SemSearch'11 - the 4th Semantic Search Workshop

[Workshop Summary]

Marko Grobelnik
Jozef Stefan Institute
Ljubljana
Slovenia
marko.grobelnik@ijs.si

Peter Mika
Yahoo! Research
Barcelona
Spain
pmika@yahoo-inc.com

Thanh Tran
Institute AIFB, KIT
Karlsruhe
Germany
ducthanh.tran@kit.edu

Haofen Wang
Apex Lab
Shanghai Jiao Tong University
China
whfcarter@apex.sjtu.edu.cn

ABSTRACT

The use of semantics and semantic technologies for search and retrieval has attracted interests both from academia and industry in recent years. What is now commonly known as Semantic Search is in fact a broad field encompassing ideas and concepts from different areas, including Information Retrieval, Semantic Web and database. This is the fourth edition of the Semantic Search workshop which aims to bring together researchers and practitioners from various communities, to provide a forum for dissemination, discussion, and for the exchange and transfer of knowledge related to the use of semantics for search and retrieval. This year's workshop will continue to push and promote efforts towards an evaluation benchmark for Semantic Search systems.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Information Storage, Information Search and Retrieval

General Terms

Theory, Algorithms, Design, Experimentation, Performance, Human Factors

Keywords

Information retrieval, document retrieval, data retrieval, semantic search

1. SUMMARY

In recent years we have witnessed substantial exploitation of search technologies, both at web and enterprise scale. However, the representation of user queries and information in existing search appliances is still almost exclusively achieved by simple syntax-based descriptions (i.e. keyword queries matched against bag-of-words document representation). While these systems have shown to work well for many common search needs, they work on the basis of rough approximations and usually fail to address more complex tasks such as aggregation and information analytics.

On the other hand, recent advances in the field of semantic technologies have resulted in tools and standards that allow for the articulation of domain knowledge at a high level of expressivity. Semantic repositories and reasoning engines have now advanced to a state where querying and processing of this knowledge can scale to large-scale scenarios. As such, semantic technologies are posed to provide significant contributions to IR problems. More expressive descriptions of resources are achieved through the representation of the resource content in terms of concepts and structured data (OWL, RDF). The recent media interest around Wolfram Alpha, PowerSet (acquired by Microsoft Bing) and Yahoo SearchMonkey show the expectations regarding the impact of semantic search.

The other way around, we have also seen the successful adoption of ideas from IR to the problem of search in semantic (Web) data, which is due to the increasing size of the Semantic Web. Popular examples include the Linking Open Data project, the large body of data in forms of Microformats and RDFa data associated with text. Common to these scenarios is that the search is focused not on a document collection, but on semantic data (which may be possibly linked to or embedded in textual information). Search and ranking large amount of semantic data on the Web is another key topic addressed by this workshop.

The WWW'11 conference is the best place for this workshop as it effectively links together three of the key topics of the conference program: Search, Semantic Web, and Bridging Structured and Unstructured Data. Our workshop also builds on the successes of our first and second workshop on semantic search. The first workshop was among the biggest ones at ESWC'08 (i.e., 50 participants) and has attracted the highest number of submissions (i.e., 22 submissions with 11 accepted papers, 50% acceptance rate). The scale of our second workshop on "semantic search" at WWW'09 grows in terms of the number of participants (i.e., more than 80 persons attended the workshop) and more submissions with lower acceptance rate (i.e., 28 submissions with 6 regular papers and 8 posters). The WWW'10 workshop was even more successful with 10 percent more in paper submissions. After

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

the LOD workshop, it was the one with the highest number of participants. Our current workshop proposal reflects the continuing maturity of this field by putting a greater emphasis on semantic search evaluation. We will provide a benchmark for participants to evaluation their semantic search solutions.

2. CHALLENGES

In this context, challenges for Semantic Search research will include, among others:

- How can semantic technologies be applied to the IR problems?
- How to address scalability and effectiveness of data Web search (by applying IR technologies)?
- How to allow web user to exploit the expressiveness of the semantic data on the Web? I.e. how to lower the technical barriers for users to ask complex questions and to interact with web data to obtain concrete answers for complex needs?
- And most importantly, how can this new generation of search systems that successfully exploit semantics for IR or for data Web search can be evaluated and compared (with standard IR systems or semantic repositories)?

3. TOPICS

Semantic Search is defined through two main directions. First is Semantic-driven IR, the application of semantic technologies to the IR problem. The second is Semantic Data Search, which mainly deals with the retrieval of semantic data. Main topics of interest for the envisioned workshop contributions include (but are not limited to) the following:

Semantic-driven IR [1, 2, 3]

- Expressive Document Models
- Knowledge Extraction for Building Expressive Document Representation
- Matching and Ranking based on Expressive Document Representation
- Infrastructure for Semantic-driven IR

Semantic Data Search [4, 5, 6]

- Crawling, Storage and Indexing of Semantic Data
- Semantic Data Search and Ranking
- Data Web Search: Search in Multi-Data-Source, Multi-Repository Scenarios
- Dealing with Vague, Incomplete and Dirty Semantic Data
- Infrastructure for Searching Semantic Data on the Web

Interaction Paradigms for Semantic Search [7, 8, 9]

- Natural Language Interfaces
- Keyword-based Query Interfaces
- Hybrid Query Interfaces (A Combination of NL, Keywords, Forms, Facets, and Formal Queries)
- Visualization of Semantic Data / Expressive Document Representation on the Web

Evaluation of Semantic Search [10]

- Evaluation Methodologies for Semantic Search
- Standard Datasets and Benchmarks for Semantic Search
- Infrastructure for Semantic Search Evaluation

4. REFERENCES

[1] J. Chu-Carroll, J. M. Prager, K. Czuba, D. A. Ferrucci, and P. A. Duboué, "Semantic search via xml fragments: a high-precision approach to ir," in *SIGIR*, 2006, pp. 445–452.

[2] P. Castells, M. Fernández, and D. Vallet, "An adaptation of the vector-space model for ontology-based information retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 2, pp. 261–272, 2007.

[3] D. Trieschnigg, W. Kraaij, and M. J. Schuemie, "Concept based document retrieval for genomics literature," in *TREC*, 2006.

[4] T. Tran, H. Wang, and P. Haase, "Hermes: Data web search on a pay-as-you-go integration infrastructure," *J. Web Sem.*, vol. 7, no. 3, pp. 189–203, 2009.

[5] H. Wang, Q. Liu, T. Penin, L. Fu, L. Zhang, T. Tran, Y. Yu, and Y. Pan, "Semplore: A scalable ir approach to search the web of data," *J. Web Sem.*, vol. 7, no. 3, pp. 177–188, 2009.

[6] G. Cheng and Y. Qu, "Searching linked objects with falcons: Approach, implementation and evaluation," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 49–70, 2009.

[7] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer, "Towards portable natural language interfaces to knowledge bases - the case of the orakel system," *Data Knowl. Eng.*, vol. 65, no. 2, pp. 325–354, 2008.

[8] T. Tran, H. Wang, S. Rudolph, and P. Cimiano, "Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data," in *ICDE*, 2009, pp. 405–416.

[9] X. Zhang, G. Cheng, and Y. Qu, "Ontology summarization based on rdf sentence graph," in *WWW*, 2007, pp. 707–716.

[10] J. Pound, P. Mika, and H. Zaragoza, "Ad-hoc object retrieval in the web of data," in *WWW*, 2010, pp. 771–780.