Using Web Activity to Support Collaboration in a University Environment

Roman Stanchak
Dept. of Computer Science
University of Maryland
College Park, MD, USA-20742
roman@cs.umd.edu

Saket Navlakha
Dept. of Computer Science
University of Maryland
College Park, MD, USA-20742
saket@cs.umd.edu

ABSTRACT

Interdisciplinary collaboration is an important component of academic research. It serves as a means to pool physical and intellectual resources, gain social status and knowledge, and bridge ideas from different areas. However, it is difficult to find such collaborators because many researchers do not know what others are working on, nor how their interests might overlap. This problem is paramounted by the fact that office spaces usually consist of people with homogenous sets of interests. In this work, we attempt to support collaboration amongst the various subfields within the computer science department at the Univerity of Maryland (UMD). To this end, we introduce UMDRecorder, a Firefox browser plugin that captures user web activity on "educational" domains. We attempt to use web activity to automatically generate user profile pages which summarize user interests, and match users to other similar users based on these interests. We also employ data mining techniques on top of this data to structure the data into relevant "topics". Our tool was used by 9 graduate students at UMD for 6 weeks. The results of our evaluation study suggest that web activity is a useful indicator of a user's research interests, but more sophisticated data mining techniques are required to filter and organize the information in a meaningful way.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation:Group and Organization Interfaces**]: Collaborative computing, computer-supported cooperative work

General Terms

Human factors, Design

Keywords

Scientific collaboration, creativity

1. INTRODUCTION

Since the very foundation of the scientific method, scientists have formed alliances with others to work towards a common goal. Indeed, scientists seek collaboration for a variety of reasons, including (but not limited to) [19]: the pooling of physical and intellectual resources, improved enjoyability and quality of research, reinforcement of existing relationships, and as a means to gain status and knowledge. All of these reasons are important, but from a macroscopic perspective, every scientist is evaluated on the innovative ideas he produces. In *How Breakthroughs Happen*, Andrew Hargadon argues that innovation is the result of recombining ideas from different areas in a new way [15]. More specifically, innovation occurs when the network of people, ideas and objects surrounding a set of technologies recombines in a novel way to produce a fundamentally different structure [15]. Hardagon examines a number of organizations renowned for their innovative ability, ranging from Thomas Edison's Menlo Park corporate research lab to the technology brokerage firm IDEO. From them, the characteristic Hargadon identifies as being integral to innovation is a tight group of collaborators with expertise in a wide variety of disciplines. Today, a large number of scientists are housed in universities.

Although Hargadon does not explicitly address research being performed by scientists in universities, if his argument is accurate, then there are wide-ranging benefits to interdisciplinary university collaboration.

Moreover, the importance of collaboration can be easily inferred from the study of scientific collaboration networks. These networks can be thought of as a graph with nodes as authors and an edge between two authors if they have collaborated on at least one paper together. In [23] Newman presents several scientific collaboration networks, and for each computes the average number of authors per paper. His results are summarized in Table 1. Clearly, most papers have at least two collaborators, and sometimes many more. Although we can see which fields these papers lie in, Newman does not tell us which fields their authors came from. It could be the case that a physics paper was written by collaborating authors in different fields. Nonetheless, it is safe to say that collaboration indeed exists and is often a necessary component of successful scientific research.

In order for collaboration to occur, however, some form of contact must occur [19]. Traditionally, this initial contact has occurred through simple physical or geographic proximity, institutional norms like social events, or through special events such as conferences [19]. On the surface then,

universities appear to be the ideal place for collaboration. However, the typical university hierarchy is actually sub-optimally configured. Each department is housed in a separate building: the mathematicians are segregated from the physicists; the computer scientists are segregated from the biologists. This same hierarchy also holds true at the level of individual departments. For instance, in the computer science department at University of Maryland (UMD; where our study is focused), each research group is segregated in their own office space. This makes it easy to bounce ideas off like-minds, however, it also creates a physical separation that makes it difficult to (1) build relationships with others; (2) recombine ideas from disparate fields; and (3) learn what other people in the department are working on. For example, how can a student in artificial intelligence find where his expertise overlaps with computer vision problems? This requires the student to become aware of computer vision problems, techniques used therein, and the people working on them. We found, however, that most people do not know what others in the department are working on, sometimes not even their friends. By knowing what other people are working on and thinking about, everyone can begin to try and make these border-crossing connections that fill structural holes and bridge small worlds [15]. This can be especially useful for younger graduate students who have not yet found a niche field to work in.

Remedying the existing sub-optimal configuration of the typical university at the organizational level is probably "the right way" of approaching the problem, but is difficult because these structures have been established for decades. An alternate approach to physical reconfiguration is creating virtual connections. As the norms for communication grow increasingly virtual (email, the Web, VOIP, etc), it is logical that these forms of communication will augment the traditional forms of collaboration [18].

