Third Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR)

CIKM 2010 Workshop

Jaap Kamps
University of Amsterdam

Jussi Karlgren
SICS Stockholm

Ralf Schenkel
MPI/Saarland University

ABSTRACT

There is an increasing amount of structure on the Web as a result of modern Web languages, user tagging and annotation, and emerging robust NLP tools. 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. Unleashing the potential of semantic annotations requires us to think outside the box, by combining the insights of natural language processing (NLP) to go beyond bags of words, the insights of databases (DB) to use structure efficiently even when aggregating over millions of records, the insights of information retrieval (IR) in effective goal-directed search and evaluation, and the insights of knowledge management (KM) to get grips on the greater whole.

The Workshop aims to bring together researchers from these different disciplines and work together on one of the greatest challenges in the years to come. The desired result of the workshop will be concrete insight into the potential of semantic annotations, and in concrete steps to take this research forward; synchronize related research happening in NLP, DB, IR, and KM, in ways that combine the strengths of each discipline; and have a lively, interactive workshop were everyone contributes and that inspires attendees to think "outside the box."

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Algorithms, Experimentation, Theory

Keywords: Semantic Annotation

1. SCOPE

The goal of this proposed workshop is to create a forum for researchers interested in the use 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.). The aim of this workshop is not semantic annotation itself, but rather the *applications* 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.

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.

The latest ESAIR workshop ended with the suggestion that semantic annotations might be the way to provide a path towards *making sense* of data on very various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space. This is a first thought of how to conceptualise a framework to integrate the various analyses we have recourse to. But we believe further research is needed before we can unleash the potential of annotations! This workshop is intended to provide a focus point for discussion on future directions, this year involving the CIKM audience, with its useful mix of IR, KM and DB competence.

2. CHALLENGE QUESTIONS

This workshop is intended to address the following challenge questions. The previous two workshops were exploratory workshops to discuss the research space around the topic; this workshop intends to propose future directions for the benefit of the field as a whole. Specifically, we aim to bring together a varied group of researchers covering NLP, IR, DB, and KM, and together identify the *barriers* to success and work on ways of addressing them.

A provisional list of themes for the workshop:

Application/Use Case What are *use cases* that make obvious the need for semantic annotation of information? What tasks cannot be solved by document retrieval using the traditional bag-of-words? What are the prerequisites of successful application?

Annotation What types of annotation are available? Are there crucial differences between author-, software-, user-, and machine-generated annotations? Named entities, temporal expressions on the one hand and sentiment and hedging on the other are examples of analyses beyond topic that have moved to profitable application. What is holding back the widespread use of these annotations? Are there other types of annotations that are within our grasp?

Result Aggregation Whereas IR focuses almost exclusively at finding individual chunks of information, DB naturally focuses on results that combine information and produce aggregated results (think of OLAP queries), and KM naturally deal with the whole information space. How can we fruitfully combine these strengths?

CIKM'10, October 25–29, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.

Searcher/Query With shallow 2.4 word navigational queries, there may be little benefit in semantic annotations. What expressive power is hidden in the semantic annotation? What is keeping searchers from exploring these powerful search request?

The Workshop will conclude with a final session addressing the best way forward to unleash the potential of semantic annotation.

3. ACCEPTED PAPERS

We requested the submission of short, 2 page papers to be presented as boaster and poster. We accepted a total of 17 papers out of 19 submissions.

Anh and Yukawa [1] investigate automatic annotation using a "concept base," and the use of the annotations for CLIR, to retrieve, to detect mistranslations, and to rerank results.

Azzam and Roelleke [2] propose the classification of queries in classes of varying semantic complexity. This classification can then be used for several purposes, one might be to call an appropriate search engine after a query is parsed and classified.

Badia [3] asks whether formalizing events is necessary for their full exploitation, and studies the merits of different axiomatizations.

Baskaya et al. [4] discuss "WebExplorer" a tool for constructing search ontologies containing synonyms and translations, and using this tool for cross-language information exploration.

Bowers et al. [5] introduce a system for adding semantic annotations to observational datasets, with a use case from ecology, and discuss the use of the resulting annotation framework.

de Vries et al. [6] advocates "search by strategy," a novel user-driven interactive search formalism that helps searchers construct complex queries exploiting the semantic annotations.

Fortuna et al. [7] study predicting user demographics (such as age and gender) of a news-site from their visiting history. Performance is shown to improve when named entities and/or editorial annotations are taken into account.

Gey et al. [8] discusses the use of semantic annotations in geo-temporal search, as done in geo-IR in general and at the NCTIR GeoTime Task in particular, and argues for date-stamped topics.

Lagos et al. [10] discusses the role (semi) automatic annotations can play in solving the e-discovery problem.

