Enriching Information Retrieval

Paul N. Bennett
Microsoft Research
paul.n.bennett@microsoft.com

Khalid El-Arini
Carnegie Mellon University
kbe@cs.cmu.edu

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

Krysta M. Svore
Microsoft Research
ksvore@microsoft.com

1 Introduction

Most information retrieval systems and tasks are now embedded in a rich context. Documents no longer exist on their own; they are connected to other documents, they are associated with users and their position in a social network, and they can be mapped onto a variety of ontologies. Similarly, retrieval tasks have become more interactive and are solidly embedded in a user's geospatial, social, and historical context. We conjecture that new breakthroughs in information retrieval will not come from smarter algorithms that better exploit existing information sources, but from new retrieval algorithms that can intelligently use and combine new sources of contextual metadata.

The goal of the Enriching Information Retrieval workshop at SIGIR 2011 was to explore how new and emerging sources of contextual metadata can be used for improving information retrieval, including ranking, personalization, diversification, and faceted search. In particular, we focused the workshop on three themes:

- The identification of novel types and sources of contextual metadata (e.g., new ontologies, usage patterns, locality information, readability, temporal).
- The automatic acquisition and distillation of metadata (e.g., via learning or through implicit data).
- The design of methods for exploiting new metadata sources in IR tasks. A special focus of the workshop was on metadata and retrieval tasks associated with social networks.

The workshop program committee consisted of twenty researchers from across academia and industry, with experience in information retrieval, machine learning, evaluation methodology, natural language processing, recommender systems and social networks. Around 30 people attended the workshop in Beijing.

The purpose of this report is to summarize the workshop, in particular to highlight the common themes that arose during our discussions and relate the outcomes of the discussions to future directions for research.

2 Workshop Structure

The workshop was organized with the specific intention of fostering discussion, while avoiding the prospect of a "mini-conference." As such, all submissions were limited to two pages (plus a page for references), and the schedule was heavily tilted towards both informal and targeted discussions.

In total, we had two invited talks, four contributed talks, a poster session, and several free-flowing discussion periods. The first session consisted of an invited talk by Jennifer Neville from Purdue University, titled "Mining Social Network Activity to Understand and Predict User Behavior," and a contributed talk from Michael Cole of Rutgers University, "Physiological Data as Metadata: A Position Paper." Before the lunch break, we had the first of three discussion sections, followed by a poster session consisting of seven posters. These posters — from both industry and academia — touched on topics ranging from emotion detection for use in music retrieval to topic retrieval from Twitter data.

The afternoon session started with an invited talk by Microsoft's Filip Radlinski on "Evaluating Rich Models in Context," followed by three contributed talks: "Personalizing Local Search with Twitter" (Matthew Lease, University of Texas in Austin), "Enriching Information Retrieval with Reading Level Prediction" (Kevyn Collins-Thompson, Microsoft Research), and "Future Retrieval: What Does the Future Talk About?" (Ricardo Campos, Polytechnic Institute of Tomar).

The workshop culminated in an hour-long discussion period and brainstorming session, where, in an attempt to summarize the day's proceedings, the discussion focused on three aspects:

1. What are possible sources of metadata that can be used to enrich information retrieval?
2. How can these sources be used?
3. What are the primary concerns for using this metadata in these ways?

The remainder of this workshop report documents these takeaway points from the discussion. (More details on the specific invited and contributed talks can be found in the online proceedings of the workshop, available at http://select.cs.cmu.edu/meetings/enir2011.)

3 Sources and Uses of Contextual Metadata

One outcome of this workshop was the construction of an annotated list describing different sources of metadata that could be used to enrich existing information retrieval problems. This list is as follows:

- **Social networks**. The documents retrieved in modern information retrieval systems have authors, and these authors have social relationships with other people and entities (e.g., companies, universities, locations, etc.). As emphasized by Jennifer Neville in the opening talk of the workshop, social networks provide rich information that can be leveraged to predict future relationships between people, and thereby potentially play a role in document relevance for an information retrieval task. At the same time, incorporating dynamic, large-scale graph structures into retrieval tasks, whether for personalization or other uses, poses interesting technical and research challenges. For instance, can we model a personalized

notion of trust using a user's social network? Are there different retrieval settings that benefit more from such social augmentation than traditional keyword search?

Matthew Lease's contributed talk provided additional points for discussion; he discussed how a user's social network can be extracted from Twitter and then be used to improve the specific task of Local Search.