In this work we address the difficulty in finding collaborators by using web activity to automatically capture a user's research interests. Intuitively, a user's web activity is a good method to capture research direction and thoughts because researchers are constantly searching for and reading work related to what they are thinking about. We then apply data mining techniques to the web activity and populate user profile pages which identify connections between users.

With the recent emergence of online communities [12] [21], it's possible that student-and-professor online profiles would serve as a better way to connect users virtually. There are some problems with this approach, however:

- People might not know whose profiles are important to visit.
- People might not know what is important to include in their own profiles.
- People might not have time nor the will to browse around and find collaborators[m/v link].
- It is disruptive, and breaks flow.
- It requires extra work which users may forget or not prioritize.
- It requires a change in habit.

Our method, on the other hand, attempts to automatically identify potential collaborators with minimal effort from the user. Existing tools are either labor intensive or targeted at general audiences [32]. This is the first work we know of that attempts to leverage web activity to aid collaboration in a university setting with minimal input from the user.

We completed a study with 9 users from the computer science department at UMD, and find that web activity is indeed a good method of capturing research interests. However, because our tool does not require manual input from the user, it contains more noise than otherwise. To improve quality, more data mining techniques are needed.

[REDO AT THE END] The remainder of this paper is organized as follows: Section 2 motivates our investigation of collaboration in a university setting. Section 3 briefly reviews related work and other collaboration and creativity support tools in the literature. Section 4 discusses the design goals for the proposed tool. Section 5 describes how we automatically capture a user's interests and thoughts. Section 6 describes how we use this data to connect two users together. Section 7 presents details about our experimental setup and a summary of the data we've collected. Section 8 presents results on our user evaluation study. Section 9 provides a discussion, conclusion, and ideas for future work.

2. RELATED WORK

Work related to our approach spans several different areas, including social-bookmarking, bibliographic citation tools, collaborative filtering, expert-recommendation, as well as general collaborative creativity support tools.

Social-bookmarking websites typically allow users to save bookmarks to other websites and share them with other users. Some, such as `del.icio.us`, rely on user-submitted *tags* to categorize content [10]. Others, such as Digg [11] and Reddit [25], rely on user voting to identify interesting content. Some social bookmarking services targeted at scientists are CiteULike [8], Connotea [9], Bibsonomy [5], and 2Collab [1]. In addition to the ability to share and tag bookmarks, these services can also be used as citation management tools. Most of these services can automatically extract relevant contextual information about scientific publications (such as authorship, year of publication and venue), and export citations in a variety of popular formats (BibTex,RefWorks,etc). None of these services offers explicit recommendation of similar users or publications, though one can typically browse users and publications associated with particular tags. Moreover, none of these sites has any facility to transparently monitor the sites a user visit; they all require the user to explicitly create a bookmark and enter tags.

Desktop bibliographic citation management tools have richer citation management features, but lack the social features. Some notable commercial desktop citation-management tools are RefWorks [26], EndNote [27], Reference Manager [29], ProCite [28], and Biblioscape [16]. Some open-source tools are Bibus [6], JabRef [17], Pybliographer [24] and Zotero [37]. Some of these tools do automatically extract contextual information from publisher's web sites, but do not offer collaborative capabilities.

Collaborative filtering is an approach that bases recommendation of items (users, products, services, etc) upon the habits of similar users [35]. A variety of e-commerce services use collaborative filtering to help users find similar products, people or media. For example, Amazon.com [2] has a feature

Dataset	Avg. # authors/paper
MEDLINE	3.75
astro-ph	3.35
cond-mat	2.66
hep-th	1.99
SPIRES	8.96
NCSTRL	2.22

Table 1: Scientific collaboration statistics, from Newman[23].

Customers Who Bought This Item Also Bought..., which lists other items commonly purchased together with the current item. The social-networking website Facebook [12] has a similar feature *People You May Know*, which suggests people a user may know but is not yet friends with, based on common connectivity among the user's friends. `stumbleupon.com` is a web service which uses collaborative filtering to recommend web pages [32]. Our approach utilizes aspects of collaborative filtering to identify similar users, but differs from the work mentioned above in that document content is also utilized.

CodeBroker is a system to help software developers find relevant portions of code in public software repositories [14]. This is similar to the proposed work in that it increases awareness of related projects and their respective authors, but is limited to programming. Our work targets a more general scientific audience.

Another area of related work is in automated expert recommendation. For instance, in the domain of Java software development, Expert Finder recommends people based on the quality and type of code they have written [34]. This is similar to the proposed work, but rather than identifying *experts*, we are interested in identifying *collaborators*.

