SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/

Microsearch: An Interface for Semantic Search

Peter Mika

Yahoo! Research
Ocata 1, 08003 Barcelona, Spain
pmika@yahoo-inc.com

Abstract. In this paper we discuss the potential for semantic search and focus on the most immediate problem toward its realization: the problem of the sparsity and relatively low quality of embedded metadata. We suggest that a part of the solution is to expose users to embedded metadata as part of their daily activity of searching the Web. We present the publicly available microsearch system which enriches search result presentation with metadata extracted from search results and report on some of the early feedback we have received.

1 Introduction

The current generation of search engines is severely limited in its understanding of the user's intent and the Web's content and consequently in matching the needs for information with the vast supply of resources on the Web.

For Information Retrieval purposes, both queries and documents are typically treated at a word or gram level, with minimal language processing involved. In other words, the search engine is missing a semantic-level understanding of the query or the content: it is as if one would try to understand the content of a document by picking out the most commonly occurring or underlined words.

The fact that search is still considered as a technology that largely 'works' has to do with a number of factors. First, a number of queries are easy in the sense that they belong to the class of navigational queries, where there is a single known item sought, e.g. 'air france'. At the other end of the scale, in answering very broad queries (such as 'hotel paris') there are typically a vast array of similarly relevant documents.

Second, search engines have managed to mask their limitations by a number of techniques. Foremost, the unit of retrieval is limited to individual documents, as the statistical methods applied degrade quickly when considering smaller units such as paragraphs or sentences. Situations of ambiguity are solved by applying measures such as PageRank which automatically zoom in on the most common interpretation of a query. (For example, the query 'George Bush' returns results related to the famous politician, irrespective of the number of persons named George Bush.) Further, users are aided in refining their query, although not on the basis of an explicit understanding of a query, but on the basis of the refinements made by other users starting with the same query.

Yet there are a number of situations where one can clearly see the limits of a syntax-based approach to search. Here we list but some of the examples. Interestingly, users have adapted to the limitations of search engines to the extent that some of these queries are rarely entered anymore.

- The *ambiguous queries* mentioned above are the most straightforward examples, in that it becomes almost impossible to find an object that relates to the secondary sense of a term, in case a dominant sense exists. In the example, consider searching for George Bush, the beer brewer. Note also that in widely scoped information spaces nearly all terms are ambiguous.
- The capabilities of *computational advertising*, which is largely also an information retrieval problem (i.e. the retrieval of the matching ads from a fixed inventory), are clearly impacted because of the greater sparsity of advertizements.
- Search engines are also unable to perform *queries on descriptions of objects*, where no clear key exists. For example, one might want try to search for the author of this paper as "semantic web researcher working for yahoo". A typical, and much important example of this category is product search. For example, search engines are unable to look for music players with at least 4GB of RAM without understanding what a music player is, what it's characteristics are, etc.
- Current search technology is also unable to satisfy any complex queries requiring *information integration* such as analysis, prediction, scheduling etc. An example of such integration-based tasks is opinion mining regarding products or services. (While there have been some successes in opinion mining with pure sentiment analysis, it is often the case that one would like to know what specific aspects of a product or service are being described in positive or negative terms.) Information integration is not possible without structured representations of content.
- Lastly, *multimedia queries* are also difficult to answer as multimedia objects are typically described with only a few keywords (tagging) or sentences. This is typically too little text for the statistical methods of IR to be effective.

Clearly, these problems cannot be addressed without moving toward *semantic search*, which we define as information retrieval with the capabilities to understand the user's intent and the Web's content at a much deeper, conceptual level. We believe that building on the results from Information Retrieval and the Semantic Web, with important contributions from the field of Natural Language Processing, semantic search could become a reality in the coming years [2]. However, before we could move to consider methods for semantic search we have to face the problems related to the *sparsity and low quality of metadata* on the Semantic Web.

