Effective Browsing and Serendipitous Discovery with an Experience-Infused Browser

Sudheendra Hangal **Abhinay Nagpal** **Monica S. Lam**
Computer Science Department
Stanford University
{hangal, abhinay, lam}@cs.stanford.edu

ABSTRACT
In the digital age, users can have perfect recall of their digital experiences. In this paper, we explore how this recall can be leveraged during web browsing.

We have built a system called the Experience-Infused Browser that indexes a user's digital history from email and chat archives. As the user browses the web, it observes the contents of pages viewed, and highlights named entities on the page that the user has encountered in the past. This browser has two benefits. First, it highlights terms on the page that occur frequently in the user's archive, effectively personalizing the page for the user. Second, the system can remind the user of names that she has encountered in the past but may not remember.

We evaluated how users reacted to the browser during organic web browsing. Our users have reported that it was useful on crowded web pages to surface content that they otherwise may have missed, and in recalling serendipitous connections to people that they had forgotten. Most of our users said they would use the browser beyond the experimental study, indicating that they derived clear benefit from it.

Author Keywords
Web browsing, Personal Digital Archives, Email, Annotation, Personalization

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous—*Optional sub-category*

INTRODUCTION

In the digital age, the memories of our lives are etched in silicon. Eventually everybody's digital life experiences can be captured in his or her personal archive, leading to the promise of "Total Recall" [4, 7]. This paper explores the question of how we might use this ability of recall to help users in routine tasks, and specifically, in the task of web browsing.

IUI '12, February 14–17, 2012, Lisbon, Portugal.

Users are overwhelmed today by the massive amount of information in pages they encounter on the web, such as crowded news portals, social networking feeds, long web pages, and even entire books. In response, they frequently resort to rapid skimming and selective reading of web pages. We propose to improve web browsing efficacy by exploiting a simple observation: when skimming a piece of text, whether online or offline, terms we know jump out at us and grab our attention, based on factors such as our experiences, memories, interests and cultural background. The same content can evoke very different reactions from different users, based on their personal context.

We propose to enhance a user's browsing experience by infusing the browser with her archive of digitally captured experiences. This approach allows terms of special interest to her to be automatically identified from the archive and to be quickly brought to her attention if they are on a page. In this way, an experience-infused browser can enhance the effect of a user noticing personally relevant terms on a page, and scale it to long pieces of content that are difficult to skim manually. As an additional benefit, the browser can help revive memories from the personal archive related to the web page that the user may have been forgotten.

Personal Digital Archives

A personal digital archive, especially of social interactions, reflects a great deal about the owner of the archive. People use email to communicate with friends or use instant messaging to chat with them. Reportedly, many users spend more time connecting with people on online social networks than in real life [21]. These interactions can often be implicitly and continuously captured in personal archives with no additional effort needed by a user. Over a period of time, the archive captures many entities, such as the names of people, places, and organizations associated with the user. It also contains signals that indicate the relative strength of the user's relationship with these friends or entities [8]. Therefore, the archive captures in some sense a detailed profile of the user, which he can exploit for his own benefit. There are many uses for these archives from reminscence to recollecting information and remembering events [25].

In this paper, we focus specifically on textual digital archives. A major advantage of a textual archive is that it is highly granular and searchable. Therefore automatic tools working on behalf of the user can use the archive to effectively augment and illuminate any information encountered by the user.

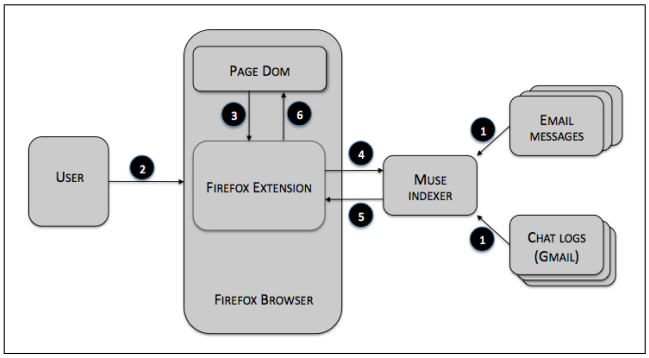

Figure 1. Prototype system design. 1) The MUSE program accesses the user's email archive and chat logs, and creates an index. This index is created once. 2) The user initiates a page load. 3) Our browser extension extracts text from the page Document Object Model (DOM) and 4) posts it to the MUSE index. 5) The MUSE index looks up and ranks named entities in the posted text, and returns them to the extension. 6) The browser extension modifies the page DOM to highlight terms of interest, insert a call-out, and provide links to the actual message contents behind a hit on the page.

In the archiving context, email is particularly significant since it is used by nearly 2 billion people [22] and much of it is frequently archived, especially with the advent of the "Never Delete Anything!" mindset. Email is used as a tool of record, and people often use their email accounts as an informal backup device. It contains rich records in the form of names of friends and interesting people, shopping receipts, travel history, ticket confirmations, movie rental history through services like Netflix, and so on. Tools to analyze this archive can generate a detailed list of terms important to the user, and moreover, can keep the list updated on an ongoing basis with no burden on the user.

Infusing a Browser with the User's Experience

In this paper, we introduce the notion of a browser, *infused* with a user's experience as captured by her personal digital archives. The browser annotates every web page, helping her identify terms of interest without any explicit action on her part. Of course, the user may not always be receptive to annotation, so it is important to minimize interfere with normal browsing, to the extent possible.