Finally, the proposed work falls in the general category of collaborative creative support tools. Some related approaches follow: The *Envisionment and Discovery Collaboratory* [3] (EDC) is an environment where people collaborate to solve problems, but focuses less on traditional computers and more on other forms of technology such as whiteboards, games, and physical prototype modeling. The *Caretta* [33] system is similar to the EDC but incorporates feedback into the design process by having users negotiate and mutually agree on certain designs. Caretta allows users to reflect individually on problems, and subsequently come together into a shared space for group discussion. [22] describes two systems for graphic design collaboration. The basic idea is that designers can attach words and images to design ideas in a shared whiteboard in order to inspire collaborators. In contrast to the proposed work, each of these tools assume collaborators have already been found and mainly serve to make the collaboration experience more productive.

3. APPROACH

We propose UMDRecorder, a tool that aims to help collaboration in university environments by automatically capturing user's interests and connecting them with other users with similar interests. A user's interests are captured by recording their web activity using a simple web browser plugin. This data is used to create *User Interest Profiles* which represent a snapshot of the user's current and historical interests. By applying data-mining techniques to the browsing history and document content, we aim to identify individuals with similar or complementary interests.

3.1 Design Goals

The design of our tool was motivated by three major goals. The tool should (1) automatically infer a user's interests, with (2) minimal interaction/disruption, and (3) match the user's interests to those of similar users. The motivation for these goals is described below:

1. **Automatically Infer a User's Interests**
   A user can easily set aside time to record and share his/her interests, but a system that automatically does so would give scientists more time to pursue actual research. In this work, we use the record of academic web sites a user visits as an indication of his interests.

2. **Minimize User Disruption**
   The user should have control over the tool at all times. Losing control implies that the user is no longer embedded in the task domain, which implies that he's thinking not about what he's doing, but about how the software works. Our tool captures a user's web activity in the background, enabling him to focus on the task at hand.

3. **Match a User's Interests to those of Similar Users**
   In many cases, researchers are physically separated; they may be in different countries or simply different parts of the same building. We aim to bridge this separation by automatically connecting users based on the academic web content they view. These connections can be utilized by users to explicitly find collaborators or to simply increase awareness of other scientists with similar interests.

3.2 System Overview

In this section we describe a system overview of our tool, and some details about the current implementation.

3.2.1 Capturing A User's Interests

The primary method we propose to *automatically* capture a user's current interests is based on their web activity (reference webpages they visit and academic papers they read. One of the first steps required in any research project is a comprehensive literature review. The internet has evolved into a powerful research tool where a simple search phrase can nearly instantly provide access to a wealth of information. Most traditional academic journals now have online database indexed by popular search engines, and a variety of open access online journals are gaining popularity.

This trend toward universal document access from the web, the instant gratification provided by search engines, and the ubiquity of internet access suggests that scientists will increasingly use this mode of research. Therefore, observing the academic web sites and papers a scientist downloads provides a window into his/her thought process.

There are two primary means by which this can be accomplished: from the client-side or server-side. Client-side monitoring implies that every user must install a small program that performs this monitoring, and then periodically uploads the user's history to a central server. This central server then stores these statistics for all the users. The disadvantage to this method is that a user must explicitly choose to install the software, and it will utilize some amount of resources on his/her computer. Server-side monitoring is more transparent to users, but would require setting up a proxy server that filtered HTTP requests, and must therefore occur on an organizational level.

As a proof of concept, we chose to implement client-side monitoring using a web browser plugin. We targeted the popular Mozilla Firefox web browser based on our own preferences, and the availability of the open-source Attention Recorder [4] plugin. Attention Recorder records which websites are visited, and is capable of uploading this *Click Stream* to a central *Attention Server*. However, asking users to provide us with their entire browsing history might (understandably) make some users uncomfortable, so Attention Recorder was modified to only record visits to a subset of network domains. This subset is primarily limited to: reference sites such as Wikipedia; online journals and citation indexes such as Citeseer, Arxiv, and PLOS; and common academic paper formats on academic (`.edu`) domains. For a complete list, see in Table 2. This means that only sites visited which match a prefix from Table 2 will be recorded. This list was generated based on personal experience, and by recommendations from several graduate students in various computer science subfields.

Data is automatically sent to our centralized server running a MySQL database. The list of IP fields stored for each visit are listed in Table 3. Notice how our tool requires no manual input from the user to capture his thoughts or interests. After installing the plugin the user will be asked for his username (see Figure 1), after which no other intervention is required. If a user wishes to *become invisible*, our plugin can be temporarily disabled by clicking on the green diamond in the Navigation Toolbar (see Figure 1).

3.3 Data Presentation / Connecting Users

Ultimately, the goal of our tool is to connect a pair of users based on some mutual content that both users have expressed interest in. In the previous section we have described how we captured expressed interest. In this section, we describe how we use the information collected to connect a pair of users. Our approach is based on automatically creating user profile pages populated with the web activity data collected thus far.