Even after ten years of the publishing of the first Semantic Web standards, the technology has largely failed to impact the way information is encoded on the Web. In fact, in recent years the focus has shifted from a vision of the Annotated Web that characterized early Semantic Web research to one that is

focused almost exclusively on Linked Data, i.e. on databases instead of documents. Interestingly, at the point where Semantic Web researchers have almost but given up on the idea of an annotated web, significant advances have been made in this area by the Web 2.0 movement, in particular through the introduction of microformats. Microformats lower the barrier for manually authoring metadata or implementing metadata production by simplifying the knowledge representation paradigm and reducing choice. (In particular, each microformat is a fixed vocabulary designed to describe one information type without possibilities of extension. From the user's perspective this makes it almost trivial to choose and follow a format.) Microformats have also earned the support of major participants in the Web industry with Yahoo! alone publishing over one billion microformat enabled pages. Encouraged by this development, the W3C has also moved forward rapidly with the standardization of RDFa, a format for embedding RDF into XML (including XHTML) in a similar way that microformats are encoded in HTML. Yet we can still consider metadata sparse when considering the fraction of metadata-enabled web pages.

The quality of embedded metadata is also of concern as it will have significant impact on any semantic search effort. While Linked Data is typically exposed in fully automated ways and thus it is no lower quality than the original data, manually created metadata suffers a loss of quality at the point of encoding. Unfortunately, users expect that the same way browsers tolerate errors in HTML markup, mistakes made during microformat authoring would also be easily corrected automatically by the processing agent. However, while forgetting to close an angled bracket in HTML is relatively easy to correct, incorrect microformat markup is much harder and often impossible to spot by automated means, e.g. in cases where the wrong class is applied to a particular information as a result of forgetting to close a DIV or SPAN element.[1] This situation is likely to be worsened by further complexity introduced in RDFa.

In our judgment the problems of sparsity and data quality on the Semantic Web are tied together by a common solution: bringing metadata to the surface of the Web. At the moment the Semantic Web is what many refer to as a *shadow web* where users almost never see metadata displayed in any shape or form. This means that users no see incentive to create new metadata. Just as importantly, users have no ways to correct incorrect metadata as this would require the mistakes to be visible. Last, to unleash collaborative effects it should be possible to correct erroneous metadata by any user not just the user who created and maintains the page with the incorrect metadata.

In this paper we present microsearch, a research prototype that demonstrates ways to bring metadata to the surface by incorporating it in the result display of a search engine. Microsearch also showcases some of the early benefits of

[1] In practice, auto-correction of microformat data is not even attempted: both microformat and RDFa data are typically processed by means of XSLT typically after running Tidy on the page. While Tidy corrects HTML markup it is not concerned with microformats and the XSLT stylesheets used are engineered for correct markup.

metadata-enabled search engines when it comes to information integration and spatial-temporal visualization.

2 The microsearch system

The microsearch system enriches the search experience by visualizing embedded metadata. First, for result pages that contain embedded metadata a summary of the data is presented as part of the abstract ('snippet'). Further, the user can take direct actions based on the semantics of the information, such as adding an address to his/her local address book, starting to compose an email or directly dialling a telephone number. Second, it is often possible to relate pages through metadata in which case the related pages can be visually grouped together. Figure 1 illustrates these features using the query 'ivan herman'. (Ivan Herman is W3C's Semantic Web Activity Lead.)

