06	together access rx montel williams freds.com
3/2/06	hsbc.com householdbank.com
3/3/06	
3/4/06	
3/5/06	ford probe gt 97
3/6/06	how to train a puppy
3/7/06	clone dvd
3/8/06	- 08
3/9/06	e books cosmetology cosmetology in south car- betty stevens betty stevens school of - .citysearch.com http www.betty stevens
3/10/06	
3/11/06	
3/12/06	
3/13/06	
3/14/06	
3/15/06	
3/16/06	tigerdirect.com
3/17/06	
3/18/06	
3/19/06	
3/20/06	
3/21/06	
3/22/06	
3/23/06	
3/24/06	
3/25/06	
3/26/06	
3/27/06	nordictrack sequoia manu-
3/28/06	robbie wrap copy paper mime
3/29/06	
3/30/06	
3/31/06	
4/1/06	
4/2/06	
4/3/06	
4/4/06	
4/5/06	
4/6/06	www.toner-cartridge.bz eakin-austin chemical re- eakin-austin 20chemical -
4/7/06	tenitis u
4/8/06	
4/9/06	
4/10/06	planetsms.com
4/11/06	
4/12/06	tonsoftoner toner-cartridge.biz toner grahicwizard.com reiner type 1 ink pads graphic wizard ink pads - rollem numbering machine-
4/13/06	
4/14/06	
4/15/06	
4/16/06	
4/17/06	
4/18/06	
4/19/06	
4/20/06	
4/21/06	driving directions concussion
4/22/06	
4/23/06	
4/24/06	clickart free clickart
4/25/06	
4/26/06	
4/27/06	
4/28/06	
4/29/06	
4/30/06	
5/1/06	
5/2/06	
5/3/06	my.ebay.com wholesale t-shirts t-shirt printing busines-
5/4/06	
5/5/06	
5/6/06	
5/7/06	
5/8/06	
5/9/06	
5/10/06	
5/11/06	sunquest tanning bed sunquest tanning bed par- sunquest tanning bed gas- alcoholism alcoholism natural recor- natural cures for alcoho-
5/12/06	
5/13/06	
5/14/06	
5/15/06	t shirt printer
5/16/06	betty stevens cosmetolog-
5/17/06	
5/18/06	t-shirt printing presses-
5/19/06	
5/20/06	
5/21/06	
5/22/06	
5/23/06	
5/24/06	hewlett packard laserjet- veath tanning beds tanning bed gas springs - 1997 sunquest tanning be- ebay stores south caroli-
5/25/06	bb&t bank security finance ebay business florence sc automotive a-
5/26/06	muscle and fitness muscle media body building magazines -
5/27/06	florence darlington tech-
5/28/06	
5/29/06	
5/30/06	sc department of highway- sc department of nationa- syvelster stallon sly ma- sly magazine
5/31/06	-

Fig. 5. A modified Seesoft visualization of search queries (user #574625).

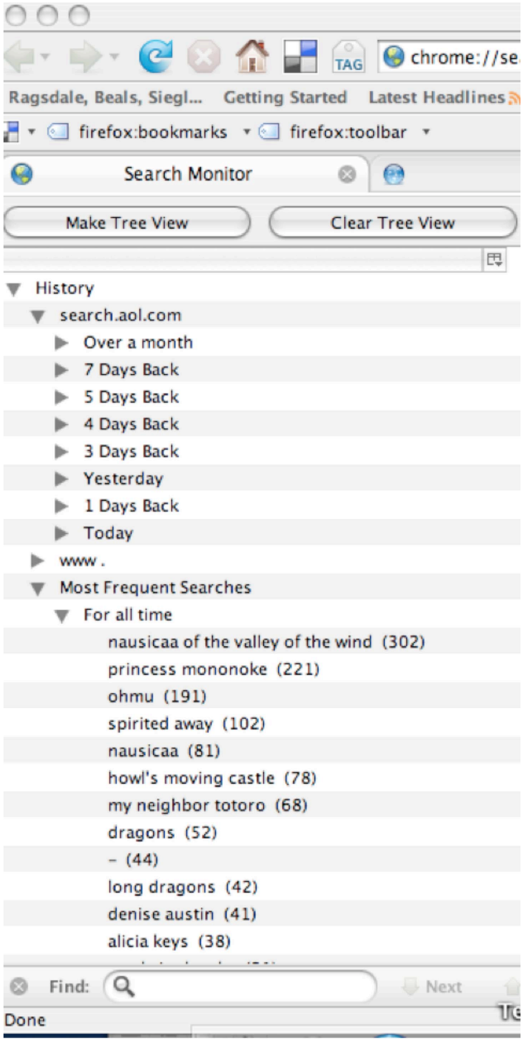
VII. Analysis

The surveys included sets of questions shown with mock-up images. Therefore, there was no testing of interaction functions, except for the bubble charts generated at the ManyEyes website. For the file tree visualization, which is like Windows Explorer, users were familiar with the concept.