3.3.1 *User profile page*

Every user of our tool has a profile page which serves as a docking point where his connections and web activity are listed. Below we show a series of screenshots displaying the functionality of profile pages. Figure 2 shows the home page, where each user is listed along with a count

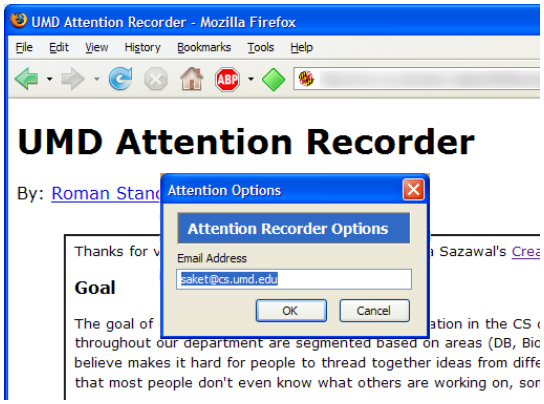

Figure 1: Plugin installation, asking for username.

of the number of webpages captured by our plugin. Some users' names are anonymous, by request of our participants. Anonymous users can still be contacted, however; We can create an "anonymized" email address that secretly maps in the backend to the user's true email address. This is the same scheme that Craigslist uses.

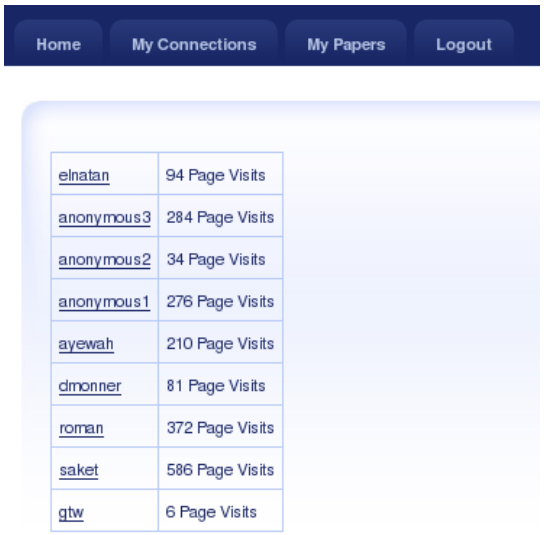

Figure 2: Homepage with all profiles listed along with # of page visits.

Figure 3 shows the user profile page for *roman*. This page has three components:

1. **Most recently browsed**: this corresponds to the user's most recent activity. A complete list can be viewed by clicking on "See More".

2. **Similar Users**: a summary of connections or common page visits, displaying the number of common pages that every other user had with *roman*. This list is generated as follows: At the end of every day, our tool looks at all the pages visited by user u_1 during

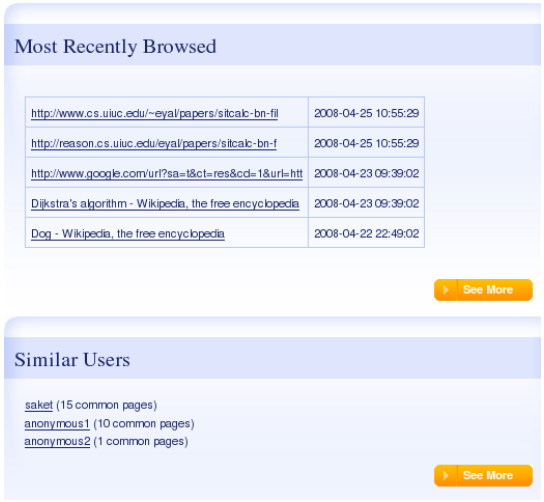

Figure 3: User profile home page, with three components.

the day and checks to see if any of those pages have already been visited by another user u_2. If so, on u_1 page a connection is automatically listed with u_2 along with the common page. When "See More" is clicked, a detailed list is brought up listing the actual common pages; see Figure 4.

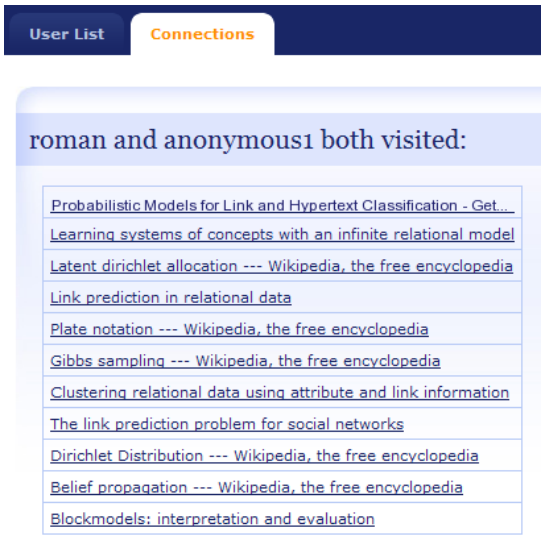

Figure 4: Other users who have visited the same pages as *roman*.

