Finding people and their utterances in social media

Weerkamp, W.

Citation for published version (APA):
We started this thesis by introducing social media and giving examples of the various platforms that are currently available. Information contained in these social media platforms is interesting for numerous reasons, some of which were listed in Section 1.1 (page 2), and research into accessing this type of information is therefore a necessity. We observed that the data in social media is noisy, because of a lack of top-down rules and editors to oversee the publication process. The noisiness of the data poses additional challenges to accessing the information.

The main motivation for the research in this thesis was described as follows. We want to allow for intelligent access to, and analysis of, information contained in the noisy texts of social media. To this end, we need to determine topical relevance of social media documents, while countering the specific challenges posed by the noisy character of these documents. We explored various ways of improving retrieval effectiveness, regarding both people and their utterances. We showed that our proposed methods help improve retrieval performance for various information access tasks.

In the next two sections we revisit our research questions and provide answers to each of them. The last section is dedicated to future research directions, following from work in this thesis.

9.1 Main Findings

The goal that we have addressed in this thesis is to improve searching for people and their utterances in social media so as to offer intelligent access to information in those media. We followed observations from Figure 9.1 and explored access to information in social media from two points: (i) people and (ii) their utterances. In Section 1.2 (page 4) we have presented five sets of research questions and in this section we provide answers to each of the five main research questions.

We began the research part of this thesis by investigating how people behave when searching for people and how this behavior relates to social media. We asked:

**RQ 1** How do users go about searching for people, when offered a specialized people search engine to access these people’s profiles?
We have found that people search differs from regular web search in two important ways. First, we observed a much higher percentage of single-query sessions in people search as compared to web search and second, we found a much lower click-through ratio. Another interesting observation is the significant number of searchers that use just one term (i.e., only a first or last name) and start exploring results, this type of exploratory search is so far unsupported by our people search engine.

Looking at the type of queries that users of the people search engine submit, we found three different types: (i) event-based high-profile queries that ask for information on people who are related to some event (e.g., a murder victim or talent show participant), (ii) regular high-profile queries that deal with celebrities and other public persons, and (iii) low-profile queries that ask for information on non high-profile people. The latter type, low-profile queries, takes up over 90% of queries in our dataset, followed by event-based high-profile queries, which occur three times as often as the regular high-profile queries. We experimented with automatic classification of queries into the three types and found that distinguishing between high and low profile is feasible, but that the three-way classification is much harder. The most important features include out clicks, search volume, and news volume.

On a session level we found that most sessions contain queries of different types according to the classification given above, which indicates that we should look into different ways of session detections. Other common session types are repetitive sessions (e.g., spelling variants) and family sessions (i.e., searching for various family members). We have found that on the result side, most users of the people search engine click on social media results, like social networking sites or microblog platforms. Finally, we have used a case study to show that a circle exists from social media, via traditional media and people search, back to social media.

We then shifted our attention from people search with the goal of finding information
9.1. Main Findings

about a person, to finding people based on their utterances (i.e., who is important given a topic?). Here, we looked at the task of finding bloggers and asked:

**RQ 2** Can we effectively and efficiently search for people who show a recurring interest in a topic using an index of utterances?

We have introduced two models, based on previous work in expert finding. Our Blogger model is a blog-based model and aims to rank bloggers directly, based on their utterances. Our Posting model is a post-based model that first ranks individual utterances and then aggregates post scores to a final blogger score. In combining these models, we introduced our two-stage model.

We have shown that by using various pruning and representation techniques we can not only improve the efficiency of our models, but also maintain (and even increase) effectiveness of our models, especially that of the two-stage model. Our two-stage blog feed search model, complemented with aggressive pruning techniques and lean document representations, was found to be very competitive both in terms of standard retrieval metrics and in terms of the number of core operations required.

Moving to the other entry point, viz. people’s utterances, we looked at ways to counter the effects of lacking top-down writing rules and editors in social media. We built upon a previous framework for credibility in blogs and asked:

**RQ 3** Can we use the notion of credibility of utterances and people to improve on the task of retrieving relevant blog posts?