Our browser brings users' attention to terms of interest by highlighting them on the page, making their background yellow, as if someone intimately familiar with the user's background and interests had prepared the page for the user and marked it up with a highlighter. This allows users to notice these terms easily as they skim and scroll around the page. In addition, the user can see at a glance all highlighted terms on the page in a subtle call-out row at the bottom of the page. This functionality is especially useful on busy web pages, where a user may be interested in only a fraction of the available content. Of course, highlighted terms are linked back to the text in the archive that contains them, so users can always explore the original messages in case they have forgotten about the terms. Especially for users with long-term archives, this functionality can serendipitously help them recall connections to their past life experiences that have been forgotten.

We note that our current system is limited to highlighting and promoting terms that a user has already seen in the past; in other words, it does not help in discovering new or related content.

Privacy Preserving Personalization

Fig. 1 illustrates the working of a prototype system that we have created to experiment with the concept of an experience-infused browser. The primary components are a Firefox browser plugin and a specialized version of MUSE, a program that analyzes the user's email messages (and instant messaging logs, if available). Both these components run entirely on the user's own computer. This design offers users the benefits of a highly personalized experience without compromising privacy, since no third party is involved.

Privacy is one of the major concerns people have with commercial personalization mechanisms. Apart from allaying privacy concerns, client-side personalization has two major benefits: a) It can be relatively complete, since only the user has the entire view of all his data and b) It removes the burden on each web page or service of implementing personalization.

We invite interested readers to try out our system which is publicly available at the URL:
http://mobisocial.stanford.edu/exp-browser/.

Contributions

The contributions of this paper include the following:

- We propose the concept of an experience-infused browser that brings a user's entire digital archives (perhaps accumulated over a long-term) to bear on the task of web browsing. The browser automatically annotates visited pages by highlighting terms of potential interest to the user. Our approach honors users' privacy by running locally on their own computer, yet provides the benefits of personalization.
- We have created a publicly available prototype of our approach, and report on design and implementation tradeoffs discovered while building this prototype.
- We show that our system is fairly effective in achieving its goals. Seven out of the nine users in our user study indicated that they were interested in using the system beyond the duration of the study. While our study is relatively small and preliminary, we provide qualitative examples to show that a simple engine, infused with a massive amount of personal information, can be useful in many different ways.

RELATED WORK

Total Recall

Recently, the idea of Total Recall has popularized the notion that a lifetime's worth of digital experiences can be recorded and stored [4, 7]. However, most research focuses on the capture, preservation and explicit use of this personal data. Our research goal is to investigate scenarios and techniques

to make implicit use of this data for ordinary tasks such as web browsing.

Web browsing

A closely related system to ours is Rhodes's Margin Notes [24]. Margin Notes takes the paragraph currently being viewed on a web page and tries to find messages in an archive that have similar words in them. While our system employs a broadly similar architecture, our goals and techniques are different, and our primary motivation is in some sense almost the opposite of that of Margin Notes. Margin Notes aims to suggest related material from the archives that a user may have missed. In contrast, we aim to help users rapidly select the few parts of long pages that they may be interested in. Further, the context tested for Margin Notes was to supplement the more or less linear reading of a document; our context is one of skimming rich websites that agglomerate all kinds of names, events and stories on a page, laid out in an arbitrary manner and difficult to pick out manually.

Our differences with Margin Notes's design goals lead to different design decisions in the relevance matching algorithm and user interfaces. Margin Notes tries to find the most closely related message to the words in a paragraph. As we will describe later, our experience-infused browser looks for named entities, which potentially makes our analysis less sensitive to the language used in the text. While Margin Notes shows a sidebar of related content, we use in-place highlighting of terms on the page, so the user can identify them without affecting the normal flow of reading. Our browser inserts a call-out at the bottom with a summary of terms highlighted throughout the web page; this is often used by users as an index to the personally relevant content on the page.

There are several other systems that attempt to display documents similar in content to the page a user is browsing [27]. The Google related toolbar (http://google.com/related) suggests content on other pages that may be related. Another method to focus a user's attention within a page is to highlight the content that has changed on the page since the user's previous visit [1]. There has been research on enhancing browsing experience by flagging irrelevant links and suggesting useful links based on browsing patterns [16]. Tools like SparTag.us [13] let users manually highlight text snippets insitu and allow them to refer to these tags easily. Semantic web browsers like Magpie use ontologies to find out related concepts and point the user to such content [5]. However, none of these systems exploit users' personal archives.

A lot of work has gone into making web browsing effective by considering how users generally browse a particular website and then analyzing this information to guide users in browsing (e.g., [14]). The YouPivot system facilitates searching for contextually associated activities on a desktop computer [10].

Personalization

Many web services attempt to personalize content for a user (e.g. [17]). Gauch et al. survey techniques and related work in this area [6]. While the mechanisms used for personalization are generally opaque to users, these services need to either build up a profile of users' behavior on that site, or ask people to enter coarse-grained profile information up-front. Though personalization can be very useful, users are often hesitant to participate in such services, or provide incorrect data in their profiles, due to the privacy issues and the additional burden of providing information.

Our approach of client-side personalization has benefits in that it does not necessarily require the user to identify himself to the website, and give up personal details. In addition, the personal history used to illuminate the web page can be relatively complete, since it is under the user's control and may aggregate different sources of personal data.

Email analysis

The MUSE system has been shown to be effective in reviving memories using long-term email archives [11]. Systems like Xobni and Rapportive bring in specific content from the web (like a public profile of the message sender) to supplement messages in the inbox; our system can be viewed as an inversion of this model, where email archives are used to illuminate web pages instead.

