

Finding People and their Utterances in Social Media

Wouter Weerkamp
w.weerkamp@uva.nl

ISLA, University of Amsterdam, Science Park 107, 1098 XG Amsterdam

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Social media, information retrieval

ABSTRACT

Since its introduction, social media, “a group of internet-based applications that (...) allow the creation and exchange of user generated content” [1], has attracted more and more users. Over the years, many platforms have arisen that allow users to publish information, communicate with others, connect to like-minded, and share anything a users wants to share. Text-centric examples are mailing lists, forums, blogs, community question answering, collaborative knowledge sources, social networks, and microblogs, with new platforms starting all the time. Given the volume of information available in social media, ways of accessing this information intelligently are needed; this is the scope of my research.

Why should we care about information in social media? Here are three examples that motivate my interest. (A) Viewpoint research; someone wants to take note of the viewpoints on a particular issue. (B) Answers to problems; many problems have been encountered before, and people have shared solutions. (C) Product development; gaining insight into how people use a product and what features they wish for, eases the development of new products. Looking at these examples of information need in social media, we observe that they revolve not just around *relevance* in the traditional sense (i.e., objects relevant to a given topic), but also around criteria like credibility, authority, viewpoints, expertise, and experiences. However, these additional aspects are typically conditioned on the topical relevance of information objects.

In social media, “information objects” come in several types but many are *utterances* created by *people* (blog posts, emails, questions, answers, tweets). People and their utterances offer two natural entry points to information contained in social media: *utterances* that are *relevant* and *people* that are *of interest*. I focus on three tasks in which the interaction between the two is key.

The first task concerns **finding utterances** in social media. Although this resembles a traditional ad hoc retrieval task, the lack of top-down rules and editors in social media entails the use of unexpected language: spelling and grammar errors are not corrected,

and the vocabulary is unrestricted, unlike edited content. In my research, I use two features of social media to overcome the problem of unexpected language: (i) I translate several *credibility* indicators to measurable features, and implement these in the blog post retrieval process to enhance retrieval effectiveness [2]; (ii) Utterances are surrounded by their environment, and this influences their content. I use this observation to introduce a model for query modeling using external collections [5], and investigate which context levels are useful in email retrieval [4].

The second task, **finding people** in social media, is operationalized in *blog feed search*: identify blogs that show a recurring interest in a given topic. This task shows similarities with expert finding, and models from this field have been successfully adopted. Blog-based models build and rank representations of bloggers based on their utterances, whereas post-based models rank utterances and aggregate scores to construct a ranking of bloggers. The former capture the centrality of the topic to the blog, but are not very efficient; the latter can identify interesting posts and are efficient. In [6, 7] I introduce a two-stage model that ranks utterances and constructs models for the blogs these utterances belong to and ranks these blog models. My two-stage model is more efficient than blog-based models, and more effective than post-based models.

The third and final task builds on the previous two, and focuses on **finding utterances using people**. Applied to search in email archives, personal profiles can be constructed from people’s utterances. On top of these personal profiles, communication profiles are built, containing information extracted from threads, quotes and replies as well as linguistic clues. Communication profiles can indicate the role of people in a topic-dependent way. For a given topic, I use both communication and personal profiles of people to find utterances that are relevant to the topic [3].

References

- [1] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.
- [2] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *ACL-08: HLT*, pages 923–931, 2008.
- [3] W. Weerkamp and M. de Rijke. Communication and personal profiles in email search. In *To be submitted*, 2010.
- [4] W. Weerkamp, K. Balog, and M. de Rijke. Using contextual information to improve search in email archives. In *ECIR 2009*, 2009.
- [5] W. Weerkamp, K. Balog, and M. de Rijke. A generative blog post retrieval model that uses query expansion based on external collections. In *ACL-ICNLP 2009*, 2009.
- [6] W. Weerkamp, K. Balog, and M. de Rijke. Blog feed search using a post index. *Submitted*, 2010.
- [7] W. Weerkamp, K. Balog, and M. de Rijke. A two-stage model for blog feed search. In *SIGIR 2010*, 2010.