More generally, we hypothesize that people will find it useful to be able to see what kind of things others are working on, regardless of whether it represents a connection. If user u_1 can see what other papers user u_2 has read, even if u_1 hasn't explicitly read that paper or a similar paper, it might be interesting nonetheless as a source of new and different knowledge. So, each page is populated with a summarized list of the page owner's web activity. In general, this enables users to *browser and discover* new content that they would have otherwise missed. It is often the case that people don't know exactly what they are interested in, but by seeing what *other* people are working on they can begin to develop their own sense of interest. A list of each user's web activity can be seen by clicking on the user's username from the home page; see Figure 5.

Figure 5: All papers for *roman*

We found that simply searching for users who visited the *exact* same website or read the *exact* same paper was too stringent and resulted in sparser-than-expected connectivity. However, this lack of exact connectivity does not necessarily mean that there is nothing in common between two viewers. A more complete model would also consider the *content* of the documents viewed.

3.4 Document Content Analysis

In order to identify more general commonality between users than simply looking for common page views, we also consider the content of the documents they viewed. One method, called *Topic Modeling*, takes a generative probabilistic approach to document representation [31]. The basic idea is that a document consists of a mixture of *topics*, each of which are associated with a distribution of words that are likely to occur. A document of N words can be projected onto a K dimensional topic-space to provide a low-dimensional representation of the original document. Typically K is around 100, compared to N of around 10,000 or more.

In our work, this low-dimensional representation is used to find the 10 nearest-neighbors for each document. Figure 6 shows an example of the nearest neighbors of a document, along with some basic information (title, author, URL, Bibtext entry), a list of others users who have browsed that page, and a list of similar documents. In our implementation we use an approach called Latent Dirichlet Allocation

Figure 6: Each document has a profile page with bibliographic information (such as title, authors, publication venue and year), who has viewed the document, and similar documents. Document similarity is based on full content analysis using Latent Dirichlet Allocation.

(LDA) [7] provided in UCI's Matlab Topic Modeling Toolbox [30]. We run LDA at the end of every day on the new set of pages visited, and then identify new connections.

4. DATA AND EVALUATION

We captured web activity using our UMDRecorder plugin for six weeks; from the end of March to the middle of May 2008. Users were selected by mass-emailing the UMD computer science department social listserv for participants. The email included a link[1] to our project webpage which contained a description of the goals of our project, our approach, information on how to download and install our plugin, and a brief privacy statement which said that we would not share the data without getting consent first.

7 users responded to the email and installed our plugin. In addition to these 7 users, the authors also allowed their web activity to be recorded. Collectively, the 9 users come from the following computer science sub-fields: neural networks (2), relational AI (1), software engineering (1), bioinformatics (1), programming languages (1), semantic network (1), computer vision (1), and HCI (1). Our final data set consisted of approximately 2,600 "educational" webpage visits, with an average of 290 webpages visited per user.

[1]http://www.cs.umd.edu/∼saket/UMDRecorder

Following data collection, we built the user-profile-page front-end and emailed our 9 users with the link. In the email, we briefly described the features of the user profile page, and included a survey questionnaire at the end. Below are the questions we asked, along with a summary of their answers.

1. How important do you think it is to collaborate with those outside your own research group? (1 = not important at all; 5 = very important)

1 2 3 4 5

The average response was 4.28, indicating that people do indeed value collaboration and find it a necessary component of academic life.

2. How active are you in finding collaborators? (1 = not active at all; 5 = very active)

1 2 3 4 5

The average response was 2, indicating that most people remain mostly inactive when it comes to finding collaborators.

3. How difficult is it to find collaborators? (1 = very difficult; 5 = very easy)

1 2 3 4 5

The average response was 2. Putting the first three questions together we find that people find collaboration very useful, but appear to not put much effort into finding collaborators because it is "difficult". There is a clear need then for collaboration support tools, which our work attempts to provide.

4. Our primary project goal was to help people find collaborators. How well do you think our project fares to this end? Note: we realize that we only have a few users currently. But consider the question if, say, we had half the CS department participating.

Responses to this question varied. Most people believed that the approach showed promise, but the profile pages lacked structure. In particular, people wanted to see more data mining which not only identified the topics of webpages, but also the relationship between topics. For example, one of our users is interested in studying static analysis tools, but would like to find users who are interested in other related things, such as web development or embedded computing. Another user thought that some of the web activity was too specific and needed to be better generalized. This user also thought that it would be useful if we could somehow capture a person's intent when visiting a page; in other words, if someone visited a page on "Gibbs Sampling" it'd be useful to also know what the user wants to use "Gibbs Sampling" for. One person thought looking at papers read to identify collaborators was not a very good idea at all.

5. Our secondary goal was to increase awareness about what kind of work is going on in our department. This not only includes recent publications, but also the research direction and thoughts people have as they try to solve problems. How well do you think our project fares to this end? Does web

activity capture this effectively?

Most everyone believed that web activity is one of the best indicators for gauging interest, short of asking people directly. Most also thought it increased awareness, but would be helpful if we filtered out irrelevant sites, and tried to group websites into better-defined topics. One person thought more data was needed to make a conclusion.