There has been work done to show emails containing the URL of the currently visited page to the user [18]. However this is limited by the fact that, more often than not, an email message may not contain the exact URL of the page being visited. In contrast, in this paper we present a technique to "intersect" the terms on a web page with the actual contents of the user's email archive and highlight terms that may potentially interest the user. We use standard natural language processing and information retrieval techniques to achieve this goal. Thus our experience-infused browser overcomes some of the limitations of earlier work.

Serendipity

It is known that serendipitous discoveries are important, and moreover that they entertain and captivate users [15, 9]. But it is also acknowledged that serendipity is hard to define, let alone engineer [2], and therefore somewhat hard to study formally. However, in an indicative experiment, Andre et al. found that users rated 1 in 5 results of web searches "interesting", even when they were not "directly relevant".

Ramakrishnan et al. outline the data mining techniques that could be used to discover serendipitous events [23]. Meerkat and Tuba are recent systems designed for promoting serendipity, and the authors make the intriguing observation that *"the design for serendipitous experiences has to some extent disappoint as well in order to be effective"* [12]. Beale presents techniques to model serendipity and interest based on pages recently viewed by the user; the author uses this prior information to highlight interesting links on the page using special coloring mechanisms [3].

USER EXPERIENCE

Our browser augments every page browsed to bring the users' attention to potential points of interest, without interfering with the normal reading of the page. The user can always

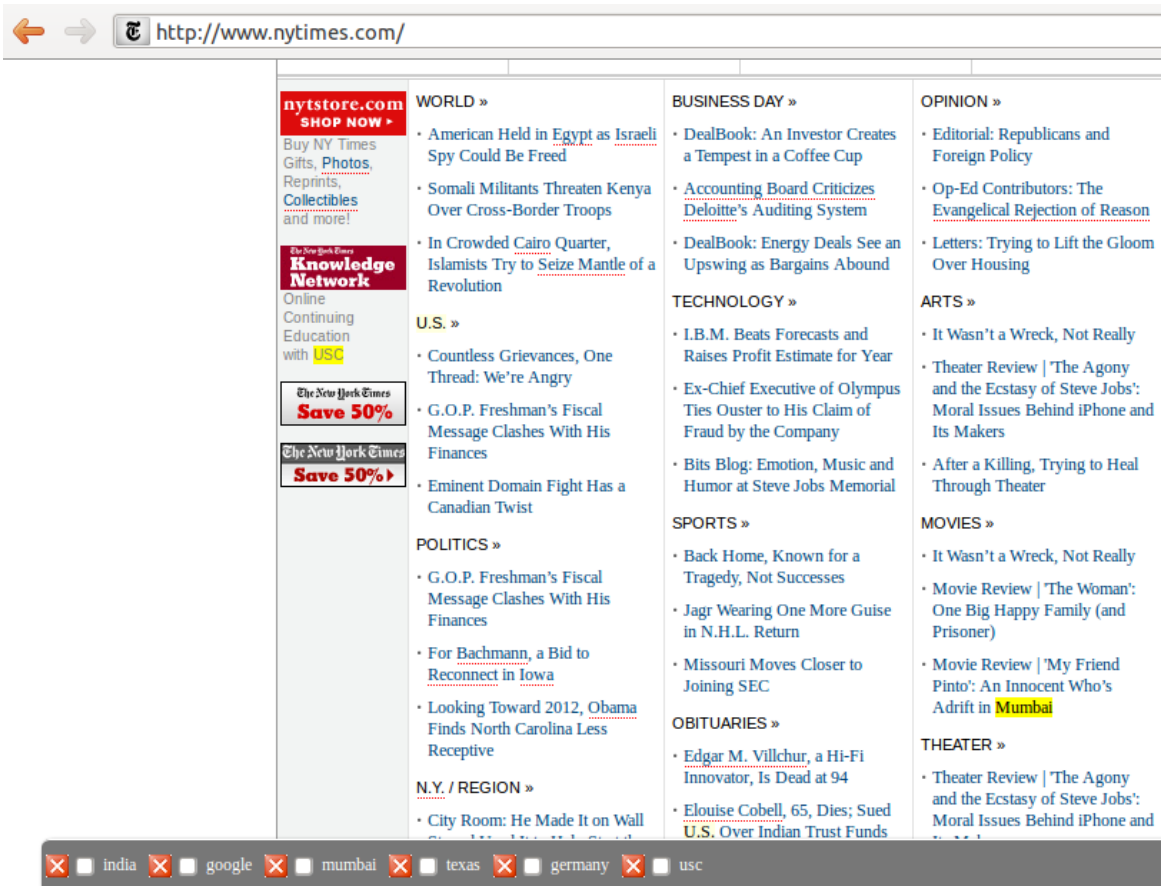

Figure 2. Highlighted terms and a summary of terms in the call-out bar at the bottom of the page.

ignore the highlight and read the page as usual. This experience can be likened to watching a game of American football on television, where the line of scrimmage and the first-down line are highlighted in the field virtually.

Our browser brings the user's attention to terms that are likely to be meaningful to him. Examples of such terms are proper nouns and names of friends or places that appear in the user's email archive. Figure 2 shows a screenshot of the browser on the New York Times home page for one of the participants in our study. Here the terms *Mumbai* and *USC* are highlighted; it turned out that this user comes from the city of Mumbai and had applied to USC for admission.

