

# Implicit Search of Social Network Data to Discover Collaborative Search Opportunities

Emre Kıcıman  
Microsoft Research  
emrek@microsoft.com

## 1. INTRODUCTION

A prerequisite for collaborative information seeking is the identification of a group of people with a shared information need. In some cases, people may identify their shared information need based on a task they are collaborating on, such as colleagues working on a project together, or a family planning a vacation. In other cases, however, a shared desire for information may not be readily apparent. For example, two friends both seeking to buy a new camera might independently be searching for the same information without realizing that they could be collaborating.

In this paper, we hypothesize that the information people post to social networking sites such as Facebook or Twitter can be automatically analyzed to surface potential collaboration opportunities to a user. As people find interesting and useful information on the web, they often share it with their friends and colleagues using Facebook, Twitter, and other social networking sites. Other kinds of information that people post to such social networking sites include status updates, commentary and discussion, questions, and photographs. All of these messages are often an expression of interest in a particular topic or information item and some are an explicit expression of information need as well.

Connecting with friends and colleagues and keeping up with their status updates and other messages on social networking sites would seem to be a good way to infer whether a shared information need exists. Unfortunately, to keep track of all of these social messages is often overwhelming. Unless a person immediately reads their friend's message, it will disappear to be replaced by newer messages. Even if the user does read a message, it may not be relevant to their own current information needs and too easily dismissed.

To bridge this gap, our tool, the Social Web Experience (SWE) toolbar, is designed to surface relevant information from a user's social network while the user is performing an information seeking task, either casually or formally, on the web<sup>1</sup>. This tool performs a text analysis of both the contents of a user's social network data and the contents of their web browser. The goal of the analysis is to discover and surface to the user any social networking data that is discussing the same entities or topics that are on the displayed web page. For example, a user researching which new camera

to buy might discover friends' posts asking for advice about new cameras, saying they've recently broken their existing cameras, commenting about their newest camera purchases, or even friends who are domain experts blogging about the latest camera news.

We believe surfacing such information can help users performing solitary information seeking tasks by leading them to transition to a more fruitful collaborative activity. In addition, this tool can aid information finding by identifying shared information needs separated in time—i.e., a user may take advantage of the results of an information seeking task a friend has performed in the past. More broadly, surfacing relevant social information may aid throughout different phases of a social search task. At all phases of their model of the social search process [4], Evans and Chi found users performing social interactions. Before a search task begins, SWE may help identify area experts to aid in task formulation and guidelines. After a search task has been completed, SWE may help users identify their friends who would be interested in the discovered results.

### 1.1 Background

A primary function of social networking sites such as Facebook, LinkedIn and Twitter is to enable people to broadcast messages to their friends and communities. These messages allow people to distribute information to their friends, such as updates, opinions, messages and photos, as well as links to events, articles, blog posts and other resources. Such sharing occurs relatively frequently. According to Facebook's reported statistics, their 300 million users post over 2 billion pieces of content (web links, news stories, events, etc) every week. Sharing occurs frequently over other social networking services as well. For example, one survey reports that 11% of Twitter messages consist of information such as commentary or reporting of news and events [10]. Indeed, our own analysis of 1.8M messages (tweets) downloaded from Twitter, we find that 21% included a URL (a strong indicator of information sharing, given that tweets themselves are limited to 140-characters).

Today, these social networking sites display messages in time-order, this having the advantage of simplicity and highlighting of time-sensitive messages. However, this time-ordering also makes it difficult to view and find older messages. If a user of a social network does not check their stream of messages often enough, they will certainly miss messages, and furthermore will not be aware of the existence of a shared information interest or need. While services exist for searching information shared by your friends [7, 1], we believe that a user is unlikely to explicitly search for collaborative informa-

---

<sup>1</sup>Throughout the paper, we'll refer to a *user* as well as to *friends*. *User* refers to the person directly using our tool, while a *friend* is a contact of the user on one or more social networks. The friend may or may not be using our tool, but it is expected that they are sharing information with the user over a social network.

tion seeking opportunities without either an *a priori* reason to believe they will find results, a significant information need, or both.

## 1.2 Related Work

Considered broadly, our work on surfacing relevant social networking data to users as they browse the web is an instance of an implicit social search, as a user's information finding process is being augmented by relevant social networking resources. As the resulting information highlights not just information posted by friends, but also highlights the connection to a person in a user's social network, we believe this tool can also help users transition from into collaborative search tasks, where two or more people with a shared information need actively collaborate to find what they are looking for.

Implicit search has been studied in several contexts [13, 6, 3] such as triggering a search of personal data based on current tasks, or triggering web search based on email contents. In addition to facing similar text analysis and information retrieval challenges as this prior work, the Social Web Experience project is also considering how discovery and awareness of information shared by a friend may trigger communication or otherwise influence a collaboration among users.