6. Even if you do not have many direct connections with others, how useful is it to browse the site and discover what sites people have visited? What could you gain from this?

Most people said it was too time consuming to browse through the list of other sites, and could be improved by making the list more concise. Also, people did not like how some sites had no title and only a long URL associated with them. The only way to figure out what the site is about is by clicking on it, which is slow.

7. What other features would you like to see presented on profile pages?

There were several good suggestions, the most dominant of which was to cluster, classify, and present the data so that it can be presented based on more meaningful semantics. Other suggestions: add contact information for each user; add RSS feed of content; add another section on "most frequently visited" sites; incorporate an alert-based system which notifies the user of connections while he browses.

Overall, the survey results suggested that scientists (1) thought collaboration was important; and (2) thought finding collaborators was difficult. Furthermore, most participants believed that using web activity to automatically infer interests was a promising approach for finding potential collaborators, and generally increased awareness of what other researchers were working on. Finally, all responses agreed that our current presentation of the data was too difficult and time consuming to be of practical use, and gave several useful suggestions for remedying this problem.

The general goal of this work was to help scientists identify potential collaborators. Andrew Hargadon's theory, as discussed in the Introduction, suggests that interdisciplinary collaboration may be the most fruitful. Unfortunately, the small number and homogeneous background of our participants makes it difficult to draw any conclusions about the usefulness of our tool in identifying collaborators in disparate fields. However, given the difficulty our participants had identifying common topic areas, it is probable that researchers in different fields will have similar or greater difficulty. Further research and experimentation is needed in this area, and is discussed in Section 6.

5. CONCLUSION

We have presented a software tool which supports collaboration in the university environment. Andrew Hargadon [15] in *How Breakthroughs Happen* argues that innovation often occurs when ideas from different fields are synthesized. Recombinant innovation, however, is only possible when experts from various fields are able to intermingle and share knowledge with each other. This is difficult to do in the computer science department at the University of Maryland because researchers are homogenously placed such that they only share office space with others in their same field. Our tool attempts to bridge this physical divide by capturing user's web activity, and identifying pairs of potential collaborators based on the types of webpages visited.

To this end, we have written a Firefox plugin called UMDRecorder which sends information about user website visits on "educational" domains to a centralized server. With this data, we automatically populate user profile pages for each users which contains recent webpages visited by the user, and common webpages visited with other users (used to find potential collaborators). Common pages themselves, however, resulted in fewer than expected connections amongst users. As a result, we used a topic modeling algorithm to generalize the webpage from a unique identifier (namely, it's URL) to it's "topic". Consequently, users do not have to visit the *exact* same page to be connected; they just need to visit pages on the same "topic". This approach appears more likely to find potential collaborators because it is less stringent and more practical.

From our survey responses it seems people agree that web activity is indeed a good method to capture a user's general interests. However, most everyone suggested that unless we do more than just topic modeling to filter and classify the data, the user will suffer from information overload.

6. FUTURE WORK

As suggested above, using more sophisticated data mining techniques which better structures and classifies web activity is an important next step for us. Another interesting future work direction would be to incorporate link prediction algorithms to identify additional meaningful connections. If two users u and v demonstrate similarity, it is likely the case that there are other relevant webpages or papers that u visited that v hasn't, or vice-versa. It's possible then that v would like to be notified of such content. Link prediction methods have been studied extensively in the networks literature [20], and could serve as a source of further connection.

It was also suggested that we make the list of domains captured customizable. Our current list 2 was constructed from our experience and from requests from others, but allowing users to add and remove domains could result in more comfortable users, and a more complete list of domains.

We have also begun preliminary work on two other features: ambient alerts and RSS feeds for personal webpage updates.

6.1 Ambient Alerts

In the physical world, someone's neighborhood is composed of the people he "runs into" during his daily routine. This can happen, for example, while going to class, eating lunch, etc. Similarly, we would like to use our tool to create a virtual neighborhood based on the same idea, except that "running into" someone means visiting a same or similar webpage.

To this end, we have created mock-ups for three types of ambient alerts which passively increase awareness of one's virtual neighborhood. Figure 7 (right side, top) shows the first of these alerts. Here, if a user u navigates to a page which has been previously visited by another user v, a small pop-up window appears notifying u that this page has previous visitors. By clicking on *More*, u is taken to v's profile page from which u can learn more about v's browsing his-