The majority thought the visualizations were easy to understand (70%) and the text and time scale decisions were reasonable (75%). Many said the Seesoft view made better use of space than the histogram view while maintaining context better by including a space for non-active days (73%) . About half (47%) wanted the ability to get information such as the time the query was made and what the resulting action was after the query submission. The majority liked the file tree view (74%) vs. the bubble chart because it was user-friendly, well organized and broke down the information in a useful way. Those that preferred the bubble chart did because it was aesthetically pleasing and provided a clear way of seeing category activity but some of those that did not prefer the bubble chart thought it wasted space. Seeing overall categories with the option of viewing the detailed query list was seen as useful. The overall favorite was the tree view (55%), then bubble charts (31%), and the Seesoft view last (14%). Since the top visualization design only received 55% of the votes, this could be taken to indicate that potential solutions should include several display formats. Most thought that not showing duplicative queries loses some meaning (63%). One idea to counter this is to show the query and a number next to it to represent how many times that query was used. Truncating queries was not seen as a problem by anyone.

The majority thought it was useful to monitor their own query activity (86%) to see what information was being revealed, and additionally, for personal information management purposes, like the ability to go back to queries they already made and access those results again. Looking at these mock-ups, the users were able to give accurate assessments and opinions on the AOL user despite their initial unfamiliarity. Some additional suggestions were made including: allowing visualization of queries by different parameters, such as frequency, time and date in the same view as well as letting the user know how much time was spent at a particular site.

Our results show that we were successful in meeting the requirements of providing a time-based view of disclosed queries, categorizing queries, and monitoring more frequently used search terms as well as raising user awareness overall.



▼ History
▼ search.aol.com
▶ Over a month
▶ 7 Days Back
▶ 5 Days Back
▶ 4 Days Back
▶ 3 Days Back
▶ Yesterday
▶ 1 Days Back
▶ Today
▶ www .
▼ Most Frequent Searches
▼ For all time
nausicaa of the valley of the wind (302)
princess mononoke (221)
ohmu (191)
spirited away (102)
nausicaa (81)
howl's moving castle (78)
my neighbor totoro (68)
dragons (52)
– (44)
long dragons (42)
denise austin (41)
alicia keys (38)

Fig. 6. Firefox extension: tree view of heavy query AOL user #671641. The days are set relatives to today.

VIII. Firefox Extension

As a result of the feedback obtained from the mock-ups, we have developed a Firefox extension (Figure 6)6 based on the file tree view to visualize search terms for the individual user. The first levels are the top four search engines (Google, Yahoo, Msn, and AOL) and the most frequent search terms (duplicate search terms are counted and shown in descending order of frequency). The second level groups the terms by the day, the week and the month that the terms were searched. This addresses visualizing queries grouped by when they occurred and queries most frequently used.

We had 10 users use the extension and give feedback by answering the following questions:

- How do you feel about search terms being collected on your computer?
- How do you feel about tools like ”Web History” from Google, which keeps track of your search history, along with what web sites you have visited?
- Which of these do you feel more sensitive about, this extension or the online Web history?
- How else would you want to interact with this tool?
- How effectively could you self-monitor your search activity? What percentage of your queries would you consider sensitive?
- What number of queries do you think this could handle before it becomes too crowded or unusable?
- Would you find any of these functions useful and why?

1) Clearing the entire history of search terms
2) Clear X days of history
3) Right clicking and deleting individual terms.

All were comfortable with search term collection on their computer, and thought it was useful for keeping track of past searches. One did not because he normally prefers history and cache data deleted regularly. Two thirds were uncomfortable with online web history due to personal privacy and possible data disclosure; half of them still thought the function was useful. Of those who were comfortable with web history, most still worry about privacy and being profiled. Two people felt the same about the web history and Firefox extension stating that sensitive key terms could be exposed with both. Two said information is disclosed through other means such as credit card use and online purchases, so query collection was not unique. The rest thought the web history was more sensitive because it is online, and a sense of control was lost. One third thought 50% of their terms were sensitive, one thought all their terms were as complete privacy was desired, and the rest thought none or 5-10% were sensitive.

Everyone felt the extension effectively monitored their search terms and would find it useful as is or with extra features, though one third think its not necessary for home use. Two thirds wanted to keep track of the clicked link after a search term was entered. One third were unsure how many queries would inundate the extension, but most thought it could handle hundreds or more. The nested feature helps to focus on a day even if the total terms are high.

Two thirds wanted to clear the entire search history; the main reason stated was for using a shared computer even within family, e.g., you were shopping online for a surprise gift. Half thought clearing X days would be useful, while the rest thought it made no difference. One thought clearing terms individually was not useful since they want to clear everything and one thought this would be too tedious to use. The rest thought it would be useful, especially for clearing key sensitive terms.

