Seventh Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'14)

CIKM 2014 Workshop

Omar Alonso
Microsoft
Mountain View, CA

Jaap Kamps
University of Amsterdam
The Netherlands

Jussi Karlgren
KTH & Gavagai
Stockholm, Sweden

ABSTRACT

There is an increasing amount of structure on the Web as a result of modern Web languages, user tagging and annotation, emerging robust NLP tools, and an ever growing volume of linked data. These meaningful, semantic, annotations hold the promise to significantly enhance information access, by enhancing the depth of analysis of today's systems. The goal of the ESAIR'14 workshop remains to advance the general research agenda on this core problem, with an explicit focus on one of the most challenging aspects to address in the coming years. The main remaining challenge is on the user's side—the potential of rich document annotations can only be realized if matched by more articulate queries exploiting these powerful retrieval cues—and a more dynamic approach is emerging by exploiting new forms of query autosuggest. How can the query suggestion paradigm be used to encourage searcher to articulate longer queries, with concepts and relations linking their statement of request to existing semantic models? How do entity results and social network data in "graph search" change the classic division between searchers and information and lead to extreme personalization—are you the query? How to leverage transaction logs and recommendation, and how adaptive should we make the system? What are the privacy ramifications and the UX aspects—how to not creep out users?

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: Graph Search; Query Suggest; Semantic Annotation

1. THEME AND TOPICS

The goal of the seventh ESAIR workshop is to create a forum for researchers interested in the use of application of semantic annotations for information access tasks. By semantic annotations we refer to linguistic annotations (such as named entities, semantic classes or roles, etc.) as well as user annotations (such as microformats, RDF, tags, etc.).

There are many forms of annotations and a growing array of techniques that identify or extract information automatically from texts: geo-positional markers; named entities; temporal information; semantic roles; opinion, sentiment, and attitude; certainty and hedging to name a few directions of more abstract information found in text. Furthermore, the number of collections which explicitly identify entities is growing fast with Web 2.0 and Semantic Web initiatives. In some cases semantic technologies are being deployed in active tasks, but there is no common direction to research initiatives nor in general technologies for exploitation of non-immediate textual information, in spite of a clear family resemblance both with respect to theoretical starting points and methodology. We believe further research is needed before we can unleash the potential of annotations!

The previous ESAIR workshops made concrete progress in clarifying the exact role of semantic annotations in support complex search tasks: both as a means to construct more powerful queries that articulate far more than a typical Web-style, shallow, navigational information need, and in terms of *making sense* of the retrieved results on very various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space.

2. OBJECTIVES, GOALS, AND OUTCOME

The ESAIR'14 workshop will have far more focus than the earlier ESAIRs. While the goal remains to advance the general research agenda on this core problem, there is an explicit focus on the main remaining challenge of exploiting semantic annotations in the coming years.

One of the main outcomes of the previous ESAIRs has been not only an overview of various domains of application and experiments on real life data, but also a clearer "theoretical" view on the role of semantic annotations. The starting point, based on discussions at previous ESAIRs is a view of semantic annotation as a *linking procedure*, connecting a *content analysis* of information objects with a *semantic model* of some sort. All three are objects of study in their own right; the point of the ESAIR series is linking those three activities into a coherent and practical whole.

The obvious next step in the discussion is how to leverage known semantic resources (such as knowledge bases, ontologies, folksonomies, lexical resources, hand-annotated or not) to streaming realistic-scale data ("big data"), to be processed in real time, with incrementally evolving knowledge models. The challenge is to use an existing resource as a semantic model, provide an effective and practicable content analysis, and a scalable linking procedure which can handle the data flows we can expect in real life data.

Whilst the exact scope and reach of the emerging knowledge resources (such as DBpedia, Freebase) is not yet clear, there is a clear

CIKM'14, November 3–7, 2014, Shanghai, China.
ACM 978-1-4503-2598-1/14/11.
http://dx.doi.org/10.1145/2661829.2663539.

focus on enumerating factual content that can fruitfully be complemented by non-topical aspects. Over the last years there has been a massive interest in annotations on non-topical dimensions, such as opinions, sentiment or attitude, reading level, prerequisite level, authoritativeness, credibility, etc, both at the level of individual sentences or utterances as well as at more aggregative levels. It is clear that such annotations contain vital cues for matching information to the specific needs and profile of the searcher at hand, yet it is an open question how such annotations can be fruitfully exploited in information retrieval, either as additional criteria on the "relevance" of results in traditional search tasks, or in specific use cases where non-topical cues are key, or in contextual or personalized search that takes the searcher's state into account.

Both in terms of knowledge bases and in terms of non-topical annotation significant progress have been made in recent years. The main remaining challenge is on the user's side—the potential of rich document annotations can only be realized if matched by more articulate queries exploiting these powerful retrieval cues—and a more dynamic approach is emerging by exploiting new forms of query autosuggest. How can the query suggestion paradigm be used to encourage searcher to articulate longer queries, with concepts and relations linking their statement of request to existing semantic models? How do entity results and social network data in "graph search" change the classic division between searchers and information and lead to extreme personalization—are you the query? How to leverage transaction logs and recommendation, and how adaptive should we make the system? What are the privacy ramifications and the UX aspects—how to not creep out users?