Even for terms that appear on the page and in the archive, users may have varying degrees of interest and association with them. We attempt to capture the strength of the user's relationship with each term and convey it in the interface; stronger terms can be highlighted more prominently. Our current implementation employs only two levels of highlighting, strong and weak. Terms are scored according to their entity type on the page and their frequency in the corpus, and classified as having either strong or weak interest. Strong-interest terms are highlighted in bright yellow, making them hard to miss as the user scrolls around the page. Weak-interest terms are highlighted in light yellow, allowing the user to see the terms if they are so inclined, but they do not jump out as prominently at the reader. While we currently highlight all matching terms at one of these two levels, it is easy to implement ranking algorithms to avoid highlight overcrowding in case it occurs. The next section describes how scores for a term are computed.

The browser also displays at the bottom of the screen a call-out listing all the terms found in the page that also match the archive. This call-out serves as an index to convey to users what is on the page without requiring them to scroll through the page. This feature is particularly useful for long web pages. One of our users has commented that it feels like a personalized tag-cloud of the terms on the page. As shown in Figure 2, the call-out is shown in a fairly unobtrusive way at the bottom of the page. The user can easily ignore the display of terms if so inclined, or hide it. In addition, the user can dismiss a term by clicking the cross icon next to it, and the browser will remove the term from consideration in subsequent pages.

The browser also lets users find out why it highlighted the terms shown. By simply clicking any highlight terms or a term in the call-out, the user can see a preview of the email messages that contain the terms of interest. This is very important especially for people with long-term email archives,

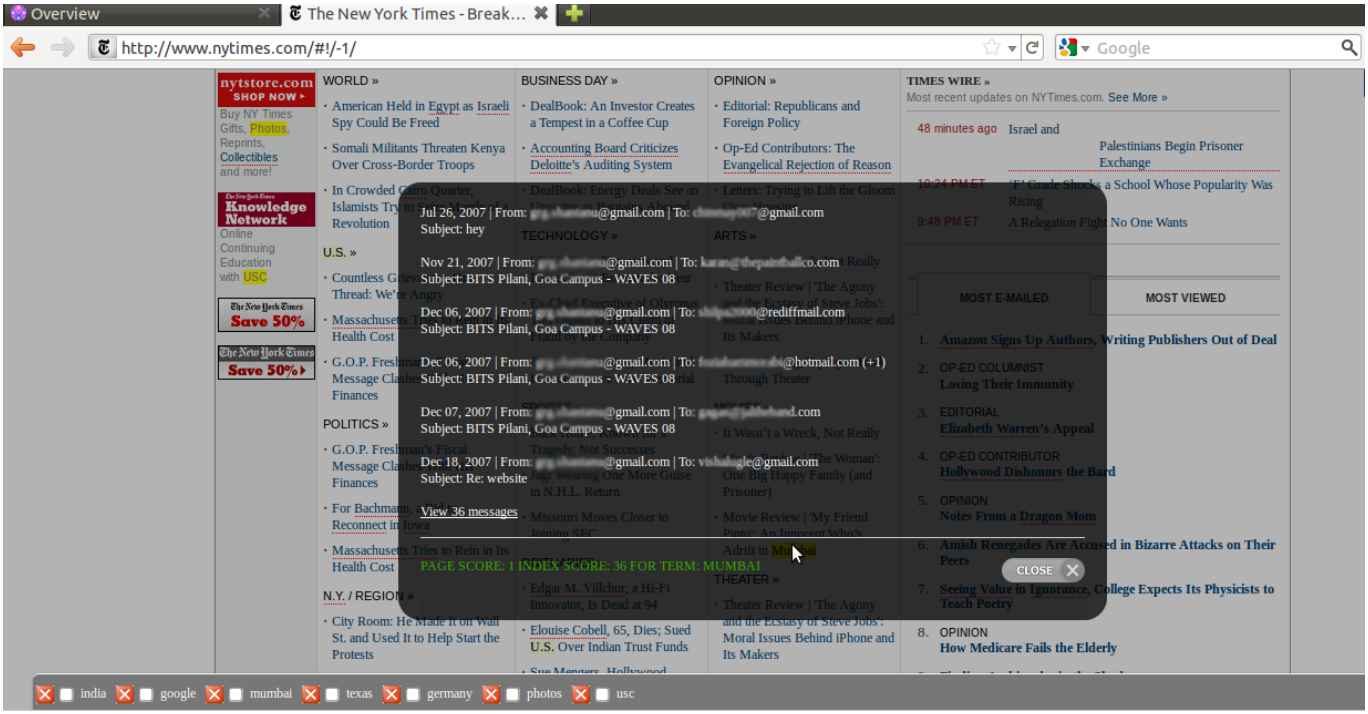

Figure 3. The browser lets users click on a highlighted term to display content in the archive that is related to that term. Each entry can be clicked to view the full message containing the term. Some details are blurred in this screenshot to preserve user privacy.

because they may have forgotten their encounter with the highlighted term. In this way, users can discover old connections and recall forgotten facts. As shown in Figure 3, the browser displays snippets for up to 5 messages. For more, the user has to click on "view messages" and go to a new browser tab. Clicking on any of the message snippets will show the entire message.

The brower applies the highlights immediately after a page is loaded. If page scripts subsequently change its contents, the user has the option to click a refresh button, which re-highlights the page. Users can also hide the call-out box if they find it distracting.

SYSTEM DESIGN

Our system consists of two parts – a background service that accesses and indexes the user's personal history from email files or online email accounts, and a Firefox browser plugin that performs user interface tasks. We chose this separation because the indexing component needs to be highly performant due to the size of personal archives, often running into tens of thousands of messages and spanning hundreds of megabytes of text. The background service is implemented in Java. The front-end part is implemented in Javascript, and is kept relatively lightweight so it can be ported to other browsers in the future. Fig. 1 illustrates the components of the system and the flow when a typical web page is loaded.

