

# Leveraging Social Context for Searching Social Media

Marc Smith  
Microsoft Research  
Mountain View, CA  
masmith@microsoft.com

Lise Getoor\*  
University of Maryland  
College Park, MD  
getoor@cs.umd.edu

Vladimir Barash\*  
Cornell University  
Ithaca, New York  
vdb5@cornell.edu

Hady W. Lauw  
Microsoft Research  
Mountain View, CA  
t-hady@microsoft.com

## ABSTRACT

The ability to utilize and benefit from today's explosion of social media sites depends on providing tools that allow users to productively participate. In order to participate, users must be able to find resources (both people and information) that they find valuable. Here, we argue that in order to do this effectively, we should make use of a user's "social context". A user's social context includes both their personal social context (their friends and the communities to which they belong) and their community social context (their role and identity in different communities).

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Search Process*

## General Terms

Design

## Keywords

social media, social search, social context

## 1. INTRODUCTION

The web is rapidly shifting from content contributed by nameless authors to a "social" web in which almost all content is linked to an author's name. This shift to social media requires a shift in the nature of the search tools needed to effectively extract value from these growing repositories. Users are shifting from just consuming information published by professional editors to contributing blog posts and twitter messages, updating their profiles on Facebook and

\*Work done while at Microsoft Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SSM'08, October 30, 2008, Napa Valley, California, USA.  
Copyright 2008 ACM 978-1-60558-258-0/08/10 ...\$5.00.

MySpace, asking and answering questions on Yahoo! Answers, authoring and editing articles in Wikipedia, tagging and rating pictures in Flickr and videos in YouTube, and voting for news items on Digg, etc. The amount of such user generated content created each day is staggering. Some estimates put the figure at 10GB/day for user-generated content on social media sites, and 3TB/day for private text [4]. While some of this content is spam, there is a significant amount of valuable high quality content embedded in these socially constructed repositories [1, 3]. However, the structure of these collections is distinct from the structure of the previous web. Clearly, search will need to adapt to play a crucial role in surfacing relevant and timely information from such repositories.

Social media is composed of multiple "social" data structures that are often conflated. Social networking sites like Facebook and MySpace have highlighted an emerging form of social context composed of a digital record of an individual's set of connections to other users of a given system. These forms of social network services have attracted attention to the idea of personal social networks, directed graphs connecting individuals to their friends, and how these networks have value for forging connections between users. Once generated, these personal social graphs can also serve as a novel form of reference set for collaborative filtering, replacing the association "people who like this also like that" with the association "people who like me also like this". We refer to this social context as the *personal social context*.

The personal social network is only one part of the ways social contexts are relevant for the improvement of search over collectively authored content repositories. In many social media systems, as users interact with others users and their content, a set of linkages is created and recorded in the form of transaction records. When linkage data is extracted from these repositories a social graph can be constructed that reveals patterns associated with the social roles and dynamics of that community. These patterns can be leveraged to identify key contributors and sources of value as well as the opposite. We refer to this social context as the *community social context*.

Each social context may initially be independent but later gain data based on their interactions. The system can model connections between users with similar interests or behaviors providing guidance about which content is deserving of their trust, and which authors have gained reputations for different patterns of contribution within the community.

Increasingly, new users come to social media spaces with existing social contexts, defined by records of relationships that exist in the social tools they already use; examples include email contacts, instant messenger buddies, social network service contacts, and linkages in wikis and blogs. This information can be used to construct a set of egocentric social networks in which the user is at the center of a set of relationships with others who may also have ties to one another. Over time, as these tools are used in concert, an integrated social context can be constructed. A user's personal social context can increasingly influence how he or she views both the relevance and the quality of content and interaction within social media. The Web is evolving into a "PeopleWeb", reflecting a shift from a web composed of pages to one populated by people and their artifacts and interactions [5].

Most existing systems fail to leverage either source of social structure effectively. In the following, we present a set of categories of social context and describe the ways each can be leveraged both independently and in concert to improve social search. Users may approach some content repositories without divulging any information about their personal social context while still leveraging the community social context that arises from patterns present in the content store. Insights into the structures present in a content repository can be combined with data about the social relationships a user maintains to further improve search results.