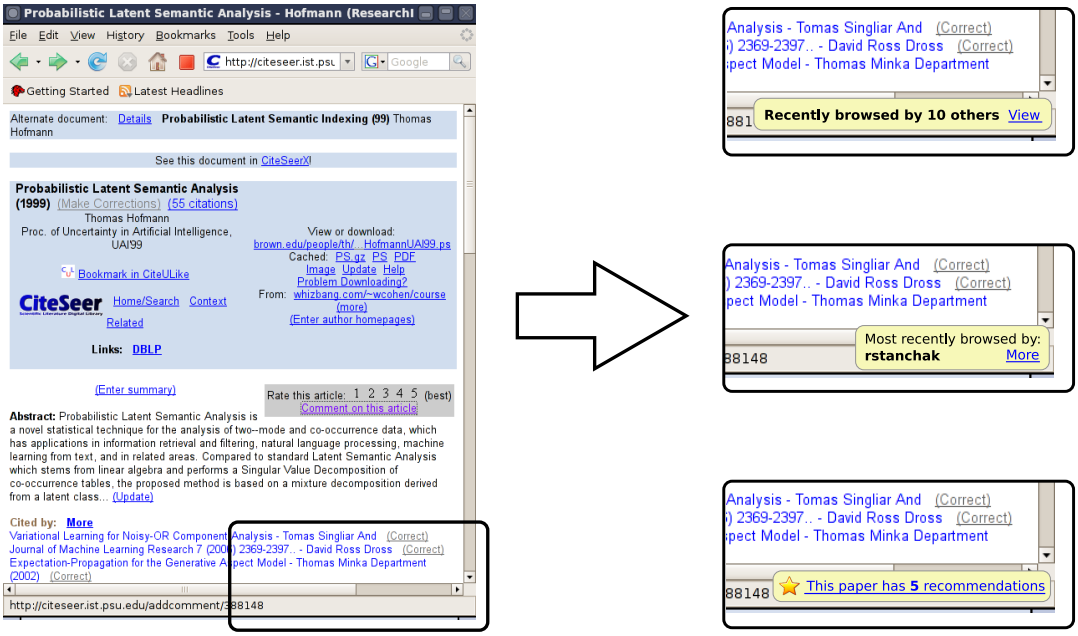

Figure 7: Future work: ambient alerts provide contextual information about a document with minimal distraction.

tory. Figure 7 (right side, middle) shows the second type of ambient alert, which presents u with a list of all other users which have recently browsed the particular site. These two types of ambient alerts are meant to passively (but immediately) notify u if a page visited represents a connection with another user. In this sense u has "bumped" into v, virtually.

Given this rich set of webpages users have viewed that our plugin has collected, it seems only natural then that we take advantage of user expertise to *recommend* webpages to other relevant users. The WWW has amassed more than 10 billion webpages, all of which cannot be browsed by a single person. However, if u and v have similar interests, and u "vouches" for a certain webpage or paper, it could be meaningful to v. Figure 7 (right side, bottom) presents a mock-up of what this feature might look like. Here, when u visits a webpage, a small window appears notifying u that other users have left comments about that page. When u clicks the link, she is taken to the document's profile page where comments are listed. A user can "vouch" for a docment by visiting the document's profile page and clicking on a "vote" button.

6.2 Personal web pages

Personal webpages usually contain a good overview of a researcher's interests, and nearly all graduate students and professors have them. However, updates can be sporadic, and no formal means of notification alerts interested parties of the changes. This problem of data dissemination in the blogosphere has been circumvented via Really Simple Syndication, or RSS. RSS is a web feed format (usually based on XML) that is used to publish frequently updated content [36]. A blog will host an RSS file on their domain, and update it whenever a new post is made. Those interested in the blog's content can then subscribe to the RSS feed via an RSS reader. RSS readers regularly grab the website's XML file and simply check to see if new content has been added since the last time checked. If so, the content is downloaded and delivered to the user.

Our situation is more complex, however, because most graduate students and professors do not maintain RSS feeds. Luckily, a new technology has emerged which allows anyone to create RSS feeds for websites that do not explicitly provide them [13]. We have begun to utilize this technology to keep track of webpage updates in real-time. The primary advantage of tracking webpages is being able to see when authors publish new papers. A new publication, almost by definition, addresses a problem at the forefront of a research area. Consequently, new publications bring forth new ideas and opportunities for discussion and collaboration. Figure 8

Figure 8: Future work: updates to personal webpages and common publication indexes are captured and aggregated.

presents a mock-up of this feature as a third component on the user's profile page. Other forms of presenting this data to the user still need to be investigated.

7. REFERENCES

[1] 2collab. http://www.2collab.com. Online; accessed 4/27/2008.

[2] I. Amazon. Amazon. http://www.amazon.com. Online; accessed 5/11/2008.

[3] E. Arias, H. Eden, G. Fischer, A. Gorman, and E. Scharff. Transcending the individual human mind and creating shared understanding through collaborative design. *ACM Trans. Comput.-Hum. Interact.*, 7(1):84–113, 2000.

[4] Attention recorder and approved services. http://attentiontrust.org/services. Online; accessed 4/27/2008.

[5] Bibsonomy. http://www.bibsonomy.org. Online; accessed 4/27/2008.

[6] Bibus. http://bibus-biblio.sourceforge.net. Online; accessed 4/27/2008.