The background service is built upon MUSE [11], a system that accesses and indexes email message contents (and chat archives if available through an IMAP interface such as that of Gmail). MUSE processes loosely organized email archives – for example, it can access multiple sources of email (accessed online or stored locally), perform data cleaning and entity resolution, and eliminate duplicate messages. The indexer runs in the background on the user's own computer. It has to run only once, and for most users, indexing is a one-time operation that takes less than 30 minutes. The index is serialized and saved to disk, so that it can be quickly loaded after a computer or browser restart.

To allow easy experimentation, we implemented our browser extension as a Greasemonkey user script[1], in about 1,000 lines of Javascript. After the web page is completely loaded by the browser, the user script extracts page contents and posts it to MUSE that is running on the same machine in the background. We extended MUSE by enabling it to look up the textual contents of a web page in its index. For most web pages, the lookup completes in about a second. The results of the lookup are used to inject Javascript into the page to highlight the relevant terms. In any case, users can begin to interact with the page as soon as it is loaded, without waiting for the highlights to appear.

Named Entities

From our early experimentation, we realized that by blindly matching terms on the page with terms in the archive, it is easy to inundate users with terms of little or no interest. A key design decision we made is to focus on named entities. This is a useful heuristic – while names are a relatively small

[1] `http://www.greasespot.net`

fraction of the text on the page, they capture a large number of personal associations as compared to ordinary words. Therefore, the use of named entities reduces the number of hits on a page to a manageable number, improving precision and reducing noise, while still allowing a relatively high degree of recall.

We extract named entities from the page text using the Stanford NLP toolkit [26]. The named entity recognizer can identify names from a piece of text, and also classify them as likely to belong to a person, place or organization. We have to format contents appropriately in preparation for the recognizer, since it depends on signals such as capitalization, nearby words and whether the term appears at the start of sentence. Therefore, we traverse all HTML elements in the web page, extract text from them, and concatenate them to form the page text. As a special case, we insert full stops if needed after the text extracted from division elements ( ), list elements (<li>), paragraphs (<p>) and headers (<h1>, . . . , <h6>) since it is unlikely that normal sentences would span across these elements.

Ranking and Filtering Hits

To determine the importance of each term that hits in the personal archive, we assign it a score. Through experimentation, we observed that users almost always find people names more interesting than other named entities (such as names of places and countries, which can be fairly generic). We therefore bias the scoring based on entity type. The scoring formula for a term t on the page is:

$$score(t) = w(Type_t) \times |\{d|t \in d\}|$$

where $Type_t$ is the weight assigned to the name type of t, and d is a document in the archive that contains t (whether in its headers or in its body). Our default weight is 1,000 for the person type, and 1 for all other types, strongly boosting the score of person names. To decide whether to highlight the term strongly or weakly, we check whether its score is above or below a threshold (the default threshold is 5). The call-out at the bottom displays terms in order of decreasing score. We expect this scoring function may become more sophisticated in future implementations.

An additional filtering step we found necessary was due to false hits generated by common single-word names (e.g., names like *James* and *Mary*). We eliminate single-word names that belong to the list of each of the 1000 most common (English language) last names, male first names and female first names.

RESULTS FROM A USER STUDY

We designed our browsing system by conducting formative studies with a few lead users and by using it ourselves for routine browsing for a few weeks. When our design was reasonably stable, we conducted a formal round of studies by inviting 9 users (3 female) to test out our system. The users were in the age group 23-29. We deployed the system on the participants' own machines to assure them about privacy. We asked the participants to download MUSE and run it on email folders of their choice, which were typically sent message folders, chat messages if available (via Gmail), and other folders that they thought contained messages interesting or important to them.

Our users had between 2,600 and 21,835 emails and chat logs. We helped them with the retrieval and indexing of their archives which typically took between 2-3 hours. Participants were then asked to install our browser extension. They were given a tour of the features of the system and asked to browse normally for an hour and report any interesting terms that the browser highlighted for them that they otherwise might have missed. They reported these terms by clicking a check-box presented in the call-out at the bottom of the page. They were also told that they could remove any term deemed noisy or irrelevant by clicking the cross icon next to it. After an hour, we collected their browsing logs and gave them a questionnaire asking for their opinion about various aspects of the browser.

We chose this setup to give our experiment broad ecological validity, and because we wanted to learn about the utility of the system on a diverse range of web pages. While the fact that users were aware that their browsing was being logged may have slightly altered their normal browsing behavior, we do not think these changes were significant enough to change our conclusions.

The overall results of this user study were very encouraging. When asked if they would want to continue to use the browser beyond the study period, 7 of 9 users replied in the affirmative. One participant replied "not sure" whereas another wanted to keep using the system "on search and news websites".

Qualitative feedback

The primary goal with this user study was to obtain qualitative user feedback at an early stage of development. We therefore asked users for their detailed impressions of the system, as well as specific examples that did or did not perform well for them. We present and categorize below some examples that are representative of their experiences. Users in the controlled study are referred to as P1 to P9, though we have also included feedback from other users (including the authors) who have informally used the system outside the context of this study.

People names

A common theme was for users to discover the names of someone they either knew directly or whose name they had encountered in the body of an email message. One user, while reading an op-ed article on a news site, found that it mentioned the president of his university. This name was mentioned in the bottom half of the article ("below the fold"), so the user would have missed it had it not been presented in the browser call-out. In this instance, the user had not directly exchanged email messages with the president of the university, but his name was present in email conversations with friends.