## 2. SOCIAL MEDIA

From a sociological perspective, social media can be described as "collective goods produced through computer-mediated collective action." For instance, in the case of Wikipedia, the collective goods are articles, and the collective action is the coediting process of article writing. In the case of Digg, the collective goods are news stories, and the collective action is the effort of finding, voting for, and commenting on stories that pushes the most important stories (as determined by Digg users) to the most prominent position on the website (i.e. the front page). In the case of Facebook, the collective goods are social capital, measured in the number and kinds of people active in the social network, and the collective action is the process of developing individual profiles and of the links between them.

The term social media includes such a diverse a collection of tools and services that systems that should be further distinguished are often confused. Social media can vary along several dimensions. A key dimension is the size of the social groups that are producing and consuming the social media. Figure 1 shows the how the different sized groups of producers and consumers generate different kinds of social media, from emails generated by dyadic communication to forum posts produced by individuals but consumed (read) by crowds, to collective search engine optimization both produced and consumed by large groups of people (searchers).

Additional dimensions focus on the ways different social media expose different granularity of content creation and control. Media granularity is an attribute of two dimensions of social media: the size of the unit of exchange and the interactivity (synchronous or asynchronous) of that exchange. Figure 2 suggests three levels of granularity for social media: fine (wherein users have control over minimal units of media: bits, words, or pixels), medium (wherein users have indirect control over units of media via the manipulation of

abstract objects, like avatars in virtual worlds), and coarse (wherein users have limited control over particular blocks of content, like photos or documents, and cannot alter media outside these blocks). These three levels cover many forms of social media from collaborative documents to virtual worlds to blogs. At any level of granularity, media can be synchronous or asynchronous, which allows us to distinguish between exchanges like IMs (coarse, but synchronous) and e-mail (coarse, but asynchronous).

## 3. ROLES AND IDENTITY IN SOCIAL MEDIA

The different scales of producer and consumer groups for social media not only specify different types of media produced, they also induce differences in the composition and structure of communities that spring up around these artifacts. A small discussion board or email list is usually composed of individuals with strong mutual ties (friends, relatives, coworkers). Almost all participants on the board are likely to know everybody else, and this dense web of interconnections creates a strong social pressure for all members to act in a productive way. Spammers and trolls (individuals who post provocative messages on purpose) are rare in small communities, precisely because the people they would be spamming and trolling are their friends, and the bonds of mutual trust involved act as strong barriers to disruptive behavior. In social media like Wikipedia and Digg, however, the sheer number of producers and consumers makes it hard to form tightly knit groups. Anonymity is the norm in these large spaces, and the social pressure on community members to make useful contributions is relatively weak. In order to survive, therefore, social media that attract a lot of users need distinct mechanisms to promote constructive behavior and prevent spam and other forms of community vandalism.

One such mechanism is the appearance of distinct roles in social media. Participants often take up a set of behaviors and activities that have a stable self-similarity but are distinct from other roles or patterns present in the social space. For example, dedicated "answer people" often emerge in some forms of social media spaces. These are participants who reply to user questions and concerns in large numbers, providing a reference librarian like service to potentially hundreds or thousands of people seeking help. Other participants may take a contrasting role as a "discussion person," who engages in exchanges on often contentious topics resulting in the production of many messages and connections among a group of similar "discussion people". In discussion-oriented social media communities participants enacting this role generate much of the content through prolonged interactions with one another. Other roles exist, some of which appear only within specific social and technical contexts. For example, the features present in Wikipedia enables the emergence of specialized patterns of behavior that rely on features unique to wikis. Studies of Wikipedia has demonstrated the essential role "vandal hunters" play in the development of successful Wikipedia pages. These participants are people who dedicate themselves to reversing vandalism (the adding of inappropriate content to, or deletion of, Wikipedia articles) by restoring backed up versions of article pages. Often this is the only or dominant behavior in which the participant will engage. Similarly, in open source software reposi-