- **Physiological data**. Michael Cole's talk touched on the potential for using physiological data at query time to determine such things as the difficulty level of the search task, the level of a user's domain knowledge, etc. A discussion followed on the inevitable privacy concerns that would be raised by having eye tracking active for every web search issued by a user. Nevertheless, with a variety of sensors available, both in mobile devices and in computer systems, that have the potential to measure physiological information, the technical and ethical questions of using such sensors to personalize search results are important to answer.

- **Sentiment and emotion**. There is significant work outside of the information retrieval community on sentiment and emotion detection, and the feeling of the attendees was that this work could be leveraged to improve search results. Such context can be incorporated in two manners: (1) by detecting the sentiment of the document being retrieved, a diversity of results can be presented (e.g., a representative selection of comments on a blog post or news article), and (2) by detecting the emotional state of the user, retrieval results can be personalized accordingly (e.g., Lijuan Zhou's poster on enriching music retrieval with emotion).

- **Supervised/crowdsourced ontologies**. Sources such as Wikipedia can be used to augment the often unsupervised problem of organizing the content of a corpus into main ideas or topics (e.g., via a latent variable topic model such as latent Dirichlet allocation). This type of augmentation to a corpus has often served as a basis for personalization or clustering in information retrieval studies. Along these lines, the poster by Min & Jones presented an examination of the use of Wikipedia categories for personalized retrieval. Demonstrating the use of both social interactions and topical representations for enrichment, the poster by Pochampally & Varma looked at extracting a "topical keyword" representation of twitter users by leveraging both their tweets and their strength of interactions with other users (mentions, re-tweets, and follows).

- **Reading level**. As Kevyn Collins-Thompson described in his contributed talk, reading level can play an important role in the utility of a retrieval system. For example, search snippets that are at an easier reading level than the page they represent often lead to abandonment. Moreover, the global reach of the Web leads to users perusing pages in languages other than their native tongue, at various levels of proficiency. Even within a single language, experts (e.g., a biology professor) have different expectations for the documents they search for than novices (e.g., middle school biology students). Collins-Thompson described that different aspects of reading level can be more easily modeled and predicted, and that research in this area has promise for improving search performance.

The following sources of metadata were also mentioned, but were not a key part of our discussions (mainly due to the condensed nature of a single day workshop): **temporal and geospatial**

context; **desktop content**; **browsing/search history** (already commonly used to augment and personalize search results); **language**; **parallel corpora**; **mobile phone sensors**.

4 Challenges & Research Directions

Much of the discussion focused on the opportunities that new sources of metadata provide, as well as the research that is necessary to enable their use. However, a substantial part of the discussion also considered potential problems and stumbling blocks:

- **Overpersonalization and privacy**. Perhaps the most obvious way to enrich information retrieval by using the above sources of contextual metadata is to personalize results to individual users. Whether taking into account who their friends are, what their reading level is, or which emotional state they are in, users can expect to see results more tailored to their tastes when such side information is employed. While the potential for providing significantly improved results is great, this should be weighed against real concerns of overpersonalization and privacy. For example, if these sources of metadata can provide such specific user models that each user sees results only from a very (ideologically) narrow slice of the Web, would there be social consequences?

  More obviously, some of the discussed metadata sources, such as EEG or eye-tracking devices, represent not just the use of information that is often already available (e.g., one's Twitter friends), but an entirely new dimension of personal data to be collected. This inevitably raises significant privacy questions that must be dealt with in any future system that employs such technologies.

- **Data availability**. Some of the most promising sources of metadata that could be used as side information in an information retrieval setting are hard — if not impossible — to access for academic researchers. In some cases, companies like Twitter make portions of their data available for academic use, but in many cases, such data is difficult to come by. This is a long-standing problem in information retrieval with no immediate solution, but one option discussed that is readily available and recognizably successful is having graduate students intern at industrial companies with large amounts of data. There, students have access to data over an extended time period and can learn important lessons from working on real retrieval settings. Ultimately, such lessons, gathered from real-world scenarios and data, will influence them in their academic research.