IX. Conclusions and Future Work

This paper presented and expanded on visualization techniques that allow users to self-monitor their information disclosure and laid the groundwork for other visualization and interface designers. Self-monitoring is a powerful tool that raises awareness to the threat and empowers both individuals and enterprises to regulate the amount of information they disclose, minimally impacting web usage.

We have found that self-monitoring is technologically feasible and that users were receptive to our approach in both the mockups and initial Firefox extension. In the future, we see potential for widespread deployment of self-monitoring technologies for both individual browsers and stand-alone enterprise level appliances. We have focused on self-monitoring at the user level, but a logical next step is to extend our research to include self-monitoring of enterprise scale datasets. In addition, while we have focused our current work on search queries, we believe that future tools should incorporate and effectively display data from all potential types of web-based information disclosure including, but not limited to, mapping, patent research, email and online commerce (such as Ebay and Amazon). For a comprehensive list see http://www.google.com/intl/en/options/. Our visualizations provided satisfactory results, well beyond the current browser history function. Our Firefox extension is a first step in self-monitoring using individual browsers.

References

[1] G. Conti and E. Sobiesk, “An honest man has nothing to fear: user perceptions on web-based information disclosure,” in *SOUPS ’07: Proceedings of the 3rd symposium on Usable privacy and security*. New York, NY, USA: ACM, 2007, pp. 112–121.

[2] World internet usage statistics news and population stats. [Online]. Available: http://www.internetworldstats.com/stats.htm

[3] D. Sullivan. Nielsen netratings search engine ratings. [Online]. Available: http://www.internetworldstats.com/stats.htm

[4] G. Gilder, “The information factories,” October 2006. [Online]. Available: http://www.wired.com/wired/archive/14.10/cloudware.html

[5] G. Piatetsky-Shapiro. (2005, December) Interview with usama fayyad, yahoo chief data officer. [Online]. Available: http://www.acm.org/sigs/sigkdd/explorations/issues/7-2-2005-12/fayyad.html

[6] E. Chickowski, “Google to anonymize older search data,” *SC Magazine*, March 2007.

[7] P. R. Clearinghouse. A chronology of data breaches. [Online]. Available: http://www.privacyrights.org/ar/chrondatabreaches.htm

[8] E. Burns. Top 10 search providers. [Online]. Available: BasedonNielsen//NetRatingshttp://www.clickz.com/showPage.html?page=3624821

[9] K. Abdullah, G. Conti, and E. Sobiesk, “Self-monitoring of web-based information disclosure,” in *WPES ’07: Proceedings of the 2007 ACM workshop on Privacy in electronic society*. New York, NY, USA: ACM, 2007, pp. 56–59.

[10] C. Harrison. Searchclock: Visualizing searches over time. [Online]. Available: http://charrison.net/projects/searchclock/

[11] Page addict. [Online]. Available: http://www.pageaddict.com/

[12] Packet garden. [Online]. Available: http://www.packetgarden.com/

[13] Etherpeg project, last accessed. [Online]. Available: http://www.etherpeg.org/

[14] Driftnet project homepage. [Online]. Available: http://www.ex-parrot.com/~chris/driftnet/

[15] D. Maynor and R. Graham, “Data seepage: How to give attackers a roadmap to your network.” Blackhat DC, 2007.

[16] Websense. [Online]. Available: http://www.websense.com/

[17] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson, “Understanding eight years of infovis conferences using paperlens,” in *INFOVIS ’04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS’04)*. Washington, DC, USA: IEEE Computer Society, 2004, p. 216.3.

[18] X. Lin, “Visualization for the document space,” in *VIS ’92: Proceedings of the 3rd conference on Visualization ’92*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1992, pp. 274–281.

[19] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, “Themeriver: Visualizing thematic changes in large document collections,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 08, no. 1, pp. 9–20, 2002.

[20] Firefox plug-ins and add-ons. [Online]. Available: https://addons.mozilla.org/firefox/plugins/

[21] Mork file format. [Online]. Available: http://en.wikipedia.org/wiki/Mork_(file_format)

[22] G. Conti, “Googling: I’m feeling (un)lucky.” Defcon 14, 2006.

[23] C. Benninger. (2007) Finding gold in the browser cache. Blackhat USA.

[24] Squid web proxy cache. [Online]. Available: http://www.squid-cache.org/

[25] G. Sadetsky. Aol search data mirrors. [Online]. Available: http://www.gregsadetsky.com/aol-data/

[26] G. Pass, A. Chowdhury, and C. Torgeson, “A picture of search,” in *Proceedings of the 1st International Conference on Scalable Information Systems*, no. 1, 2006.

[27] S. G. Eick, J. L. Steffen, and J. Eric E. Sumner, “Seesoft-a tool for visualizing line oriented software statistics,” *IEEE Trans. Softw. Eng.*, vol. 18, no. 11, pp. 957–968, 1992.

[28] Many eyes project. [Online]. Available: http://services.alphaworks.ibm.com/manyeyes/home

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, the Department of the Army, the Department of Defense or the United States Government.