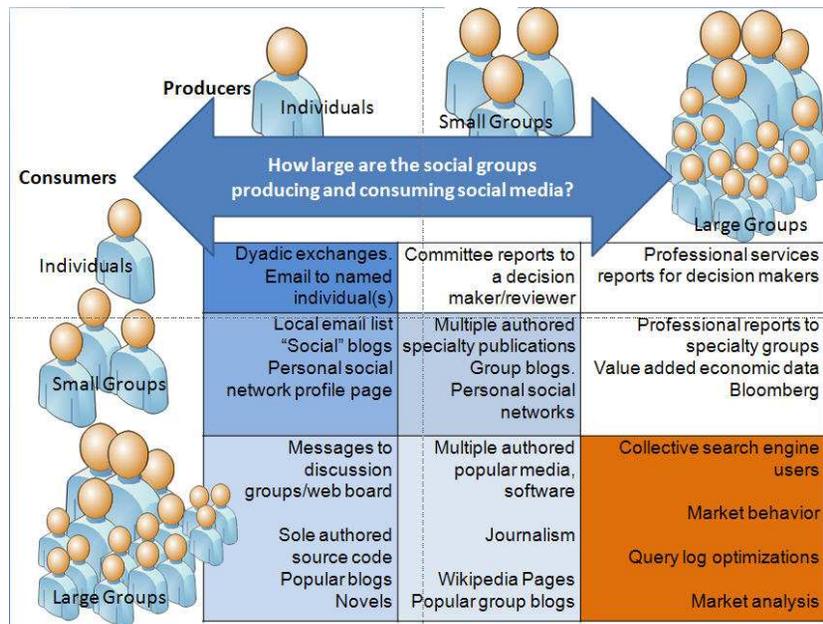


Figure 1: Variations in social media in terms of the size of the producer and consumer populations.

tories (e.g. Sourceforge), research shows that some users become dedicated bug-fixers who contribute little novel code, but fix mistakes in existing software, in contrast with other contributors who specialize in new code contribution and yet others who only engage in discussion about the software project.

The interaction of these roles in social media creates healthy communities: when enough roles exist to fill the community's needs, and enough individuals fill each role, in concert they can both create valuable content and censure disruptive behavior. In the absence of pre-existing institutions of social control, these social roles allow large-scale communities to create collective goods in an organic way, without relying on an explicit organizational structure for task delegation.

Roles also help users to forge and manage their online identities in social media spaces. Particularly active answer people, discussion people, and other role representatives become well-known in their respective communities. This "e-fame" endows the individual in question with some measure of respect from fellow community members, who are more likely to read, and respond to, the posts of an established answer person than to those of a newcomer. On the other hand, the community expects its most active members to conform to the roles they play, so a vandal fighter in Wikipedia has to be careful about not (accidentally or on purpose) vandalizing pages himself, lest he lose the community's trust. Online identity management is a subtle, but very important process, in which individual behavior affects and is in turn affected by the role that individual plays in a particular social media community.

#### 4. SEARCH IN SOCIAL MEDIA

One of the things that distinguishes search in social media from more traditional types of search is the types of things that people look for. Unlike standard search, where queries are often informational, navigational or consumer-related,

when searching social media, the query is often for people (e.g., people I know, people like me, experts in certain topics), groups or communities, or content created by specific people or groups. These socially oriented queries can be supported by a novel approach to search that considers the social structure and roles of content contributors as context for the ranking and relevant process. Two forms of social context can be leveraged. Each user brings their personal social context in the form of their friendship network, the communities of which they are members, their identity and roles in each. A second form of social context is found in the details of interactions among contributors to a social media repository. This form can be useful for social ranking and relevance even when the user issuing a search query lacks or does not provide data about their own social context.

