

Answering Web Questions Using Structured Data – Dream or Reality?

Panel Discussion

Fernando Pereira, Anand Rajaraman, Sunita Sarawagi, William Tunstall-Pedoe, Gerhard Weikum, Alon Halevy (moderator)

The question of which role structured data can play in Web search has been raised from the early days of the Web. On the one hand, structured data can be used to answer factual queries. On the other, large amounts of structured data can be used to better organize web-content and therefore to improve search on a wide range of queries.

While the Information-Retrieval approach to web search has been the clear winner to date, in recent years there has been renewed activity in trying to leverage structured data in web search. The efforts encompassed both research and commercial products, such as [1, 2, 3, 4, 5, 6, 7, 8, 9]. Perhaps the key aspects of these new efforts is that they consider breadth of structured data, rather than individual verticals. This panel discussion will consider some of the fundamental questions regarding the role of structured data in search and will examine the significance of recent advances. In particular, some of the questions we will consider are the following:

- Is structured data at all even necessary or relevant in the context of web search? Search engines have thus far fared admirably well without doing much in the way of structured data. If so, why is it relevant now when it hasn't been relevant to date?
- Is it even feasible to harness structured data in a broad horizontal fashion? Are there applications apart from standard query answering? Most successful efforts to date around structured data have been in verticals with highly constrained query types.
- Exactly what do we mean when we say structured data anyway? How does this relate to the Semantic Web and the Deep Web?
- What techniques are used to build the large ontologies that underly current systems, and what evidence do we have that they work? Why do they work? Why were previous attempts to build monolithic ontologies less successful?
- How do these systems collect the data for their knowledge bases? Is it important to make sure the data is high-quality and/or curated? What role does inference play in such systems? Can we combine evidence from multiple sources on the Web to give higher-confidence answers to queries? What level of quality does the knowledge base need to adhere to in order to offer useful search? Has information extraction had any success here or is it all manual work?
- How can structured data best be used to improve the results of search queries? In particular, would you integrate techniques for answering factual queries with a traditional search engine? Are there enough factual queries that current search engines don't answer well?
- What systems issues do we need to consider? How do you index such a large and varied collection of structured data? How do you process queries efficiently? Do you need offline reasoning/view computation?
- Why couldn't we build these systems 15 years ago when the idea was first proposed? Was there some critical technology that became mature? Is it all because we now have Wikipedia? What do we need to really make structured data a success on the Web?

1. REFERENCES

- [1] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. WebTables: Exploring the Power of Tables on the Web. In *VLDB*, 2008.
- [2] R. Gupta and S. Sarawagi. Answering table augmentation queries from unstructured lists on the web. In *Proc. of VLDB*, 2009.
- [3] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. Naga: Searching and ranking knowledge. In *ICDE*, pages 953–962, 2008.
- [4] Kosmix, inc. <http://www.kosmix.com/>.
- [5] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep-web crawl. In *Proc. of VLDB*, pages 1241–1252, 2008.
- [6] A. Rajaraman. Kosmix: Exploring the deep web using taxonomies and categorization. *IEEE Data Eng. Bull.*, pages 42–48, 2009.
- [7] True knowledge. <http://www.trueknowledge.com/>.
- [8] D. S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, and M. Skinner. Intelligence in wikipedia. In *AAAI*, pages 1609–1614, 2008.
- [9] Wolfram alpha. <http://www.wolframalpha.com/>.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the publisher, ACM.

VLDB '09, August 24–28, 2009, Lyon, France
Copyright 2009 VLDB Endowment, ACM 000-0-00000-000-0/00/00.