

Making things findable: Semantics for Web Search and Online Media

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

The publishing of original content on the Web and later, Web search have developed into two of the most profitable sources of business on the Web, supported by revenue from display advertizing and search advertizing, respectively. Despite a traditional separation, what we can observe on the Web today is a gradual blurring of the lines between these two businesses as search engines gradually integrate more and more structured content, and online publishers turn to methods of search for customizing user experiences. This brings both opportunities and a new set of challenges in applying semantic technologies.

Web search engines today are going well beyond providing a list of "ten blue links" in response to user queries. Enhanced Results in Yahoo! Search are an example of how a search engine can exploit data from webpages to create richer search results with structured data (such as product ratings and prices), images and video. Another example is faceted exploration [4], which appears on Yahoo! Search for a number query types, including names of celebrities and world cities. The suggestions appear on the left bar of the search result page, and they consist of entities related to the user's original request, such as related actors or sights within a city. Effectively, this allows the user to switch to exploration mode, and launch new queries for related interests. A third example is the increasing usage of information boxes that appear at various positions on the search result page, and give relevant data from vertical content sources. The content and data used in the above services traditionally comes from professional, paid sources, but search engines are also increasingly exploiting methods of Information Extraction. This helps in covering the long tail of content (e.g. niche products), to cover domains where professional data is simply not available (e.g. high school listings) or to fill gaps in acquired data (e.g. restaurant menu items).

At the same time, online publishing is becoming increasingly sophisticated in targeting content, which is essentially the retrieval problem of surfacing the right content for the right users at the right time. This way, online publishers such as Yahoo! are increasingly exploit ranking methods

similar to those applied in web search, with the difference that the users interest in mostly implicit. An example of this is the Content Optimization and Relevance Engine (CORE, formerly COKE), which is used by Yahoo! on various sites (most prominently, on the Yahoo! homepage) to select the right content to show for each user based on click data and profile information [1]. This machine learning based engine does not substitute human editors entirely, but complements it in a powerful way. The hyperlocal pages of Yahoo! Local are another example of automated matching of content to specific audiences, in this case selecting content (news, blogs, deals, events) that are specific to a city neighborhood. A third example is event-based targeting such as the micro-sites that Yahoo! develops for key events such as the Summer Olympics or the Oscars. User profiles, geography, and the occurrence of real world events are only three particular examples of the many dimensions that may contribute to the relevance of content to particular users. We also note that the difference to search is narrowing also in the type of content that is served. As the targeting becomes more fine-grained, Yahoo and other publishers are also increasingly relying on semi-professional or Web content from select sources to fulfill niche interests.

Where these models converge is the need to retrieve, aggregate, and format for display structured data or richly annotated content acquired from a variety of sources, ranging in scale and quality from professionally curated content from trusted sources to information culled from unknown Web pages. Semantic Search, that is search using semantic technologies, holds a particular promise in addressing this need. The Semantic Web has been designed from the onset to be a global data space where data from various sources could be integrated automatically, or at least with the help of machines based on the machine's understanding of the data. Similarly, this understanding of the data, and a semantic analysis of the user's query would in principle allow powerful inferences regarding the relevance of certain items to the user's request. Given content that is richly annotated with terms from shared ontologies, machines could also learn the user's long-term interests and current context at the semantic level, i.e. in terms of the ontological concepts that the user is interested in, making much precise recommendations.

There is progress in realizing this form of search, but not in all respects. The Resource Description Framework (RDF) does provide the right level of representation for managing information at fine granularity, i.e. at the level of individual facts. RDF as a standard is also a powerful interchange

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS '11, May 25-27, 2011 Sogndal, Norway
Copyright 2011 ACM 978-1-4503-0148-0/11/05 ...\$10.00.