Search is improved by making use of social context to improve retrieval in the following ways:

- **Filter.** Social context may be an additional selection or filtering criterion that is not present in the search query alone. For instance, queries such as "good music to play at my party" or "good papers to discuss with my research group" are unintelligible to a search engine that does not know either who the user is or who the user's friends are. Knowing the user's social context give a whole different meaning to the same queries, which in this case would involve selecting those content favored by the user and friends, and excluding those that the user and friends dislike.
- **Rank.** A user's social context influences how the user intuitively perceives which content items are most relevant to him/her. We see this in play in sites such as Amazon.com, where knowing what others are buying helps to inform one's own purchasing decisions. In addition, as most content in the social media are generated by others, a user may have developed a sense

	<b>Digital Object Editing Granularity</b>	<b>Fine</b> (Character/Pixel/Byte)	<b>Medium</b> (Object/Attribute/Track/Player)	<b>Coarse</b> (Document/Message/Blog Post/Photo)
		Each user can directly control smallest units of content.	Each user controls medium sized blocks of content that can only <b>indirectly</b> alter or be altered by other user's content in a larger shared data structure.	Each user controls a block of content, rarely edited or modified by others with only associative linkages.
<b>Digital Object Editing Synchronicity</b>				
	<b>Synchronous</b>	Real time Shared canvas	Virtual Worlds Multiplayer Games Real-time networked musical jamming	Chat, IM, Twitter
	<b>Asynchronous</b>	Shared docs, images, video, audio Source code Wikipedia	Contribution to collected works (album, anthology, report section, discussion group, photosets and other collections).	Email Blog posts Link sharing Photo sharing Document sharing Turn based games

Figure 2: Social media dimensions according to varying granularity and rate of exchange.

of who are trustworthy or reputable within the community. It is conceivable that a frequent participant in a discussion forum would instinctively know how to weigh the credibility of answers given by different contributors.

- **Disambiguate.** Most search queries are just one or two words long, much too short for a search engine to fully understand the user's intent. The query "IR" may be issued by an information retrieval researcher or by someone checking the current stock price of Ingersoll-Rand Company Limited. The query "jaguar" may refer to the car, the animal, or the operating system. As we tend to associate most with others like ourselves [2], our social context would be a huge hint towards disambiguating the user's intent in the face of such ambiguous queries as the above.
- **Share.** The act of sharing one's generated content (e.g., tags, ratings, edits) with others within a social medium is an important foundation to forming one's social context and using it for searching. Knowing what others share helps us to discover people with similar interests, and aids our own discovery of new and interesting content (as highlighted or generated by others).
- **Recommend.** As sites such as MovieLens, Amazon and ePinions illustrate, there is a huge market for online recommendation systems. As researchers in collaborative filtering noted from early on, user similarity is an important component in making recommendations. Making even greater use of personal social context is likely to be beneficial.
- **Match Make.** Social content search is often driven by a desire to find people similar to the user or different in complementary ways. The need for match making goes beyond the dating and romantic relationship services to extend into professional relationships and the need to build teams with diverse skills and compatibilities. Social search can deliver lists of people like me or people who might be compatible with me for business purposes.

Search has always occurred in a social context but increasingly that context is in a machine readable form that makes it available for algorithmic ranking and relevance applications. These applications can align search results with social context which improves search by considering the social structures of both the content creators and the query author.

## 5. FUTURE DIRECTIONS

There are a number of possible directions for intermediate and long-term development of search in social media. One such direction is better capabilities for managing roles and identity. Another direction is better support for sharing social context and collaboration. Finally, we see the possibility of transitioning from searching social media to searching as a social activity, enabling users to construct and collaborate in new social contexts around their existing search tasks.

## 6. REFERENCES

- [1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 183–194, New York, 2008. ACM.
- [2] K. Carley. A theory of group stability. *American Sociological Review*, 56(3):331–354, 1991.
- [3] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictor of answer quality in online q&a sites. In *Proceedings of the 26th ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 865–874, New York, 2008. ACM.
- [4] R. Ramakrishnan. Web data management: Powering the new web. In *Proceedings of the NSF Symposium on Next Generation Data Mining*, 2007.
- [5] R. Ramakrishnan and A. Tomkin. Towards a People Web. *Computer*, 40(8):63–72, 2007.