DATA DISCOVERY

Gerhard Weikum

Department 5, Databases and Information Systems, Max Planck Institute for Computer Science, Campus E1.4, 66123 Saarbrücken, Germany
Email: weikum@mpi-inf.mpg.de

1 STATE OF THE ART

Discovery of documents, data sources, facts, and opinions is at the very heart of digital information and knowledge services. Being able to search, discover, compile, and analyse relevant information for a user's specific tasks is of utmost importance in science (e.g., computational life sciences, digital humanities, etc.), business (e.g., market and media analytics, customer relationship management, etc.) , and society at large (e.g., consumer information, traffic logistics, health discussions, etc.).

Traditionally, information discovery is based on search engines, which in turn have mostly focused on finding documents of various kinds: publications, Web pages, news articles, etc. Search engine technology has been developed for both Internet/Web search and, with somewhat different requirements, enterprise search within companies and intranets of organisations. Digital library services are an example of the latter. A few commercial stakeholders such as Google and Microsoft have dominated Internet search; they provide excellent support for simple queries to satisfy popular information needs by typical users as opposed to expert-level needs by advanced users. In recent years, the spectrum of search scopes and items to be discovered has been expanded by multimedia data, such as photos, videos, and music, and by social media, such as blogs, tweets, and online forums.

The available services for information discovery are predominantly organized into vertical search facilities, with specialized focus on particular types of data, e.g., news vs. tweets vs. videos. Search engines do provide unified interfaces that have an integrated look-and-feel, but the underlying technologies for indexing and retrieving data items are often customized to particular classes of items. For example, digital library services are typically well geared to help users discover publications (both contents and metadata), but finding relevant information in discussions of online communities is usually provided by a separate system—if at all. Moreover, searchable content is represented mostly in terms of keywords, text phrases, and simple metadata (e.g., creation date, author, bibliographic attributes); there is no deeper understanding of the entities and semantic relationships that documents and data sources refer to. Thus, for discovering interesting connections across documents and data sources and for aggregated analyses over entire collections, users still need to invest substantial time and care to manually connect relevant pieces of information. For example, search engines do not understand that a query for "Merkel Sarkozy Euro Athens" is really a task of discovering and analysing positions of the German and French governments, as semantic entities, regarding the financial crisis in Greece. This lack of semantic representation and entity-level search is a major limitation of today's search technology.

This observation holds regardless of the fact that today's search engines already support a variety of data formats, including data models that allow rich annotations, namely, XML, RDF, and HTML5 (with embedded XML and RDFa data). In enterprise search, XML tags can be used for effectively retrieving semantic resources, but there is hardly any support for semantic XML in Internet/Web search. RDF data is prevalent in the Web of linked data—based on the W3C initiative for Linking Open Data (LOD)—and there are SPARQL query engines for individual LOD sources. However, there is very little support for search across sources, and there is no support for informatively ranking search results in order to discover semantically relevant data. Last but not least, none of the existing SPARQL engines allows combined search over both text and RDF triples.

2 TEN YEAR VISION

By the year 2020, we expect a quantum leap in terms of richer support for capturing the information needs of advanced users in a semantic representation, referring to entities and their relationships rather than keywords and pages. This will allow scientists, business analysts, journalists, and other knowledge workers to perform better discoveries of relevant data and deeper analytics of aggregated data, with much lower manual efforts than

today. Search, discovery, and analysis of information and knowledge will be unified across different data models (e.g., text and HTML vs. RDF) and will be uniformly and easily applicable to publications, datasets, software repositories, multimedia information, and social media alike.

The following example scenarios illustrate the expected form of semantic search and knowledge discovery. For further use-case scenarios, see, for example, Freire et al. 2011, Kersten et al. 2011, Mass et al. 2011, Michel et al. 2011).

Sciences: Consider a biomedical researcher who is looking into genetic and environmental factors in developing various forms of asthma. This requires identifying relevant publications, but these would be too plentiful to actually read. Instead, the scientist would prefer to obtain an aggregated picture of the risk factors themselves and how they relate to the different forms of asthma. This could be largely automated by identifying the relevant factors in the underlying texts and grouping them in an informative way—for example, by discovering that the gene GSTM1 plays a major role, inferring that this gene encodes the enzyme "glutathione S-transferase mu 1", finding empirical studies about this and other enzymes for different ethnic or socio-cultural groups, and identifying external risk factors, such as pollution, stress, exposure to endotoxins, nutritional habits, etc. All this will be feasible with tools that can detect entity names, correctly map them to entities in a knowledge base, and compute relations between entities. In addition, the underlying system should help with discovering relevant datasets, experimental data at the molecular or phenotype level, and so on. As scientists often have ad hoc needs for such complex knowledge acquisition tasks, the necessary data discovery should be performed online.

