Search in Conversational Social Media Collections

Jonathan L. Elsas
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jelsas@cs.cmu.edu

1. INTRODUCTION

Community generated content has become increasingly important over the past several years: blogs, Wikipedia, online forums, twitter, Yahoo! Answers, Facebook and many other online communities that foster social interaction have flourished. However, studying "Search in Social Media" as a distinct sub-field of information retrieval poses some questions. Although there is a loose consensus of the definition of *social media*, it is not immediately clear what is different or special challenges social media collections pose as compared to other text collections that have been previously studied. This abstract focuses on a specific class of social media – archives of conversational social media – and highlights facets of these collections that distinguish them from others collections that have traditionally been studied in the information retrieval research community.

2. CURRENT WORK ON SEARCH IN SOCIAL MEDIA

The information retrieval research community has shown increased interest in looking at social media collections for various information retrieval tasks: the Blog Search track at TREC focuses on blog collections [7, 6], Wikipedia is becoming an ubiquitous external collection in many types of retrieval techniques, and several search related tasks have been evaluated in the context of community question-answering [1], and online discussion forums [4, 8].

Although all these research efforts are applied to social media collections, there are not common tasks or techniques that are applicable across social media collections. From the IR researcher's perspective, what is significant about studying these artifacts of social media? Is there something that distinguishes these from other document collections? If so, how can we leverage that distinction in our retrieval models?

3. CONVERSATIONAL SOCIAL MEDIA COLLECTIONS

Many social media outlets generate potentially large volumes of archived data – some in the form of more or less polished documents, like a blog post or Wikipedia article; others, like twitter, are snippets of an often one-sided conversation and broadcast messages. Is there value in providing access to these artifacts of social media? Some, like twitter, seem to be mostly ephemeral, only (generally) interesting in the moment and quickly fading from view. Even the twitter search engine advertises: "See what's happening – right now" and the results are seemingly only ranked chronologically[1]. Services like Twitter, although potentially useful for event detection and tracking, may not lend themselves to traditional ad-hoc information retrieval tasks.

Many other types of social media – some existing long before Web 2.0 was born – generate large text archives that can be of great long-term value. There exists an online forum, public mailing lists, newsgroup or message board for virtually every special interest group under the sun – from gardening, to home-brewing, to specific electronics manufacturers. These are often heavily trafficked, populated with real subject matter experts, and host a rich information exchange. The content created through these social media outlets present an enormous value to information seekers, and information retrieval research has a lot to contribute in this corner of social media.

In these types of collections, referred to here as *Conversational Social Media*, the basic element are *threads* of discussion among different authors. The online forum or message board is a typical such collection. These collections often have rich organizational schemes, with threads grouped in topical sub-forums, which are then organized hierarchically. Each message typically has associated metadata indicating the time of creation and author.

4. RETRIEVAL CHALLENGES

In conversational social media collections, there are several retrieval challenges that require special attention.

1. **Retrieval Granularity**: In many of these collections, the unit of retrieval – what we consider a document – is not fixed, but rather dependent on the task or information need. Consider online forums, often organized into topical sub-forums, which in turn are organized into conversation threads of individual posts. Some information needs many only require a single post as a result, some require the context of the full conversation thread, and others may need to retrieve a pertinent sub-forum.

2. **Multiple Axes of Organization**: In addition to the typically topical hierarchy present in these collections, there are frequently several other orthogonal axes of organization. Individual messages are typically associated with their author and the time when the message

[1] The results are clearly also filtered for language, and possibly for spam, location, etc. But the core "ranking" algorithm appears to be only date-driven.

was written. These multiple axes naturally lend themselves to faceted search interfaces, where the user is in control of which axes are favored at results presentation time.

Multiple axes also define different searchable elements in the collection. In highly trafficked message boards and mailing lists, tens or hundreds of thousands of users with varying levels of expertise contribute to the conversation. With their explicit authorship metadata and large numbers of authors, these collections naturally lend themselves to expert-finding tasks similar to those studied at the TREC enterprise track [3].

3. **Author Modeling**: Retrieving authors is not always an appropriate task, but authorship information my be useful beyond just expert finding. The presence of authorship metadata and a large number of authors in the collection enables the aggregation of statistics about authors. For a given author, we may be able to identify which topics they have expertise in, whether they tend to ask or answer questions, and how frequently they post messages. By modeling authors in this way, we can favor threads with contributions from experts, potentially finding more reliable and useful retrieval results.

4. **Indexing Challenges**: To support this variety of retrieval tasks, there are significant indexing challenges that must be addressed. Current retrieval systems have a static view of documents, because the retrieval task is typically fixed. But, in these collections we may want to support retrieving along several different axes of organization. What is the most effective way to index the text to simultaneously support retrieval at the thread, forum and author levels?

These challenges, of course, are not entirely unique to social media search, and have to some degree been addressed in previous research. This question of identifying the granularity of the unit of retrieval has been addressed at the document level (for example in XML element retrieval at INEX)[5], but not so much at the collection level. Resource ranking in federated search and cluster-based retrieval bear some resemblance to the selection of a topical sub-collection, such as a sub-forum ranking. But, this previous work generally only takes advantage of one level of organization, not a deep hierarchical organization often present in online message boards. Author-ranking has also been studied at TREC in the Blog and Enterprise Tracks [2, 7].

In all of these examples, a single level of document organization has been studied in isolation, without much regard to the interaction between the different aspects of the collection. Conversational social media collections, however, offer these many levels of rich organization in a single collection.

5. CONCLUSION

Conversational social media collections provide multiple levels of organizational granularity, different axes of organization, and multiple types of searchable objects. Search in conversational social media collections will potentially be an interesting and fertile direction of information retrieval research – pushing the systems to support more sophisticated multi-dimensional indexing and extending existing retrieval models to handle rich relationships between documents.

6. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, New York, NY, USA, 2008. ACM.

[2] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2007 enterprise track. *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2007)*, 2007.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM New York, NY, USA, 2006.

[4] J. L. Elsas and J. G. Carbonell. It pays to be picky: an evaluation of thread retrieval in online forums. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 714–715, New York, NY, USA, 2009. ACM.

[5] M. Lalmas and A. Tombros. Evaluating xml retrieval effectiveness at inex. *SIGIR Forum*, 41(1):40–57, 2007.

[6] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 blog track. In *Proceedings of the 2007 Text Retrieval Conference*, 2007.

[7] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 blog track. In *Proceedings of the 2008 Text Retrieval Conference*, 2008.

[8] J. Seo, W. B. Croft, and D. A. Smith. Online community search using thread structure. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1907–1910, New York, NY, USA, 2009. ACM.