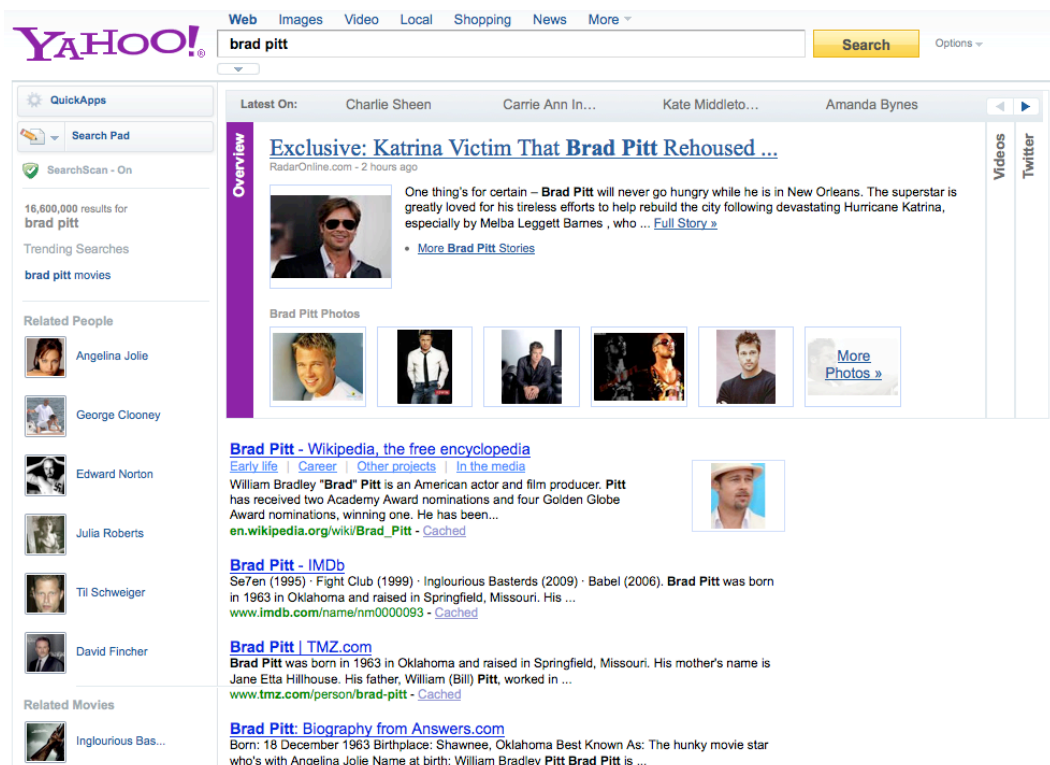


Figure 1: A screenshot of Yahoo! Search showing faceted search (left), an information box and an enhanced result.

format and as such it enables open data sharing across the Web. RDF is the backbone of Linked Data, and RDFa is used increasingly in marking up HTML pages for particular purposes such as providing data to Facebook's Open Graph API [2]. Although its adoption is more limited, SPARQL does liberate us from the tyranny of APIs by allowing anyone to submit free-form structured queries to remote data sources on the Web.

There is a yet to be fulfilled promise of pushing the boundary in a machine understanding of data, largely due to the lack of ontologies. OWL, the Semantic Web schema language is proving to be tedious to use for many, and after many years there are still no easy ways to find ready-made ontologies on the Web. Many implementers are willing to sacrifice even the basics of proper data modeling at the altar of simplicity, which we can see by looking at Linked Data usage (e.g. the misuse of owl:sameAs) or how RDFa is being used in the wild (e.g. the disappearing distinction between object- and datatype-properties in the Open Graph Protocol). Although this data first, schema second approach to building the Semantic Web may help adoption, the net effect is that the Semantic Web is being reduced to a bag of triples without much help from ontologies in interpreting this data.

There is not much to do other than embracing this new reality, and this is what we've seen in the area of Semantic Search. While early works have focused on ontology-based retrieval in narrow domains, current approaches that scale to open domains are bottom-up in learning ontologies from the data or forgoing ontologies altogether. Semantic search evaluation on web data is now focusing on the more basic

task of ranking objects [3], as opposed to generating structured interpretations of user queries for which precise answers could be given. While this brings Semantic Search closer to IR than Logic or Databases, it is still a substantially different and challenging problem, and even relatively modest advances over text retrieval can bring a potentially large pay-off in both web search and online publishing.

1. REFERENCES

- [1] D. Agarwal, B. C. Chen, P. Elango, N. Motgi, S. T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online Models for Content Optimization. In D. Koller, D. Schuurmans, Y. Bengio, L. Bottou, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, pages 17–24. MIT Press, 2008.
- [2] P. Mika, E. Meij, and H. Zaragoza. Investigating the Semantic Gap through Query Log Analysis. In *ISWC '09*, pages 441–455, Berlin, Heidelberg, 2009. Springer-Verlag.
- [3] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *WWW '10*, pages 771–780, New York, NY, USA, 2010. ACM.
- [4] R. van Zwol, P. Ng, A. Ramani, A. Sahai, S. T. Sathish, H. Vasudev, U. Vuyyuru, B. Sigurbjornsson, R. Adapala, L. Garcia Pueyo, A. Katiyar, K. Kurapati, M. Muralidharan, S. Muthu, and V. Murdock. *Faceted exploration of image search results*. ACM Press, New York, New York, USA, 2010.