We provided efficient estimations for 10 credibility indicators from the credibility framework proposed by Rubin and Liddy [160], based on textual information in blogs. The indicators were divided into two groups, on the user level (blog level) and on the utterance level (post level). Given that blog post retrieval is a precision-oriented task, we propose two reranking approaches. The first, Credibility-inspired reranking, takes the top \(n\) results of a baseline ranking and reranks these results based on their credibility score alone. Combined reranking multiplies the retrieval and credibility scores of the top \(n\) results and reranks these results on the resulting score.

We have assessed the impact of the individual credibility indicators on blog post retrieval, as well as the combinations of indicators for post level, blog level, and both levels. We have found that most post-level indicators have a positive effect on precision metrics, whereas the performance of most of the blog-level indicators is disappointing. Comparing the two reranking approaches we found that Credibility-inspired reranking is more risky, leading to higher gains and larger drops, while Combined reranking acts as a smoothed version, resulting in less dramatic, but significant (more stable) changes. Both approaches achieve high early precision performances, indicating the usefulness of credibility indicators in blog post retrieval.

We showed that the content of utterances is influenced by the real-world environment in which users live. Sources like newspapers, other social media, television shows, and many more, all give users reasons to write and produce content. We have put this information environment to use and explored it in a query modeling setting. We asked:
9. Conclusions

RQ 4 Can we incorporate information from the environment, like news or general knowledge, in finding blog posts using external expansion?

We explored the use of external collections for query expansion in blog post retrieval. We introduced a general external expansion model that, amongst others, models the query-dependent collection importance. Our External Expansion Model (EEM) can be instantiated in various ways, depending on (in)dependence assumptions one makes, and in one case it boils down to the mixture of relevance models [44] (MoRM). We have found that query expansion using external collections is very effective for blog post retrieval. Each of the external collections we have used (news, web, Wikipedia, and blog posts) led to (mostly significant) improvements and the combination of all four collections gave the best results. We furthermore found that conditioning the weight of the external collection on the query is beneficial for retrieval performance, as our EEM (including this component) outperforms the MoRM (excluding this component).

Analyses with so-called oracle runs have revealed that the impact of the query-dependent collection importance is much higher than that of the collection prior (i.e., a-priori belief a collection is relevant/useful). Although our method for estimating the collection importance was a rather simple one, it already proved very beneficial and promising.

Finally, we zoomed in on utterances and their context within the social media platform. Specifically, we moved to the task of finding emails in an email archive and explored various levels of context that these archives offer, ranging from thread structure to community members. We asked:

RQ 5 Can we incorporate information from the utterances’ context in the task of finding emails?

We identified a number of context levels surrounding emails in an email archive: quote-reply, thread, mailing list, and community levels. We have demonstrated that contextual information can improve retrieval effectiveness, even using a simple query modeling approach. Each of the three context levels we explored (threads, mailing list, and community content) retrieved unique relevant emails, suggesting that each level captures slightly different perspectives. For email search the thread level works best.

We also investigated the effects of using credibility indicators (viz. Chapter 6) in email finding. We have translated three indicators: text quality, thread size, and email length and we found that these credibility indicators can improve further the effectiveness of our retrieval model. Especially the combination of indicators showed good performance, indicating that high quality, long emails in larger threads are preferred over emails lacking these characteristics.

9.2 Future Research Directions

The research presented in this thesis motivates a broad variety of future research projects, most of them aimed at improving over the results presented in the previous chapters, by adding new methods or optimizing existing ones. We do not list each of these smaller
research directions, but focus on four major directions for future research in information access in social media.

**Beyond topical relevance** In this thesis we focused solely on topical relevance (in case of retrieval tasks): find “documents” that are about a given topic. As mentioned in the introduction, many information needs in social media require additional ranking criteria. Not only should a document be about the topic, it should also satisfy other criteria. In Chapter 2 we already referred to work on various ranking criteria that go beyond topical relevance. Opinionatedness is a popular criterion and recency, diversity, and novelty have also received a lot of attention.

We identify three ranking criteria that are challenging, and are without a large body of work so far. First, people often talk about their *experiences* in social media. They refer to things they did yesterday, theme parks they visited, or products they used. Reporting on experiences goes beyond giving opinions in that experiences include descriptions of how something was done or used. These experiences offer a wealth of information to marketeers and product developers: they give insights in aspects of the experiences that people liked or, maybe even more important, did not like. Detecting experiences on a given topic, extracting these, and summarizing them for easy access would be one future research direction.