Marrero et al. [11] propose a specific formalization of rule-based patterns for semantic annotation and information extraction.

Marshall [12] propose a graph representation of multimedia objects, including information in different media and from different sources such as user tags.

Palacios et al. [13] describe an approach to the integration of semantic information that is associated with documents with heterogeneous semantic annotations.

Said and Luca [14] investigate contextual recommendations based on hierarchical tags for various facets.

Shiells et al. [15] proposes the grouping of tweets by the URL they contain and then considering the textual content of these tweets as social annotations of the URL.

Tichy et al. [16] propose using semantic annotation as part of the software specification and lifecycle system, by using NLP to extract semantic tags from the specifications and following those tags through to the development process.

Trippel et al. [17] proposes a "simple" interface for expressive, but complex, query languages on annotated data.

Velupillai [18] discusses "electronic health records" which are a combination of structured and unstructured content, and how semantic annotation can help structure the unstructured content making it available for further analysis.

4. FORMAT

We will start the day with two keynotes that help frame the problem, and create a common understanding of the challenges. Liz Liddy (Syracuse University) will discuss the problem from a birds eye view, based on her extensive experience in NLP and IR. Maarten Marx (University of Amsterdam) will demonstrate the extraordinary power of querying annotated documents: even a little annotation will take you a long way.

After the morning coffee, we will continue with a boaster/poster session, where the papers from Section 3 will be presented. After lunch, when otherwise creativity might run low, we will schedule a *feature rally*, where every participant is given about one or two minutes (and maximum one slide) to describe their favourite envisioned technology or idea for future technologies. Next, we will have break-out sessions in parallel that focus on specific aspects or problems. After the afternoon coffee, we have reports of the break-out sessions, followed by a final discussion on what we achieved during the day and how to take it forward.

The goal of the workshop is to produce a joint statement on future directions of purpose-driven semantic analysis, taking the challenge questions above as point of departure. This joint statement is to be coauthored by all participants and published in some suitable journal in white paper form as an archival record of the deliberations of the workshop.

REFERENCES

[1] P. H. Anh and T. Yukawa. Cross language information retrieval based on concept base and language grid. In Kamps et al. [9].

[2] H. Azzam and T. Roelleke. A semantic query rating scheme. In Kamps et al. [9].

[3] A. Badia. Is formalizing events necessary for full exploitation. In Kamps et al. [9].

[4] F. Baskaya, J. Kekäläinen, and K. Järvelin. A tool for ontology-editing and ontology-based information exploration. In Kamps et al. [9].

[5] S. Bowers, H. Cao, M. Schildhauer, M. Jones, B. Leinfelder, and M. O'Brien. A semantic annotation framework for retrieving and analyzing observational datasets. In Kamps et al. [9].

[6] A. de Vries, W. Alink, and R. Cornacchia. Search by strategy. In Kamps et al. [9].

[7] B. Fortuna, D. Mladenić, and M. Grobelnik. Application of semantic annotations to predicting users' demographics. In Kamps et al. [9].

[8] F. Gey, N. Kando, and R. Larson. The crucial role of semantic discovery and markup in geo-temporal search. In Kamps et al. [9].

[9] J. Kamps, J. Karlgren, and R. Schenkel, editors. *Proceedings of the Third Workshop on Exploiting Semantic Annotations for IR, ESAIR 2010, Toronto, Canada, October 30, 2010*, 2010. ACM Press.

[10] N. Lagos, S. Castellani, and A. Kaplan. Semantic annotations for digital investigations. In Kamps et al. [9].

[11] M. Marrero, J. Urbano, J. Morato, and S. Sánchez-Cuadrado. On the definition of patterns for semantic annotation. In Kamps et al. [9].

[12] B. Marshall. Modeling betweenness for question answering. In Kamps et al. [9].

[13] V. Palacios, J. Lloréns, S. Sánchez-Cuadrado, and M. Marrero. Tagging for improved semantic interpretation of xml documents. In Kamps et al. [9].

[14] A. Said and E. W. D. Luca. Exploiting hierarchical tags for context-awareness. In Kamps et al. [9].

[15] K. Shiells, O. Alonso, and H. J. Lee. Generating document summaries from user annotations. In Kamps et al. [9].

[16] W. Tichy, S. Koerner, and M. Landhäußer. Creating software models with semantic annotation. In Kamps et al. [9].

[17] T. Trippel, C. Kirchhof, S. Awad, P. Dunkhorst, and M. Bohnes. A linguistic approach to structured search in multimodal data. In Kamps et al. [9].

[18] S. Velupillai. Semantic annotations in clinical documentation – exploring potentials for future information retrieval. In Kamps et al. [9].