Social Media and Search

Junghoo Cho
University of California, Los Angeles

Andrew Tomkins
Yahoo! Research

The past few years have witnessed the rapid rise of *social media* Web sites such as Flickr, del.icio.us, YouTube, Myspace, and Facebook, as well as the proliferation of "mashup" applications created when users combine services from multiple sources.. These sites contain *user-generated content* in various forms, from plain text to rich multimedia. In fact, most publicly available text content created during the next 24 hours will be generated by end users, rather than professional writers, journalists, corporate communications departments, or others whose job it is to create and publish content. Furthermore, end users will generate an additional two orders of magnitude more text that they will send privately to other users through a communications channel such as email.[1] The emergence of user content as the dominant content form on the Web raises various questions about the most effective approach to processing it.

Much user-generated content is hosted on *social media* Web sites, which commonly allow users to form communities based on shared interests, and to associate tags, reviews, recommendations, and comments with that content. These *metadata* are invaluable in helping assess the highly variable quality of content end users are creating. Visitors to these sites often seek not just the content itself, but also an understanding of the individuals who posted it. Furthermore, they might visit the site without any particular goal or informational need, but rather based on the simple desire to "get an update" or "be entertained" during their spare time.

Academic Landscape

Social media innovation occurs largely in the corporate sector, with many offerings arising from small Internet startups. Academic work in this area has focused primarily on studying the dynamics of social media generation or consumption. Significant literature exists on the dynamics of personal publishing through blogs and of distributed metadata generation through tagging. Academic work also exists on bulletin boards, wikis, and other creation modalities as well as on comments, reviews, ratings, bookmarks, and other forms of metadata. Workshops in various disciplines have sprung up around this

1089-7801/07/$25.00 © 2007 IEEE

area, and beginning in 2007, a new international conference on weblogs and social media (ICWSM) is being held annually.

Work in information retrieval has only recently begun to address social media corpora and to incorporate social media metadata as features.[2] General-purpose Web search engines index and return social media content in response to queries, and specialized search engines perform even more targeted analysis of particular social media — technorati.com or blogpulse.com for blogs, boardreader.com for bulletin boards, and so forth. However, these companies typically don't publish their techniques in order to maintain a competitive advantage.

At the fringes of social search are implicit techniques that capture users consumption behavior to modify retrieval for new users. Substantial literature discusses collaborative filtering in this space, and emerging work considers user click behavior as a feature in Web search.

Challenges and Opportunities

The proliferation of user-generated content and the resulting associated metadata on the Web introduce new challenges and opportunities in search. For example, the rich metadata users provide help distinguish the high-quality content from the vast amount of noise, but might also be susceptible to user manipulation. In particular, the following characteristics make searching such user-generated content more challenging:

- *Vulnerability to spam.* By letting users create and publish content without much central governance, social media Web sites have been able to amass a rich body of content. Unfortunately, user-generated content is intrinsically more vulnerable to spam and noise because the content isn't filtered by any meaningful editorial process. When no dependable third party verifies the integrity of published content or the author's motivation, a significant portion of such content inevitably exists to promote its own commercial interest, potentially without benefiting public users. Partly due to the significantly larger fraction of spam, metrics such as PageRank, which work well for "traditional" Web content, are less effective for social media content.
- *Short lifespan.* The content on social media Web sites tends to have a shorter lifespan because much of it focuses on an ongoing real-world event or a current "hot" topic. Public interest in such content subsides rapidly over time. Thus most user-generated content doesn't accumulate many incoming links or user visits before it becomes irrelevant, making it difficult to judge such contents' general "quality."
- *Locality of interest.* The large pool of potential content creators on social media sites has produced an explosion of publicly shared content, but much of it is of little interest to the general public. When publication costs are high, Web sites publish only content that's interesting to a general audience. However, in a world of near-zero publication costs, a teenaged boy's daily journal is unlikely to spark the interest of the general public, even though it might be interesting to his friends and family.
- *Access control.* Most user-generated content is "private," meaning it's sent to only a few recipients and isn't visible to anybody else. Recently, a significant middle ground is emerging — for instance, Facebook provides differentiated access to all members of a network, and these networks frequently contain tens of thousands of people. Content visible to the network isn't distributed to all members; rather, it's hosted, and Facebook verifies access credentials at access time. Searching in such an environment provides significant new challenges that existing data structures don't effectively address.