[7] D. Blei, A. Ng, M. Jordan, and J. Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[8] Citeulike. http://www.citeulike.org. Online; accessed 4/27/2008.

[9] Connotea. http://www.connotea.org. Online; accessed 4/27/2008.

[10] del.icio.us social bookmarking. http://del.icio.us. Online; accessed 5/9/2008.

[11] Digg. http://www.digg.com. Online; accessed 5/9/2008.

[12] I. Facebook. Facebook. http://www.facebook.com. Online; accessed 5/11/2008.

[13] Feed43 (feed for free). http://www.feed43.com/. Online; accessed 4/27/2008.

[14] G. Fischer and Y. Ye. Personalizing Delivered Information in a Software Reuse Environment. *Bauer et al.(2001) UM*, pages 178–187, 2001.

[15] A. Hargadon. *How Breakthroughs Happen*. Harvard Business School Press, 2003.

[16] C. Information. Biblioscape. http://www.biblioscape.com. Online; accessed 4/27/2008.

[17] Jabref. http://jabref.sourceforge.net. Online; accessed 4/27/2008.

[18] R. T. Kouzes, J. D. Myers, and W. A. Wulf. Collaboratories: Doing science on the internet. *IEEE Computer*, 29(8):40–46, 1996.

[19] R. Kraut, J. Galegher, and C. Egido. Relationships and Tasks in Scientific Research Collaboration. *Human-Computer Interaction*, 3(1):31–58, 1987.

[20] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[21] I. Myspace. Myspace. http://www.myspace.com. Online; accessed 5/11/2008.

[22] K. Nakakoji, Y. Yamamoto, and M. Ohira. Computational support for collective creativity. *Knowledge-Based Systems*, 13(7-8):451–458, 2000.

[23] M. E. J. Newman. The structure of scientific collaboration networks. Working Papers 00-07-037, Santa Fe Institute, July 2000.

[24] Pybliographer. http://pybliographer.org. Online; accessed 4/27/2008.

[25] Reddit. http://www.reddit.com. Online; accessed 5/9/2008.

[26] I. RefWorks. Refworks. http://www.refworks.com. Online; accessed 4/27/2008.

[27] T. Scientific. Endnote. http://www.endnote.com. Online; accessed 4/27/2008.

[28] T. Scientific. Procite. http://www.procite.com. Online; accessed 4/27/2008.

[29] T. Scientific. Reference manager. http://www.refman.com. Online; accessed 4/27/2008.

[30] M. Steyvers and T. Griffiths. Matlab topic modeling toolbox 1.3.2. http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

[31] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 2007.

[32] Stumbleupon. Stumbleupon. http://www.stumbleupon.com. Online; accessed 5/11/2008.

[33] M. Sugimoto, K. Hosoi, and H. Hashizume. Caretta: a system for supporting face-to-face collaboration by integrating personal and shared spaces. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–48, New York, NY, USA, 2004. ACM.

[34] A. Vivacqua and H. Lieberman. Agents to assist in finding help. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 65–72, 2000.

[35] Wikipedia. Collaborative filtering — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Collaborative_filtering. Online; accessed 5/10/2008.

[36] Wikipedia. Rss file format — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/RSS_(file_format). Online; accessed 4/27/2008.

[37] Zotero. http://www.zotero.org. Online; accessed 4/27/2008.

Website name	URL
Algorithms for Molecular Biology	almob.org
PLoS Biology	plosbiology.org
Genome Biology	genomebiology.com
BioMed	biomedcentral.com
Genome Research	genome.org
Protein Science	proteinscience.org
Genetics	genetics.org
NIPS	nips.cc
Nature	nature.com
Science	sciencemag.org
Citeseer	citeseer.ist.psu.edu
CiteULike	citeulike.org
JSTOR	jstor.org
ACM	portal.acm.org
Google Scholar	scholar.google.com
Oxford Journals	oxfordjournals.org
DBLP	dblp.uni-trier.de
DBLP	informatik.uni-trier.de
Wikipedia	en.wikipedia.org
IEEE Transactions	ieee.org
IEEE Computer Society	computer.org
Springer	springer.com
ScienceDirect	sciencedirect.com
ArXiV	arxiv.org
Proc. of the National Academy of Sciences	pnas.org
Cryptology ePrint Archive	eprint.iacr.org
PS or PDF file on edu domains	$*.edu/*.(pdf\|ps)$

Table 2: List of all educations domains captured by UMDRecorder.

Field name	Example
user_id	saket@cs.umd.edu
url_scheme	http
url_host	www.cs.cmu.edu
url_path	/%7Ejure/pubs/blogs-sdm07.pdf
url_query	NULL
http_response_code	200
http_method	GET
page_title	Cascading Behavior in Large Blog Graphs
user_agent	Firefox/2.0
user_ip	128.8.128.181
timestamp	2008-03-24 12:57:05

Table 3: Example row in our database.