P4 was looking at a company website and found a person's name highlighted by the browser. On exploring further, the user discovered that the person had given a talk some years

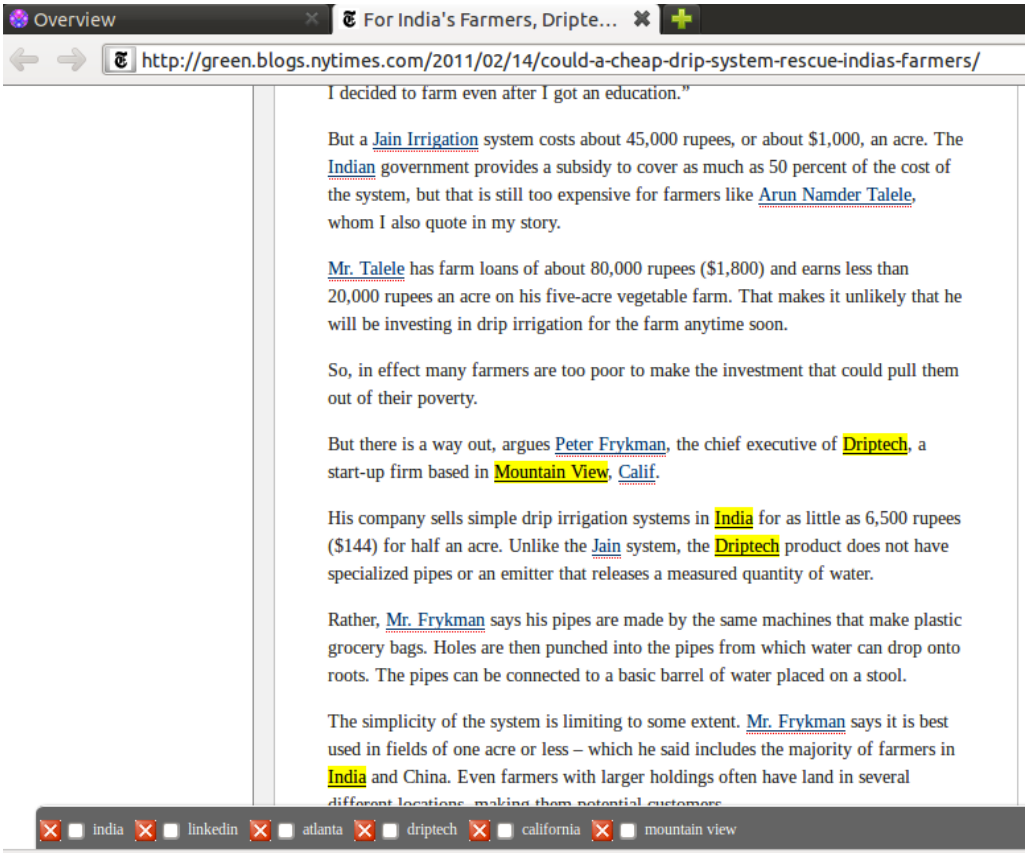

Figure 4. Example of a highlight of a company name that a user had interned at the previous summer.

ago which he had attended. He remarked, *"I would like to have such a tool present everywhere which helps me reach for such hidden information, which I have forgotten about."*

One user who is a professor, while attending a conference and browsing its program web page, found several highlighted occurrences of a person whose name was not familiar to her at all. By examining the relevant email messages, she discovered that the person had applied to her university years ago and was turned down for an interview despite her recommendation. The university should have considered the candidate after all! This story demonstrates that our browser can be effective in illuminating the serendipitous re-crossing of paths.

P2 was browsing a website for applying to a foreign internship and found a testimonial on the website by a person he knew and responded *"I would not have noticed this name had it not been highlighted."*

P9 was reading a page with a list of startup companies that had recently been incorporated when she was surprised to find highlighted the name of a former customer of hers with whom she had interacted 3 years ago.

Yet another user found that the name of the teaching assistant for a course that he was considering was highlighted. Examining the related messages reminded him that the TA had emailed him in response to a campus classified advertisement.

The above examples show that highlighting previously encountered person names, along with the messages in which they were encountered, can be useful in a variety of contexts: personal or professional, either directly known or merely discussed, and well known to the user or forgotten about after a chance encounter.

Products and commerce

P1 found the browser useful as he was browsing an airline's website and found its loyalty program named *Skywards* highlighted. The user responded *"The airlines had emailed me that I have accumulated sufficient miles to get a discount and I had almost forgotten about it."*

Another user found an article on a news site about the car *Toyota Prius*, that he owned. This indicates that personalization based on products owned by a user is possible using names present in ordinary email conversations or in email receipts.

Interestingly, our browser sometimes highlighted relevant words in page advertisements as well. This indicates that it may be effective in highlighting relevant ads without compromising user privacy.

Organization and place names

One user found that a news blog happened to mention Driptech, the name of the company he had interned at the

previous summer, and it helped him notice its presence (See Fig. 4.) Yet another user (who was from India) discovered an article that mentioned Indian cooking in the cooking section of a news site, to which he ordinarily does not pay any attention.

In these examples, users were surprised to find the particular name on a page, and were happy that our browser helped surface information that they otherwise might have missed.

General comments

Participants in the study found our browsing extension an effectively way of viewing a web page through a uniquely personal lens. They found it especially compelling on sites that they usually only skimmed, or that had a lot of content or listings of people names.