Despite these challenges, the richer context that social media content provides gives us exciting opportunities. For example, users often form explicit and implicit communities around their interests, letting us apply collaborative filtering techniques at an unprecedented scale. Users also provide a rich body of metadata, in the form of tags, bookmarks, and favorites, and they leave detailed interaction history while they explore the content.

In this Issue

The three articles selected for this special issue present some early work in understanding the characteristics of user-generated content and its metadata, and in making high-quality content more accessible and comprehensible.

In the first article, "Social Information Processing in Social News Aggregation," Kristina Lerman studies the mechanisms by which a broad user community produces a set of community recommendations for new articles. She investigates several factors influencing an article's overall popularity on news aggregation site Digg, including how well the article's author is "connected" within the online community, and tries to quantify these factors' impact by fitting the data against a mathematical model.

Next, in "Social Bookmarking for Scholarly Digital Libraries," Umer Farooq and his colleagues study the process by which users save references to objects (in this case, technical papers) for later discovery or for social discovery. In particular, they study how tags are used on a social bibliography site, CiteULike, and suggest how such a site might be improved.

In the last article, "Fighting Spam on Social Websites: A Survey of Potential Approaches and Future Challenges," Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina study an increasingly prevalent problem in search and information discovery: malicious manipulation of content or metadata to influence results. They survey three common countermeasures against such spam — detection, demotion, and prevention — from the standpoint of social media Web sites and discuss some differences and challenges of fighting spam in this context.

The articles in this special issue represent just a sample of the early findings in this fledgling research area. As users become more familiar with social media, and as service providers gain a better understanding of their users, user behavior and content and metadata characteristics are likely to change, necessitating the continuous reevaluation of what we learned before

Social media analysis is well positioned to continue advancing its understanding of content dynamics and metadata generation. At the same time, social media is becoming big business and is driving a significant fraction of worldwide pageviews on the Web. It's imperative to develop better and more effective technologies to cope with ongoing attempts by commercial users to manipulate the system to their advantage. Likewise, as competition continues to increase in these domains, social search will become a differentiating technology, resulting in continued investment and material advances beyond the state of the art today. At the same time, the high volume of social media consumption will result in another critical problem: the monetization of social media sites. Here, the problem is one of searching for relevant advertisements based on user properties as well on content properties. We expect these two problems to receive increasing attention over the next few years.

Acknowledgments

IC's prior and current editors in chief, Robert E. Filman and Fred Douglis, helped us greatly with this issue. We thank them for giving us the opportunity to create a special issue for this new and emerging research area and for their constant commitment and encouragement. We also thank the numerous academics and practitioners who submitted their excellent research results to this special issue. Finally, numerous reviewers spent an enormous amount of time evaluating the submitted articles and providing thoughtful comments and guidance to the accepted articles. Without their help, it wouldn't have been possible to reach the level of quality the articles in this issue represent.

References

1. R. Ramakrishnan and A. Tomkins, "Toward a People Web," *Computer*, vol. 40, no. 8, 2007, pp. 63–72.
2. J. Jeon, W.B. Croft, and J. Lee, "Finding Similar Questions in Large Question and Answer Archives," *Proc. 14th ACM Conf. Information and Knowledge Management* (CIKM 05), ACM Press, 2005, pp. 84–90.

Junghoo Cho is an assistant professor in the Department of Computer Science at the University of California, Los Angeles. His main research interests are evolution, management, retrieval, and mining of the Web. Cho has a BS in physics from Seoul National University and a PhD in computer science from Stanford University. Contact him at cho@cs.ucla.edu.

Andrew Tomkins is director of search research at Yahoo! Research. His research interests lie in measurement, modeling, algorithms, and analytics for large heterogeneous data sets such as the Web. Tomkins has a PhD in computer science from Carnegie Mellon University. Contact him at atomkins@yahoo-inc.com.