Humanities: Consider a researcher who is analyzing the "brain drain" phenomenon: the emigration of scientists from Europe to America in the 20th century (especially the 1930s and 1940s). This will require tracking people in a large collection of textual sources, with entity-name disambiguation, gathering biographical and background information, and organizing all this into a largely structured representation with rich connections to the underlying texts. This will enormously benefit the researchers by allowing deeper analysis of historical patterns and causes (e.g., dependent on the country/region within Europe and other specific conditions).

Business and Media Analysts: Consider an analyst who is interested in the European market for tablet PCs and its development over the last 10 years. This is an example of a general pattern in which analysts want to see, compare, and understand the behaviors of and trends about entities, such as companies, products, politicians, entertainers, and movies. The results of a search for "tablet PC" would ideally obtain information grouped by named entities (e.g., Apple Corp., Microsoft Corp.) combined with the time dimension (e.g., year). Likewise, the matches for "tablet PC" would ideally be based on product names rather than the literal text to capture also related products such as e-book readers.

Citizens: On the broader scale of society, people will obtain much more precise and concise answers to knowledge-centric questions that arise in everyday life and from people's interests in entertainment, sports, consumer products, vacation traveling, etc. Which Oscar winners are from Europe? Who covered the Leonard Cohen song "Hallelujah"? What is the highest number of goals that any Italian football player has ever scored in a single match for the national team? Against which German soccer clubs did FC Barcelona play in one of the European leagues or cups?

Finally and most importantly, **health** is a topic of great importance in society. Information about diseases and pharmaceutical drugs becomes increasingly complex while more and more (often elderly) people depend on this information. Examples of typical search requests are the following. Children can safely use which drugs against flu symptoms or flu viruses? Which of them can be taken while being pregnant? Could the H1N1 (swine flu) vaccine, Pandemrix, interfere with blood-pressure medication such as Metolazone? Answers will be drawn from knowledge bases but will also tap into online portals about health topics in which citizens discuss their personal experience.

All this will happen with a high degree of **personalization and contextualization**. For example, information that the user already knows would not be highlighted or shown at all. A biologist who already knows about the GSTM1 gene would be directly guided to the environmental factors. A medical researcher from Italy specialized in clinical studies would be guided to empirical data from Mediterranean hospitals.

Advanced information discovery and analysis of the links outlined above will be powered by a variety of **knowledge bases** and semantically **linked data** sources. This requires that search be treated as a computational

service so that questions can be answered by composing and reconciling query results from different sources in a federated way.

3 CURRENT CHALLENGES

There are big gaps between the current state of the art and the outlined vision for the year 2020.

The most important limitation is the lack of semantically understanding both contents and user questions (or more generally, users' information needs). This calls for moving up in the value chain: from keywords and pages to entities and relationships, from ten blue links to informative answers, or, in short, from information to knowledge.

This major leap in the semantic value chain needs to be accompanied by a more unified search over the full spectrum of relevant contents, as opposed to today's situation with compartmentalized engines. In particular, search over data and text, over Web and social media, and over Deep Web services needs to be addressed. This calls for new approaches to federated query architectures and service composition. In addition, better integration of search and analytics is needed, for example, when analyzing entire corpora, tracking entities over time, and for many forms of business intelligence.

As the available information keeps growing along with the need for search to tap into a wider spectrum of sources, the issue of data quality and trustworthy answers will become even more pressing than it already is today. For discovering truthful facts, users need explanations of their provenance and the authority and authenticity of the underlying sources. This is a major challenge. It calls for new models of trustworthiness and truthfulness and also new algorithms that can probe sources in order to estimate their coverage and quality.

Finally, the lack of user guidance and the mass-user nature of today's search tools is often a major impediment for data and knowledge discovery. Information needs are often not satisfied by a single query because insightful knowledge discovery typically requires interactive sessions that include search, exploration, aggregation, and other steps. On the one hand, this calls for deeper approaches to personalization and context awareness, including spatio-temporal and socio-cultural background as well as individual histories of user behavior. On the other hand, it calls for cognitive guidance in knowledge discovery tasks, and also for better user interfaces suitable for smartphones and tablet computers (including e-book readers).

4 RESEARCH DIRECTIONS PROPOSED

To address the above challenges, we propose the following research directions.

