Database Researchers: Plumbers or Thinkers?

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Categories and Subject Descriptors
H.1.0 [**Information Systems**]: Models and Principles—*General*

Keywords
Knowledge Management, Information Extraction, Machine Reading, Disambiguation, AI Applications, Scalability, Robustness

ABSTRACT
DB researchers have traditionally focused on engine-centered issues such as indexing, query processing, and transactions. Data mining has broadened the community's viewpoint towards algorithmic and statistical issues. However, DB research has always had a tendency to shy away from seemingly elusive long-term challenges with AI flavor. On the other hand, the current explosion of digital content in enterprises and the Internet, is mostly caused by user-created information like text, tags, photos, videos, and not by seeing more well-designed databases of the traditional kind.

In this situation, I question the traditional skepticism of DB researchers towards "AI-complete" problems and the DB community's reluctance to embark on seemingly non-DB-ish grand challenges. Big questions that I see as great opportunities also for DB research include: 1) automatic extraction of relational facts from natural-language text and multimodal contexts [4, 6, 21], 2) automatic disambiguation of named-entity mentions and general phrases in text and speech [10, 11], 3) large-scale gathering of factual-knowledge candidates and their reconciliation into comprehensive knowledge bases [1, 2, 8, 13, 19], 4) reasoning on uncertain hypotheses, for knowledge discovery and semantic search [9, 14, 16, 17, 20], 5) deep and real-time question answering, e.g., to enable computers to win quiz game shows [7], 6) machine-reading of scientific publications and fictional literature, to enable corpus-wide analyses and enable researchers in science and humanities to develop hypotheses and quickly focus on the most relevant issues [3, 5].

I believe that successfully tackling these topics requires efficient data-centric algorithms, scalable methods and architectures, and system-level thinking - virtues that are richly available in the DB research community. Moreover, I would encourage our community to look across the fence and get more engaged on the exciting challenges outside the traditionally narrow boundaries of the DB realm. I will illustrate these points by examples from my own research on knowledge management [12, 15, 18, 19]. Breakthroughs will require long-term stamina. In the meantime, steady incremental progress is better than not embarking on these important problems at all.

EDBT 2011, March 22–24, 2011, Uppsala, Sweden.

1. REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary G. Ives. DBpedia: A Nucleus for a Web of Open Data. *ISWC*, 2007.

[2] Michael J. Cafarella. Extracting and Querying a Comprehensive Web Database. *CIDR*, 2009.

[3] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell. Toward an Architecture for Never-Ending Language Learning. *AAAI*, 2010.

[4] AnHai Doan, Luis Gravano, Raghu Ramakrishnan, Shivakumar Vaithyanathan (Editors). Special Issue on Managing Information Extraction. *ACM Sigmod Record* 37(4), 2008.

[5] Oren Etzioni, Michele Banko, Michael J. Cafarella. Machine Reading. *AAAI*, 2006.

[6] Oren Etzioni, Michele Banko, Stephen Soderland, Daniel S. Weld. Open information extraction from the web. *Commun. ACM* 51(12): 68-74, 2008.

[7] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, Chris Welty. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3): 59-79, 2010.

[8] Eduard H. Hovy. Turning the Web into a Database: Extracting Data and Structure. *NLDB*, 2009.

[9] Gjergji Kasneci, Jurgen Van Gael, Ralf Herbrich, Thore Graepel. Bayesian Knowledge Corroboration with Logical Rules and User Feedback. *ECML/PKDD*, 2010

[10] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text. *KDD*, 2009.

[11] Diane McCarthy. Word Sense Disambiguation: An Overview. *Language and Linguistics compass* 3(2): 537-558, Wiley, 2009.

[12] Ndapandula Nakashole, Martin Theobald, Gerhard Weikum. Scalable Knowledge Harvesting with High Precision and High Recall. *WSDM*, 2011.

[13] Fabian Suchanek, Gjergji Kasneci, Gerhard Weikum. Yago: a core of semantic knowledge. *WWW*, 2007.

[14] Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum. SOFIE: a Self-Organizing Framework for Information Extraction. *WWW*, 2009.

[15] Bilyana Taneva, Mouna Kacimi, Gerhard Weikum. Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity. *WSDM*, 2010.

[16] William Tunstall-Pedoe. True Knowledge: Open-Domain Question Answering Using Structured Knowledge and Inference. *AI Magazine* 31(3): 80-92, 2010.

[17] Gerhard Weikum, Gjergji Kasneci, Maya Ramanath, Fabian M. Suchanek. Database and information-retrieval methods for knowledge discovery. *Commun. ACM* 52(4): 56-64, 2009.

[18] Gerhard Weikum. Search for Knowledge. In: Stefano Ceri, Marco Brambilla (Editors). *Search Computing Challenges and Directions*. Springer, 2010.

[19] Gerhard Weikum, Martin Theobald. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. *PODS*, 2010.

[20] Michael L. Wick, Andrew McCallum, Gerome Miklau. Scalable Probabilistic Databases with Factor Graphs and MCMC. *PVLDB*, 2010.

[21] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, Ji-Rong Wen. StatSnowball: a Statistical Approach to Extracting Entity Relationships. *WWW*, 2009.