Fifth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'12)

CIKM 2012 Workshop

Jaap Kamps
University of Amsterdam

Jussi Karlgren
Gavagai Stockholm

Peter Mika
Yahoo! Research

Vanessa Murdock
Microsoft Bing

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. Currently, we have only started exploring the possibilities and only begin to understand how these valuable semantic cues can be put to fruitful use. To complicate matters, standard text search excels at shallow information needs expressed by short keyword queries, and here semantic annotation contributes very little, if anything. The main questions for the workshop are how to leverage the rich context currently available, especially in a mobile search scenario, giving powerful new handles to exploit semantic annotations. And how can we fruitfully combine information retrieval and semantic web approaches, and for the first time work actively toward a unified view on exploiting semantic annotations.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Selection process*
General Terms: Algorithms, Experimentation, Theory
Keywords: Semantic Annotation

1. THEME AND TOPICS

The goal of the fifth 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 micro-formats, 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, and in particular the fourth ESAIR at CIKM'11 and third ESAIR at CIKM'10, 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 various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space.

2. OBJECTIVES, GOALS, AND OUTCOME

The general aim of ESAIR'12 is not the technologies for semantic annotation itself, but rather the *applications* and *contributions* of semantic annotation to information access tasks on various levels of abstraction such as ad-hoc retrieval, classification, browsing, textual mining, summarization, question answering, etc. ESAIR'12 will focus the discussion on two of main insights from earlier ESAIRs:

First, one of the main outcomes of ESAIR'10 was to recognize that semantic annotations are no panacea, and have clearly more potential in areas characterized by the need for i) rich context, ii) for interaction, and iii) for combining different types of data. The potential of semantic annotations in this setting is huge, but this may result in our searcher needing to articulate a complex information request in a complex query language, requiring full awareness of the used annotation schemes. It is crucial to prevent that the onus for exploiting semantic annotation is put on the searcher. The mobile search scenario, which is particularly context-rich, is an ideal scenario to push the ESAIR agenda. Processing data from mobile users allows a wide range of contextual information not available in many other usage situations. Besides personalization and geo-positional information, mobiles have a wide and growing range of locational, mechanical and even biometrical sensor data available to them. In an information retrieval situation this allows the system to infer task and situational context to flesh out the topical content of the query itself.

Second, one of the main outcomes of ESAIR'10 and '11 of was a clearer "theoretical" view on the role of semantic annotations. ESAIR'10 concluded with viewing semantic annotation as a *linking* procedure, connecting an *analysis* of information objects with a *semantic model* of some

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.

sort. ESAIR'11 further explored this view focusing on the "exploitation" aspects—how this can be leveraged to some gainful *task* of interest to end users. Interestingly, the resulting view still allows for a wide range of views on semantic annotations—including radically opposing views as held in *information retrieval* (relying on statistical methods modeling uncertainty) and *semantic web* (relying on knowledge-intensive methods based on certainty). These opposing views did surface during the breakout groups at earlier ESAIRs, highlighting different underlying assumptions, and different modes of information access assumed. Both views respond differently to the trade-off between the desire to enforce a messy world into clean data structures, and the need to do justice to every unique searcher and search request, in a world of partial and uncertain information. We firmly believe that the time has come to delve deeper into these underlying assumptions, clear up under which conditions each approach has benefits, and work toward an integrated view on semantic annotations for information access tasks. Hence, ESAIR'12 will focus the discussion on, simply put, an head-to-head of information retrieval and semantic web, and for the first time work actively toward a unified view on exploiting semantic annotations.

3. ACCEPTED PAPERS

We requested the submission of short, 2 page papers to be presented as boaster and poster. We accepted a total of 10 papers out of 13 submissions.

Balog and Nørvåg [1] suggest to extend existing work on entity search with the temporal dimension, i.e. searching over knowledge bases where the temporal validity of facts is well defined and the information needs may have temporal constraints.

Das and Gambäck [2] investigates the 5W annotation (Who, What, When, Where, Why) of a sentiment/opinion corpus that pilots this kind of annotation in Bengali.

Eklund [3] investigates mapping "end-user" search terms to the appropriate medical terminology using the UMLS, addressing the problem of dealing with natural language searches in systems that use controlled vocabularies.

Fujita et al. [4] presents several ways to identify query rewrites based on the click behavior of users, and a topic hierarchy of the Yahoo directory.

Mishra et al. [7] describes the creation of an important new benchmark corpus, integrating Wikipedia with the knowledge bases DBpedia and Yago. In addition it comes with 90 SParQL queries based on Jeopardy questions that are conjunctive queries on the structure part plus free text queries on the textual part of the corpus.

Nomoto and Kando [8] address the problem of labeling unstructured documents with labels generated from combinations of Wikipedia article titles and section headers.

Sellami and Rodríguez [9] address the task of measuring the quality of annotations for Semantic Web services, in terms mappings between schema elements and ontological concepts in a reference ontology.

Sojka [10] discusses the semantic annotation of mathematics in large scale digital libraries, by augmenting surface texts (including math formulae) with additional linked representations providing semantic information (expanded formulas as text, canonicalized text and sub-formulas).

Kristianto et al. [6] propose a framework for annotating scientific papers for mathematical formulae search,which in essence extracts surrounding text and classifies that text.

Yoshioka and Kando [11] presents a system that supports news searches where the user can specify hybrid structured queries involving explicit named entities, news metadata (source, date), and text keywords.

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 two keynotes (to be announced) 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 thank the EU FP7 Parlance project for sponsoring the workshop.

REFERENCES

[1] K. Balog and K. Nørvåg. On the use of semantic knowledge bases for temporally-aware entity retrieval. In Kamps et al. [5], pages 1–2.

[2] A. Das and B. Gambäck. Exploiting 5w annotations for opinion tracking. In Kamps et al. [5], pages 3–4.

[3] A.-M. Eklund. Why query annotations may help in providing accurate public health information. In Kamps et al. [5], pages 5–6.

[4] S. Fujita, G. Dupret, and R. Baeza-Yates. Semantics of query rewriting patterns in search logs. In Kamps et al. [5], pages 7–8.

[5] J. Kamps, J. Karlgren, P. Mika, and V. Murdock, editors. *ESAIR'12: Proceedings of the CIKM'12 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2012. ACM Press.

[6] G. Y. Kristianto, G. Topic, M.-Q. Nghiem, and A. Aizawa. Annotating scientific papers for mathematical formulae search. In Kamps et al. [5], pages 17–18.

[7] A. Mishra, S. Gurajada, and M. Theobald. Design and evaluation of an ir-benchmark for sparql queries with full-text conditions. In Kamps et al. [5], pages 9–10.

[8] T. Nomoto and N. Kando. Conceptualizing documents with wikipedia. In Kamps et al. [5], pages 11–12.

[9] S. Sellami and C. C. G. Rodríguez. Semantic annotation: What about quality? In Kamps et al. [5], pages 13–14.

[10] P. Sojka. Exploiting semantic annotations in math information retrieval. In Kamps et al. [5], pages 15–16.

[11] M. Yoshioka and N. Kando. Multifaceted analysis of news articles by using semantic annotated information. In Kamps et al. [5], pages 19–20.