Search for knowledge (Weikum 2009, Pereira et al. 2009, Weikum et al. 2009, Weikum and Theobald 2010): The ultimate goal of search is to return informative pieces of knowledge, not just raw information like ten blue links to Web pages. This goal entails organizing search in a fundamentally different manner so that it 1) can understand entities and relationships in users' questions and information-discovery tasks and 2) can return precise and concise results that are informative to the user. In addition, the underlying system should guide the user in querying, exploring, and discovering relevant data. This requires extensive research at all levels of a search engine: data models, statistical ranking, efficient algorithms, and interaction with users.

Search as a service (Ioannidis 2007, Ceri and Brambilla 2010): Search is a building block for achieving the overriding goal of satisfying users' information and knowledge needs. Often, several search engines for different kinds of data need to interoperate for this purpose. It should be possible to implement intelligent agent software for user guidance, recommendations, or analytic tasks on top of such an ecosystem of computational services. This requires that services be composable and provide carefully designed APIs. The languages for these APIs should be able to cope with different data representations in a unified manner (e.g., RDF data plus text plus images).

Personalization (Koutrika and Ioannidis 2010, Mokbel 2011): The relevance of information is in the eye of the beholder. Different users should be provided with different answers to the same question, depending on their

specific interests, prior knowledge, and current situation. This calls for deeper notions of personalized search and discovery, taking into consideration long-term and short-term user behavior, cognitive models for user interactions, and also awareness of the user's social context.

Space, time, and cultural context (Baeza-Yates et al. 2011, Weikum et al. 2011, Hoffart et al. 2013): These are key dimensions in understanding a user's situational context. Space and time can refer not only to the current user position but also to the interest expressed in a user's search requests. Translational and longitudinal analyses on news and Web archives require tracking entities over spatio-temporal contexts. Likewise, the cultural background of both the user and the potentially returned contents needs to be considered for true context awareness. Children should not get the same answers as scientists, and suitable advice on nutrition, hygiene, and health related issues crucially depend on socio-cultural habits.

Search and analysis of social media (Amer-Yahia 2009, Baeza-Yates et al. 2010): The latter points become most prominent in social media, such as blogs, wikis, and online forums. As they echo events in primary media, such as news, but add the people's views, social media are an invaluable asset for opinion analysis on political, health, and other societal discussions. This requires major enhancements to search technology with better integration of querying, information extraction, and sentiment mining.

Structured data on the Web (Cafarella et al. 2011, Heath and Bizer 2011, Ooi 2010): The Web of Data (the "LOD cloud") already comprises more than 20 billion RDF triples from hundreds of sources and keeps growing. Moreover, it provides extensive cross-linkage among semantic sources from a variety of domains including biology, music, movies, maps, and universal knowledge bases. Effectively querying this wealth of data over widely distributed and highly heterogeneous sources is an important and difficult problem. Moreover, for emerging applications and the possibility of tapping into embedded RDFa in HTML5, it is crucial to extend search capabilities to combinations of data and text.

Data quality and trust (Yin et al. 2009, Dong et al. 2010, Doan et al. 2011): Information sources vary highly in their quality, i.e., authority, authenticity, freshness, degree of curation, etc. Thus, it is crucial to differentiate trustworthy sources from lower-quality ones and to explain the provenance of how the results of queries and discovery tasks are derived.

User interfaces (Baeza-Yates et al. 2010, Ferrucci et al. 2010): Last but not least, although APIs are crucial for layering application programs on top of search services, applications ultimately need to communicate with human users. This calls for new kinds of user interfaces (UIs) with much better cognitive models to guide users towards answers for their information needs. These UIs will need to be integrated with the next generation of smartphones. Thus, speech-based search and natural-language question answering will have to be greatly improved, and visual output and interaction metaphors should be reconsidered as well.

5 RECOMMENDATIONS

In summary, on the basis of the above arguments, we recommend encouraging and supporting research along the following three main lines:

- Search for knowledge: Develop models, methods, and tools for entity-relationship-oriented semantic search across the entire spectrum of data representations. This should support and unify information and knowledge discovery in structured Web data, rich HTML and text with embedded RDF data, social media and online communities, multimedia formats with semantic metadata and expressive object-level features as well as repositories for software, data collections, and experiments.
- Personalization and Socio-Cultural Awareness: Develop new approaches for deeper levels of personalization and context awareness so as to provide adequate support to knowledge workers with advanced information needs and limited willingness to manually inspect and tediously post-process large amounts of diverse and mixed-quality results. This research should also consider the social and cultural background of users and models of user guidance in the discovery process as well as modern user interfaces for tablets and smartphones.
- Federation of services: Develop versatile and composable building blocks for an ecosystem of search, discovery, and analytics services. Value-added services should be quickly configurable by intelligently coupling suitable services and providing integrated behavior and benefit for advanced users while drawing

on an agile marketplace of grassroots providers.