3. ACCEPTED PAPERS

We requested the submission of short, 3 page papers to be presented as boaster and poster. We accepted a total of 11 papers out of 15 submissions after peer review (a 73% acceptance rate).

Cotelo et al. [2] investigate semantic cues to articulate more expressive queries by reviving various query operators and explore their value in a preliminary evaluation.

De Nies et al. [3] give a broad overview of the challenges in the context of entity tagged corpora, focusing on the annotation quality, appropriate similarity measures, data quality, and access problems.

Deolalikar [4] investigates within corpus text mining to cluster documents and combine cluster and document scores, demonstrating that coarse grained clusters are unable to capture specific intent of topically focused queries.

Ibrahim et al. [5] address the problem of entity linking in social streaming data, looking into the normalization of mentions due to cryptic abbreviations, the contextualization of short postings by shared hashtags, persons, and links, and the temporal trends of attention to time-sensitive entities.

Jan et al. [6] study the specific domain of searching IT service desk tickets, based on topic modeling, concept analysis, and clustering, leading to increased performance on a corpus of noisy statements of IT related problems.

Jiang et al. [7] investigate some heuristics to improve "explicit semantic annotation" by labeling documents with Wikipedia concepts.

Li et al. [8] revisit the answer type prediction problem of question answering systems, using dependency parsing and semantic role labeling rather than ad hoc heuristics.

Mao and Lu [9] focus medical literature search and return to the old problem of using controlled subject headings with a mixture language model and show that this promotes retrieval effectiveness.

Verma and Ceccarelli [10] study the problem of entity detection in non-head queries, observing similarities and differences in the types of entities occurring in slices of queries.

Yang [11] studies concept similarity measures comparing tree edit distance with textual similarity of subtrees or fragments over the open directory project's concept hierarchy.

Zuccon et al. [12] investigates reasoning with rigorous semantic concept hierarchies in medical literature search, and discusses the potential benefits of semantic-based retrieval as well as the risks of unconditionally embracing such inferences.

4. FORMAT

We start the day with a short introduction of the goals and schedule, and a "feature rally" in which each participant introduced her- or himself, and stated her or his particular interest in this area. Next, we have keynote speakers that help frame the problem, and create a common understanding of the challenges. We continue with a boaster/poster session, where the papers from Section 3 are presented. The poster session continues over lunch. After lunch, we have break-out sessions in parallel that focused on specific aspects or problems related to the four themes. After the afternoon coffee, we have reports of the breakout sessions, followed by a final discussion on what we achieved during the day and how to take it forward. The workshop will continue with a more informal part, over drinks and dinner with all attendees of the workshop.

Acknowledgments

We thanks the CIKM workshop chairs (Huan Liu and Xiaofeng Meng) and the local organization team (Lanying Zhang, Xiaoyang Sean Wang) and Sheridan Printing (Lisa Tolles and Cindy Edwards) for their great support.

5. REFERENCES

[1] O. Alonso, J. Kamps, and J. Karlgren, editors. *ESAIR'14: Proceedings of the CIKM'14 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2014. ACM Press.
[2] S. Cotelo, A. Makowski, L. Chiruzzo, and D. Wonsever. Documents search using semantics criteria. In Alonso et al. [1], pages 1–3.
[3] T. De Nies, C. Beecks, W. De Neve, T. Seidl, E. Mannens, and R. Van de Walle. Towards named-entity-based similarity measures: Challenges and opportunities. In Alonso et al. [1], pages 4–6.
[4] V. Deolalikar. Can corpus similarity-based self-annotation assist information retrieval? In Alonso et al. [1], pages 7–9.
[5] Y. Ibrahim, M. A. Yosef, and G. Weikum. Aida-social: Entity linking on the social stream. In Alonso et al. [1], pages 10–12.
[6] E.-E. Jan, K.-Y. Chen, and T. Ide. A probabilistic concept annotation for it service desk tickets. In Alonso et al. [1], pages 13–15.
[7] Z. Jiang, M. Chen, and X. Liu. Semantic annotation with rescoredesa: Rescoring concept features generated from explicit semantic analysis. In Alonso et al. [1], pages 16–18.
[8] Z. Li, P. Exner, and P. Nugues. Using semantic role labeling to predict answer types. In Alonso et al. [1], pages 19–21.
[9] J. Mao and K. Lu. Leverage the associations between documents, subject headings and terms to enhance retrieval. In Alonso et al. [1], pages 22–24.
[10] M. Verma and D. Ceccarelli. Bringing the head closer to the tail with entity linking. In Alonso et al. [1], pages 25–27.
[11] H. Yang. A fragment-based similarity measure for concept hierarchies and ontologies. In Alonso et al. [1], pages 28–30.
[12] G. Zuccon, B. Koopman, and P. Bruza. Exploiting inference from semantic annotations for information retrieval: Reflections from medical ir. In Alonso et al. [1], pages 31–33.