P5 said *"I feel like this almost presents me with a personal synopsis of the (web) page."*

P8 also responded along the same lines: *"This tool lets me skim through websites faster."*

P6 mentioned *"I like how it recognizes certain topics that I am interested in–the highlight helps me walk through the site better."*

Users generally liked how the browser helped them find interesting material to read when they were skimming through certain websites. For a particular news site, P9 also remarked, *"After I've got used to it, and know what to expect for this site, it's easy for me to see that there's no new news on it for me today."*

Most users liked the fact that the browser ran completely locally on their own computer, and hence their private information was not being shared with anyone else. P3 said, *"I like the idea of personalization without giving my data to anyone."*

DISCUSSION

Users spend a significant portion of their time browsing the Internet. This browsing activity may be focused on a particular transaction or it may involve exploratory or casual browsing. Many users spend more Internet time on social networking sites and blogs than on reading news [19]. One reason for this may possibly be that social networking feeds are inherently tailored for the user, since a user's friends often reflect his interests. In contrast, most generic news sites are curated by editors for a wide audience, and hence they may fail to get articles of interest across to specific users. We believe that the principle of supplementing organic web browsing with color gleaned from personal history can be very useful in many settings.

One way of thinking about our browsing system is that it effectively creates lightweight "alerts" for thousands of terms in the personal index. When any of these terms are present on the current web page, the browser tries to ensure that the user notices their presence.

We have found that experience-infused browsing is particularly useful for reading news or long blog posts. For instance, the website of The New York Times typically has over 120 stories on its front page. A small number of these are likely to be interesting to any particular user. Our browser can help users identify news stories particularly relevant to them. Often these are stories related to particular places, people and organizations that the user is interested in or affiliated with. Similarly, the browser can be useful for the news site of a small community (such as a campus newspaper) because there is a chance that the user knows specific people in the community who may be mentioned on a particular day. It can also be used to highlight interesting items in social network feeds like those from Twitter and Google Plus.

The results of experience-infused browsing depend on the quality of data present in the personal archive. For example, we recommend that users index both incoming and outgoing e-mail with MUSE. It is useful to include folders containing receipts of transactions such as shopping records, tickets booked, etc. Instant messenging logs are also useful as they reflect real-time discussions among people and encompass a wide range of information, from restaurant reviews to buying a car. We envision that in the future, users will actively enhance and manage their own digital archive. For example, a user can declare interest in all his classmates by importing the class directory into the corpus (perhaps just by emailing it to himself), or a professor could ensure that the names of all the students she has taught over the years are available in her personal index. Similarly, people can save business cards or LinkedIn profiles of their contacts in their digital archive. Now, whenever the name of any of these people is on the web page being browsed, the user will not miss noticing it. This often turns out to useful when looking at a page containing a lot of names – for example, attendees at a large event.

Along the same lines, users may wish to import a curated set of blogs or pages that they trust into the archive. Then they would not need to read and remember them entirely, but the browser could surface connections from the current page to the imported content as a reminder to the user. Perhaps users could subscribe to high-quality mailing lists related to their broader interests that they do not necessarily read manually, but that are used only to build up a corpus of terms related to the topic that can be used to highlight related text on web pages. Students could import the class materials for all their courses into their personal archive, so that when they visit any related web page at any time in the future, the browser can insert a link back to their original class material. Researchers can import their own papers into the personal index, so they can spot related material when reading other papers. These scenarios indicate that total recall is not only possible, it can be very useful in a wide range of applications.

We found that over time, users built up a mental model of the kind of things our browser was good at highlighting on a particular site (for example, a news site that they visited daily). They used our index to quickly guess when there was no useful content, which saved them time.

The experience-infused browsing technique would be particularly useful on mobile devices where screen area is limited and it is relatively difficult to browse a large amount of text. We would like to explore this use case in future work.

LIMITATIONS AND FUTURE WORK

We now discuss the limitations of our current browser, which need to be addressed in future work. Our current implementation has a slight speed impact when the browser is loading a heavy page. One user complained to us about this sluggishness during the user studies. We plan engineering improvements to address this issue in the future.

More importantly, the browser is relatively easily fooled by noise. For example, a relatively common name like *Michael Scott* may result in false hits, unless the user marks and dismisses the term. One way of solving this problem is to filter more terms; another is to look for co-occurring names to disambiguate each other. For example if the personal index contains messages from a certain *Michael Scott* working at *Dunder Mifflin* (and therefore both names frequently appear together in the archive), the browser could highlight the name *Michael Scott* on a page only if *Dunder Mifflin* also appears on it. Similarly, place names can be broad and non-specific; highlighting names like *U.S.A.* is probably not useful in most circumstances.

While email and chat logs are conveniently accessible and capture a rich fraction of personal history for many users, a more comprehensive archive would incorporate more social streams such as those from Facebook and Twitter. Eventually, one could also consider importing a subset of web pages the user has visited (and perhaps spent a minimum amount of time on) in a browser into the archive. Needless to say, maintaining the security of a comprehensive personal digital archive is essential. Studying ways to integrate and score terms encountered via disparate sources and at different times in such an archive is likely to be an interesting challenge.

Currently our browser does not change page layout in any way. It also accesses content only within the current page. This could be expanded to items that may be of interest to the user but are hidden deeper down on related pages. Some of our users commented to us that they would like the browser to chase navigation links from the front page of a news portal into its various sections to uncover stories that they might be interested in. A useful approach might be to identify relevant articles for the user from a news feed of all articles, and then generate a customized newspaper by laying out just the relevant articles on a single page.

Of course, not everything that the browser highlights, even if accurate, is necessarily useful, and it may not always be desirable to only promote terms that the user has already encountered in the past. Personalization can lead to online "filter bubbles" [20]. More work is needed to determine how, when, and to what extent an experience-infused browser should intervene. We hope that deploying systems such as ours will shed more light on this problem.