Two other ranking criteria, related to each other, are whether documents contain *discussions* or viewpoints. Being able to determine viewpoints in social media utterances allows search engines to present searchers with a diverse set of results, based on viewpoints. This ensures that people do not collect information from just one perspective, but get a complete overview of the views on the subject. Discussions play an important role in this, as they can be used as indicators for viewpoints: when people argue, it is likely that they differ in their views on the topic.

Being able to rank documents on other criteria besides topical relevance has an interesting application. In Chapter 6 we have already explored using credibility-inspired indicators as a ranking criterion. The next step would be to assess the ability of this framework to actually measure credibility. To this end we need a collection with credibility assessments, combined with relevance assessments. One step further would be to use credibility as an indication of whether or not to use the document for pseudo-relevance feedback. The motivation behind this is that more credible documents generate “better” query expansion terms than less credible documents. Similarly, we can use a time-based criterion (e.g., recency) to construct time-dependent query models. Initial experiments on credible query expansion have shown promising results.

**Combining people and document relevance** In this thesis we have explored retrieval tasks on the user level (Chapters 4 and 5) and on the utterance level (Chapters 6–8). Although we briefly touched on combining these two levels when we introduced expertise as one of the credibility indicators in Chapter 6, we did not specifically address this issue. In the field of expert retrieval, various attempts have been made at combining document relevance and expertise score [25, 172].

In social media we know who wrote which piece of text and this knowledge could make it easier for combining user and utterance level relevance scores. On the other hand,
social media add characteristics like blogrolls, followers, “like”-s, re-posts, and perma-links that all might play a role in identifying relevant users and utterances. Research into combining evidence from various sources, both textual and non-textual, is necessary to improve retrieval performances in the ever-growing amount of social media utterances. We could, for example, build on previous work by Serdyukov et al. [172] and consider people and utterances in a graph structure, using links, “like”-s, and re-posts as edges.

**Implicit information requests** We have looked at presenting users with results based on a query provided by the user. However, currently people are often connected to a set of streams, that continuously provide the user with new utterances. Examples of such streams are status updates on networking sites or new (micro)blog posts by people to whom the searcher is connected. Instead of waiting for the user to provide a query and search for the proper results, the tasks becomes how to filter relevant information from this continuous stream of utterances and how to present this in an efficient way.

Imagine two use case scenarios. (1) A user uses her smartphone to keep up-to-date, but is unable to keep up with each individual utterance produced in the streams she follows. Here, we could, for example, try to identify the most popular topics in the streams and provide this user with a summary of the topic. We are left with tasks like topic detection, summarization, and linking of utterances to external sources to provide context, each of which is in itself a challenging task. (2) A user follows a set of streams out of professional interest and is interested in those utterances that are actually about her profession. Here, we should take this user’s profile into account when filtering information from each stream, assuming that what she writes about reflects her interests. This second example is similar to recommender systems. These systems recommend information items to users based on their interests [2]. In the case of social media streams, however, we are dealing with a large set of streams from very different sources (e.g., very short tweets, longer blog posts, picture, videos, network updates, questions posted to forums, etc.). The challenge becomes how to develop a model that can combine the streams from these different sources into one set of recommendations.

A related research direction is real-time search. Here, searchers want information about a topic that is currently “hot” or happening. This type of search poses challenges both on the indexing side (e.g., how to perform real-time indexing) and on the accessing side. How do we know which topics are currently happening? Which utterances belong to this particular topic? How do we present all the utterances on this topic to the searcher? Again, different research topics should be combined to facilitate this type of search task.

**Prediction** Finally, we observe a shift in tasks, from retrieving and informing, to predicting. Social media allows us to gain insights in people’s behavior in large volumes and over a longer period of time. These new insights give us the opportunity to answer new questions, like, what will be popular tomorrow? Or next week? How will people respond, if they respond at all? Initial work on prediction already shows promising results (as shown in Chapter 2). We can predict number of comments for a news article, number of views for videos, and clicks for ads. Using more advanced models and larger datasets, we should be able to generate more accurate predictions and do prediction on more challenging issues, like activities, rioting, or even revolutions.