Commercial services for explicitly searching one's own social networking data, including using one's own social network to influence a search of the public web, include Delver [1], Google's Social Search [7], and OneRiot [12]. Each of these services uses information about your social network to influence your search results, such as by surfacing public web data authored by your friends or ranking higher any web content preferred by friends.

Several tools have been built to surface social network data in context. The Glue [5] browser plugin displays social network annotations on books, music and movies across top sites on the web. Headup [8] is a browser plugin that simplifies the search for social network information by allowing users to highlight and click on web content to search for related social content. Closely related are web page annotation tools, such as Spar.tag.us. In [9], Hong and Chi discuss how Spar.tag.us uses paragraph fingerprinting to spread annotations on content across multiple copies of that content on the web.

## 2. SOCIAL WEB EXPERIENCE TOOLBAR

To discover semantic matches between social network data and visited web pages, SWE uses the analysis pipeline shown in Figure 1. First, the tool periodically retrieves a user's social networking data. This data may include text messages sent to the user, messages broadcast by friends, and profile information, such as friends' interests and location.

Our pluggable analysis pipeline allows multiple analysis techniques. We currently include two techniques. One is a naive n-gram extraction, while the second is a more sophisticated entity extraction algorithm. The entity extraction algorithm, described by Cucerzan in [2], is trained using the text of Wikipedia to canonicalize and disambiguate references to people, places, and other entities with articles written about them in Wikipedia. It uses the references between articles to discover alternative names for entities (for example, that the "University of Texas at Austin" is sometimes called "UT Austin", "Texas" , and uses the text of articles

to create term-frequency statistics to aid disambiguation of text. The output of this analysis includes both a canonical representation of an entity, and the specific surface form used in the analyzed text. For example, given the message: *Looking forward to a fun weekend in Portland, Oregon. Music, food and bikes. Will look to bring the Trail Blazers back with me to Seattle.*

one of the entities extracted by the entity extraction algorithm is:

```
<entity>
  <canonical>Portland Trail Blazers</canonical>
  <surface>Trail Blazers</surface>
</entity>
```

To improve the performance of the system, social network data is analyzed once and cached locally in a pre-processed form. As the user browses the web, the same text analysis techniques are applied to the web page content. Web pages are analyzed in the background of the user's browsing.

Once a web page has been analyzed, the matching pipeline is responsible for determining what social messages should be shown with that page. The matching pipeline consists of first producing preliminary one-to-many matches between each entity found on a web page and entities found in social network data. Each match is scored according to the features of the match (the source of the social network data, the entity's relevance to the web page, the length of the social message, etc). When there are multiple pieces of social network data that match the same entity on a web page, the matching pipeline may, depending on the kinds of data involved, either choose the highest scoring message to be displayed or merge the multiple messages into a single summary. As an example, if two friends both lived in Atlanta, the matching pipeline merges the these separate pieces of information into a single message.

We have built an implementation of this analysis pipeline and integrated it as a browser extension for Internet Explorer. Our current implementation downloads profile and stream posts from Facebook, and also downloads public Twitter messages on popular trending topics. The prototype is currently available for download at <http://research.microsoft.com/swe/>

In our prototype, we attempt a non-intrusive display of the results of this implicit social search. To do this, we highlight the relevant words in the web document and show the social data when the user hovers their mouse over the highlighted word. Figure 2 is a screenshot showing how a set of social data showing friends interested in the topic of 'travel' is displayed.

In future work, we are planning augmentations of the user interface to highlight opportunities to take action directly from the social message as it is embedded on the web page. Such actions may act as miniature web applications, embedded within the social message, or hyperlinks that open up new browser windows or other tools. For example, once friends' interest in travel has been highlighted, a user may directly message them to ask for advice on travel destinations, enquire if any are interest in a joint vacation, and even share the currently viewed content. Other actions that might aid in collaborative search might include providing users and their friends an easy way to link to shared document repositories where they may share results of an information finding task, or directly open collaborative search