- **Annotated corpora**. While some user-centric metadata is difficult to access in an academic setting, document-centric metadata could be incorporated into existing corpora more easily. For example, there was discussion on augmenting the ClueWeb corpus with features describing trustworthiness (transaction), credibility (statement), sentiment, reading level, topic (supervised), aspect (unsupervised topic), genre, freshness, location (addresses, interest), spamminess, and linguistic features (e.g., noun phrases). Moreover, many of the attendees were in favor of having such a resource that would enable research in applications and scenarios that leverage these enriched data rather than trying to design the augmentation around a specific application. Moreover, by controlling for the augmentation, it reduces

variability in replicating other research results (i.e., one can be certain that differences are a result of the use of the enriched data and not in how the enriched data were generated). This is similar to a point made by Jamie Callan during his keynote at CIKM 2010 where he argued that enriching corpora in advance of a specific application need will help drive the diversity, depth, and novelty of applications that leverage the information since they reduce the bar to entry.

- **Beyond reranking**. Many of the current approaches to incorporating contextual metadata in a retrieval system involve a reranking procedure, where one algorithm decides which documents should be near the top of the list, and a separate one decides how to modify this ranked list based on the metadata of interest. There was a discussion at the workshop of moving beyond this reranking paradigm, perhaps leading to entirely new retrieval algorithms that incorporate this additional data at a fundamental level.

- **Applications.** In an effort to list information retrieval applications that would greatly benefit from using contextual metadata, the workshop attendees decided it would be a good exercise to think about which current retrieval tasks are poorly solved. Examples cited included local search (e.g., finding a place to eat or a hotel to stay at, etc.) and discovering relevant scientific literature. The point was also made that the consumer of an IR system does not have to be a person, but could also be an application itself (e.g., machine translation, intelligent tutoring, etc.). Additionally, the idea of providing a lower-level API for a retrieval system (potentially giving developers the ability to produce their own reranked results) was floated as a way to spawn innovation. Finally, workshop attendees noted that retrieval with contextual metadata should not be limited to text-oriented tasks, but could also include richer media (e.g., retrieving images, videos, music, mobile apps).

- **Collaboration with HCI**. At the end of the day, the users of information retrieval systems are trying to solve some task. There is an existing literature – particularly in the human computer interaction (HCI) community – that studies how people solve problems and complete tasks, and in the final discussion session of the workshop, we discussed ideas for exploiting this synergy. A common HCI approach for such tasks is to design a new interface or user experience, and then attach an existing IR algorithm under the hood. Would an explicit collaboration between IR and HCI researchers lead to better, more usable retrieval systems? Workshop attendees found this idea particularly exciting. For example, the idea was floated to have a TREC-style competition that involved both communities, with the final project perhaps being an app for a mobile device (e.g., iPhone, Android, or Windows).

- **Evaluation methods**. Solving novel retrieval tasks that differ from the traditional "keyword query and ten blue links" paradigm means that traditional evaluation metrics and benchmark data sets are unlikely to directly translate. Discussion was held on the importance of testing with real users, either in an academic user study setting, or on an industrial live site.

- **Presentation**. Using metadata as side information for enriching search results leads to the interesting question of how to display the results. Specifically, can the results be augmented by displaying some information from the metadata? Obvious examples include displaying

search results on a map if they are tagged with geospatial information, or displaying relevant people from the user's and/or author's social network in addition to the retrieved documents. This increases transparency and could give the users more confidence in the retrieval system.

5 Conclusion

Overall, the Enriching Information Retrieval workshop at SIGIR 2011 led to an exploration of both types of enrichment and IR applications for these types of enrichment. The work presented at the workshop demonstrated the usefulness of a variety of enrichment types including topical, temporal, reading level, sentiment, comprehensibility, phsyiological, social strength of interaction, reading level, and location. While a majority of the work focused on leveraging enriched representations in personalization, both the work presented and the discussions explored other applications such as quantifying task difficulty and user cognitive load, keyword extraction and contextual advertising, and retrieval about uncertain events (future events where speculation may be present in the corpus).

As highlighted in the subsections on Annotated corpora and Applications in Section **??** above, many of the workshop attendees were interested in investigating applications and scenarios that leverage enriched data. As a direction for the research community, a majority of the attendees felt that augmenting an existing public research corpus would serve the general interests of the research community. One concrete outcome of this workshop is that we are actively exploring concrete interest in contributing to such a resource. For those readers interested in contributing to this effort, please mail Paul Bennett (`paul.n.bennett@microsoft.com`). Participants will be expected to provide the augmentation (e.g., output of a classifier) as well as a reference that describes the methodology.