These lines of research are not fundamentally new. However, the ongoing explosion of raw information and the need for deeper forms of knowledge discovery and business intelligence suggest that the above directions should be high-priority items now. In addition, the recent advances in supporting technologies—especially the availability of large knowledge bases and other linked-data sources—have revived and intensified research in this area and have opened up new opportunities for the coming decade.

To create a strong research momentum while striving for direct and measurable impact, the political governance should emphasize and set incentives for the following.

- Services that are made publicly available as early as possible (e.g., Dbpedia/LOD-style endeavours) create an organic habitat rather than big flagship projects only (e.g., not just Quaero or Theseus style, which had limited impact relative to their size and cost).
- Healthy resource and service diversity and dynamics, avoiding quasi-monopolistic structures with market-dominating search engines or heavily centralized data/knowledge and platform providers.

6 REFERENCES

Amer-Yahia, S. (2009) Special Issue on New Avenues in Search. *IEEE Data Engineering Bulletin* 32(2).

Baeza-Yates, R.A., Masanes, J., and Spaniol, M. (2011) *The 1st Temporal Web Analytics Workshop (TWAW)*. WWW (Companion Volume), pp 307-308.

Baeza-Yates, R.A., Broder, A.Z., and Maarek, Y.S. (2010) The New Frontier of Web Search Technology: Seven Challenges. *SeCO Workshop 2010*, pp 3-9.

Cafarella, M.J., Halevy, A.Y., and Madhavan, J. (2011) Structured Data on the Web. *Communications of the ACM* 54(2), pp 72-79.

Ceri, S. and Brambilla, M. (2010) Search Computing Systems. *CAiSE 2010*, pp 1-6.

Doan, A., Ramakrishnan, R., and Halevy, A.Y. (2011) Crowdsourcing systems on the World-Wide Web. *Communications of the CM Communication* 54(4), pp. 86-96.

Dong, X., Berti-Equille, L., Hu, Y., and Srivastava, D. (2010) SOLOMON: Seeking the Truth Via Copying Detection. *PVLDB* 3(2), pp. 1617-1620.

Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., et al. (2010) Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3), pp 59-79.

Freire, J., Bonnet, P., and Shasha, D. (2011) Exploring the Coming Repositories of Reproducible Experiments: Challenges and Opportunities. *PVLDB* 4(12), pp. 1494-1497.

Heath, T. and Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool Publishers 2011.

Hoffart, J., Suchanek, F.M., Berberich, K., and Weikum, G. (2013) YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence* 194: pp. 28-61.

Ioannidis, Y.E. (2007) Emerging Open Agoras of Data and Information. *ICDE 2007*, pp 1-5.

Kersten, M.L., Idreos, S., Manegold, S., and Liarou, E. (2011) The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds. *PVLDB* 4(12), 1474-1477.

Koutrika, G. and Ioannidis, Y.E. (2010) Personalizing Queries based on Networks of Composite Preferences. *ACM Transactions on Database Systems 35*(2).

Mass, Y., Ramanath, M., Sagiv, Y., and Weikum, G. (2011) IQ: The Case for Iterative Querying for Knowledge. *CIDR 2011*, pp 38-44.

Michel, J.-B., Kui-Shen, Y., Presser-Aiden, A., Veres, A., Gray, M.K., The Google Books Team, et al (2011), Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331 (6014), pp176-182.

Mokbel, M.F. (2011) Special Issue on Personalized Data Management. *IEEE Data Engineering Bulletin 34*(2).

Ooi, B.C. (2010) Special Issue on Keyword Search. *IEEE Data Engineering Bulletin 33*(1).

Pereira, F., Rajaraman, A., Sarawagi, S., Tunstall-Pedoe, W., Weikum, G., and Halevy, A.Y. (2009) Answering Web Questions Using Structured Data - Dream or Reality? *PVLDB* 2(2), p 1646.

Weikum, G., Kasneci, G., Ramanath, M., and Suchanek, F.M. (2009) Database and information-retrieval methods for knowledge discovery. *Communications of the ACM* 52(4), pp 56-64.

Weikum, G. (2009) Search for Knowledge. *SeCO Workshop 2009*, pp 24-39.

Weikum, G. and Theobald, M. (2010) From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. *PODS 2010*, pp 65-76.

Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczur, A.A., Kirkpatrick S., et al. (2011) Longitudinal Analytics on Web Archive Data: It's About Time! *CIDR 2011*, pp 199-202.

Yin, X., Han, J., and Yu, P.S (2009) Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transaction on Knowledge and Data Engineering* 20(6), pp 796-808.

(Article history: Available online 1 July 2013)