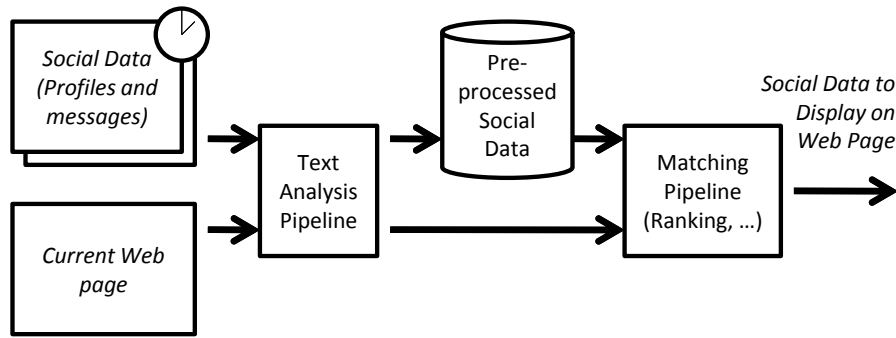


Figure 1: The analysis pipeline used by the Social Web Experience toolbar

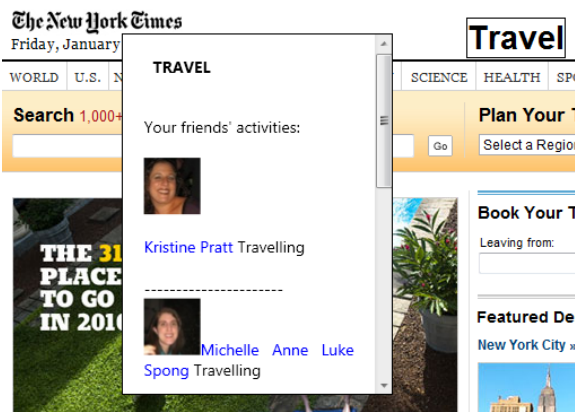


Figure 2: This screenshot shows how our tool displays a set of friends who are interested in the topic of travel. This information is shown over a New York Times travel web page.

tools such as SearchTogether[11].

### 3. CURRENT WORK

Our current work to improve our Social Web Experience prototype in several areas:

- **Improving analysis of short text messages:** Most research on entity extraction and other text mining techniques has focused on long-form text, including web pages, news stories, and e-mails. In these cases, the words in a document provide the context to correctly identify and disambiguate the key terms in the text. However, in the short messages common (and sometimes enforced) on popular social networking sites, existing techniques begin to fall short — often not recognizing the key terms in a message at all or incorrectly disambiguating the terms. We are currently investigating simple extensions of existing techniques, optimized through training on a large corpus of short messages, as well as the application of new techniques based on statistical language modeling.
- **Improving our models of communication, collaboration and relevance:** Critical to improving

our tool is improving our understanding of how and why people share or request information via social networks, and how an improved awareness of a social network data can trigger communication between users and aid collaborative tasks. Such understanding will help us improve the decision of when to show a message. Of particular interest is how our tool and the surfaced messages can help a user better understand shared interests and connections with people in their social network that might lead to increased communications and collaborations.

- **Developing On-line Metrics:** Last but certainly not least, we are investigating the kinds of usage metrics we might collect from our tool and how we may analyze such metrics to help test our experiment, without compromising user privacy.

Through participation in the Workshop on Collaborative Information Seeking, we hope to receive critical feedback on our project, especially on the topics of information sharing and serendipitous discovery in the context of information seeking.

### 4. ACKNOWLEDGMENTS

We would like to thank several people for their invaluable contributions to the social web experience toolbar: Chun-Kai Wang implemented much of the system; Silviu-Petru Cucerzan provided us with his entity extraction library; and Richard Hughes and Dan Liebling provided the framework and technical support for building an Internet Explorer extension. We also thank the workshop reviewers for their advice on improving the paper for this workshop.

### 5. REFERENCES

- [1] R. Carthy. Delver comes out of stealth with a new twist on social search. <http://www.techcrunch.com/2008/01/28/delver-comes-out-of-stealth-with-a-new-twist-on-social-2008>.
- [2] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, 2007.

- [3] S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz. Implicit queries (iq) for contextualized search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 594–594, New York, NY, USA, 2004. ACM.
- [4] B. M. Evans and E. H. Chi. Towards a Model of Understanding Social Search. In *In Proc. of Computer-Supported Cooperative Work (CSCW'08)*. ACM, 2008.
- [5] Glue. Glue: The network that sticks with you. <http://www.getglue.com/>, 2006.
- [6] J. Goodman and V. R. Carvalho. Implicit queries for email. In *CEAS*, 2005.
- [7] Google. Features: Google social search. <http://www.google.com/support/websearch/bin/answer.py?answer=165228>, 2009.
- [8] Headup. Headup: The semantic web firefox addon. <http://www.headup.com/>, 2009.
- [9] L. Hong and E. H. Chi. Annotate once, appear anywhere: collective foraging for snippets of interest using paragraph fingerprinting. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1791–1794, New York, NY, USA, 2009. ACM.
- [10] E. Mischoud. Twitter: Expressions of the whole self. Master's thesis, London School of Economics and Political Science, 2007.
- [11] M. R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In *UIST '07: Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 3–12, New York, NY, USA, 2007. ACM.
- [12] OneRiot. Oneriot: Realtime search engine. <http://www.oneriot.com/>, 2009.
- [13] J. Shen, W. Geyer, M. Muller, C. Dugan, B. Brownholtz, and D. R. Millen. Automatically finding and recommending resources to support knowledge workers' activities. In *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*, pages 207–216, New York, NY, USA, 2008. ACM.