Finding People and their Utterances in Social Media

Wouter Weerkamp
University of Amsterdam, The Netherlands
w.weerkamp@uva.nl

Abstract

At the beginning of the twenty-first century, the web entered a phase of explosive growth. Since then, a multitude of platforms has become available for users to publish information, communicate with others, connect to like-minded individuals, and share anything that they want to share. These platforms are commonly known as social media. Social media enables many-to-many communication: many people can create content, which in turn can be read by many others. To make interesting content findable for many people, machines need to be able to identify the "right" pieces of content or the "appropriate" content creators. That is, we need ways to intelligently access information in social media.

The main motivation for the research in my thesis is that we want to enable intelligent access to, and analysis of, the information contained in social media. To this end, we must determine the topical relevance of social media documents, while countering the specific challenges posed by their noisy and creative nature. We identify two entry points for accessing information in social media: (i) the people creating the social media and (ii) their individual utterances (e.g., blog posts, tweets). These entry points act as a doorway to the information in social media. The aim of the thesis is to improve searching for people and their utterances in social media, thereby offering intelligent access to the information it contains.

In the thesis we explore ways to improve retrieval effectiveness, for both people and their utterances, in five research chapters. We start with an extensive analysis of the query logs of a people search engine and show that users are mainly interested in finding social media profiles of people for whom they search. We continue with our focus on searching for people in the second chapter, but adopt a more technical perspective. We use blog posts as representation of bloggers' interests and show that for blogger finding a combination of two models, one based on individual blog posts and one on aggregated blogger models, is both effective and efficient.

In the remaining three chapters we move away from finding people and focus on individual utterances. Specifically, we look at blog post retrieval in two chapters and email retrieval in mailing lists in the last chapter. For blog post retrieval we explore the use of indicators obtained from a credibility framework. We propose ways of estimating these indicators and rerank an initial list of blog posts based on these (combined) indicator scores. We find that incorporating these indicators leads to improvements and our analysis determines indicators that are most influential.

The second chapter on blog post retrieval and the chapter on email search both use a notion of "context" to help improve retrieval effectiveness. For blog post retrieval we observe that much of what is written in blog posts is based on things happening the bloggers' environments. To use this observation we propose a generative retrieval model that uses information from a set of external sources. We test our model using a news collection, Wikipedia, and a web collection as external sources. We find that our model is capable of improving retrieval performance on most topics and that the per-topic collection importance is a very useful component of the model. Finally, the last research chapter uses the direct context of emails to refine queries; we explore the use of threads, the entire mailing list, and the pages from within the same community as sources for query expansion terms. Besides that, we also incorporate the previously mentioned credibility-inspired indicators in this setting. We find that context helps in improving retrieval performance, as do the credibility-inspired indicators.

Concluding, my thesis provides new insights into search behavior in social media and shows that a diverse set of methods can help improve retrieval effectiveness for multiple information access tasks within social media.

Available online at `http://wouter.weerkamp.com/`