CONCLUSIONS

We have found that the experience-infused browser is useful to let users efficiently browse textual content by highlighting personally relevant named entities in web pages. Our approach of using personal archives as a way to capture a user's experiences and interests appears to be effective, and the technique of matching named entities on web pages is a useful heuristic. From our user studies we see that deploying these archives in the context of web browsing can be a valuable tool for improved personalization and web browsing using purely client-side mechanisms. This is a significant departure from personalization based on user profiling by specific services.

The prototype we have developed appears to strike a good balance between surfacing interesting information and not being obtrusive. Users may not find interesting highlights on every page, but every so often it augments the browsing experience in interesting ways. Users find material that they may have missed, notice friends where they did not expect and recall people that they have long forgotten. This notion of infusing everyday tools with our digital life experiences can potentially be used in many other settings.

Serendipity is, by definition, rare [2], so how does one build a tool that surfaces potential serendipitous facts without inundating users with irrelevant or obvious information? Instead of just listing all possible connections, our tool uses the information in the archive to highlight terms familiar to a user, thus providing the primary benefit of more efficient browsing while imposing little cognitive overhead. Users can click on highlighted terms only if they are curious about the connections to those terms. The general idea of supporting serendipitous discovery in everyday tools can be applied to other contexts.

ACKNOWLEDGEMENTS

We thank Peter Chan for useful discussions, our study participants, and the anonymous IUI reviewers for excellent feedback. This research was partially funded by the NSF POMI 2020 Expedition Grant 0832820 and the Stanford MobiSocial Computing Laboratory.

REFERENCES

1. E. Adar, J. Teevan, and S. T. Dumais. Resonance on the web: web dynamics and revisitation patterns. In *Proceedings of CHI '09*, pages 1381–1390. ACM, 2009.
2. P. André, M. C. Schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: Designing for (un)serendipity. In *C&C 09: Proceedings of the seventh ACM conference on Creativity and Cognition*. ACM, 2009.
3. R. Beale. Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. *International Journal of Human-Computer Studies*, 65(5):421–433, 2007.
4. G. Bell and J. Gemmell. *Total Recall: How the E-Memory Revolution Will Change Everything*. Dutton Adult, 2009.

5. M. Dzbor, J. Domingue, and E. Motta. Magpie towards a semantic web browser. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *The Semantic Web - ISWC 2003*, volume 2870 of *Lecture Notes in Computer Science*, pages 690–705. Springer Berlin / Heidelberg, 2003.
6. S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 54–89. Springer Berlin / Heidelberg, 2007.
7. J. Gemmell, G. Bell, and R. Lueder. MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1), 2006.
8. E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of CHI '09*. ACM, 2009.
9. J. Gritton. Of serendipity, free association and aimless browsing:do they lead to serendipitous learning? `http://www.education.ed.ac.uk/e-learning/gallery/gritton_serendipitous_learning/conclusion/assets/assignment_print_version.pdf`.
10. J. Hailpern, N. Jitkoff, A. Warr, K. Karahalios, R. Sesek, and N. Shkrob. Youpivot: Improving recall with contextual search. In *Proceedings of CHI '11*. ACM, 2011.
11. S. Hangal, M. S. Lam, and J. Heer. MUSE: Reviving memories using email archives. In *Proceedings of UIST '11*. ACM, 2011.
12. J. Helmes, K. O'Hara, N. Vilar, and A. Taylor. Meerkat and tuba: Design alternatives for randomness, surprise and serendipity in reminiscing. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction*, INTERACT'11. Springer-Verlag, 2011.
13. L. Hong, E. H. Chi, R. Budiu, P. Pirolli, and L. Nelson. Spartag.us: A low cost tagging system for foraging of web content. In *Proceedings of the working conference on Advanced Visual Interfaces*, AVI '08. ACM, 2008.
14. T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the 1997 IJCAI*, August 1997.
15. J. Lawley and P. Tompkins. Maximising Serendipity: The art of recognising and fostering unexpected potential - A Systemic Approach to Change. `http://www.cleanlanguage.co.uk/articles/articles/224/1/Maximising-Serendipity/Page1.html`.
16. H. Lieberman, C. Fry, and L. Weitzman. Exploring the web with reconnaissance agents. *Communications of the ACM*, 44:69–75, August 2001.
17. J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of IUI '10*. ACM, 2010.
18. M. Mesarina, J. Jain, C. Sayers, T. Close, and J. Recker. Evaluating a personal communication tool: Sidebar. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*, pages 490–499, Berlin, Heidelberg, 2009. Springer-Verlag.
19. Nielsen. Social Media Report: Q3 2011. `http://blog.nielsen.com/nielsenwire/social/`.
20. E. Pariser. Beware online "filter bubbles", March, 2011. `http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html`.
21. B. Quinn. Social network users have twice as many friends online as in real life, May 08, 2011. `http://www.guardian.co.uk/media/2011/may/09/social-network-users-friends-online`.
22. Radicati Group Inc. Email Statistics Report, 2010-2014. `http://www.radicati.com/?p=5282`.
23. N. Ramakrishnan and A. Grama. Data mining: from serendipity to science. *Computer*, 32(8):34 –37, Aug 1999.
24. B. J. Rhodes. Margin notes: building a contextually aware associative memory. In *Proceedings of IUI '00*. ACM, 2000.
25. A. J. Sellen and S. Whittaker. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 53:70–77, May 2010.
26. The Stanford Named Entity Recognizer. `http://nlp.stanford.edu/software/`.
27. D. Wolber, M. Kepe, and I. Ranitovic. Exposing document context in the personal web. In *Proceedings of IUI '02*. ACM, 2002.