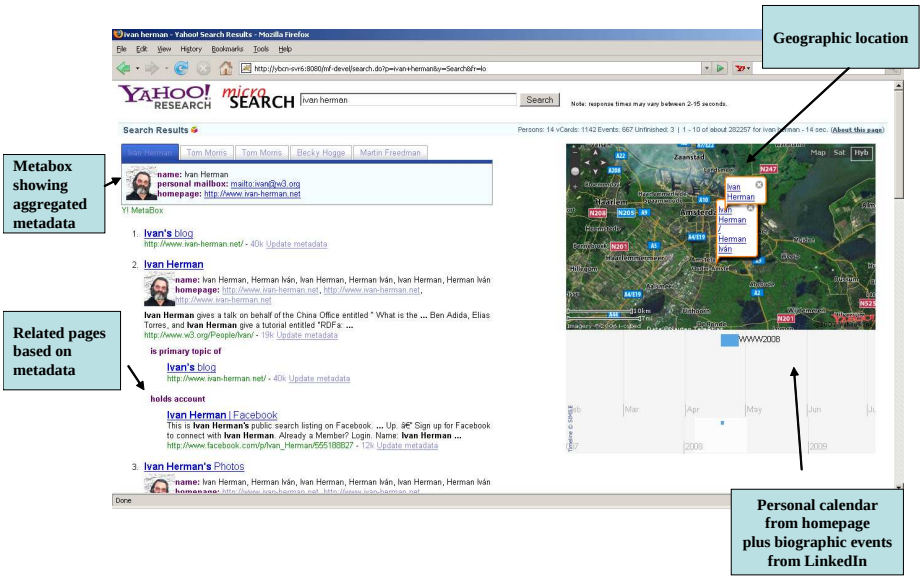

Fig. 1. Result display for the query *ivan herman*.

Microsearch also demonstrates the promise of semantic search when it comes to the aggregation of information across result pages. A Yahoo! Map shows resources which have a geographic relevance and for which a location is given (and this location can be successfully geocoded). At the moment this is limited to foaf:Person instances with geographic coordinates and vCards for persons and organizations in which case the address is geocoded using the Yahoo! Maps API itself. Figure 2 shows this feature for the query 'peter site:flickr.com', i.e. for

all the users named Peter on the Flickr web site. The map zooms and pans automatically in order to include all the nodes being visualized. Similarly, a timeline shows event information when available using the SIMILE Timeline API. The timeline can show both points in time as well as periods in time such as biographical information from profile sites such as LinkedIn. The scale of the timeline is fixed, but two bands are shown to allow scrolling by month and by year. Also, the timeline is centered on the last event displayed (which may be in the future). Figure 3 shows this feature for the query 'san francisco conference'. At the moment the map and the timeline are shown for all queries, but it would be easy to change this behaviour in a way that only relevant modules are shown.

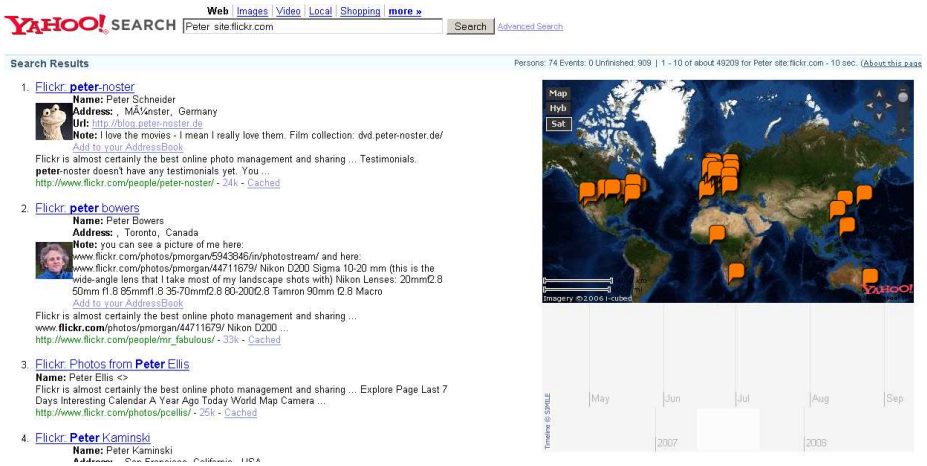

Fig. 2. Result display for the query *peter site:flickr.com*.

Figure 4 shows an overview of the architecture of the microsearch system. The dynamic behaviour of the system is as follows. On the microsearch website [2], users initiate a search the same way they would with Yahoo!'s main search engine. The query is issued against the search engine and the top results are retrieved for display. Besides retrieving regular search results, we also retrieve the top results that are known to contain certain types of microformat data. In a next step, the metadata is extracted from the displayed results and the pages that are known to microformat results. (The reason we process the display pages is that not all forms of embedded metadata are available from the search index.) After running Tidy on the pages, the extractor (known as the sponger) extracts popular microformats, linked RDF and RDFa data. (Support for GRDDL is among the future work.)

[2] `http://yr-bcn.es/demos/microsearch/`

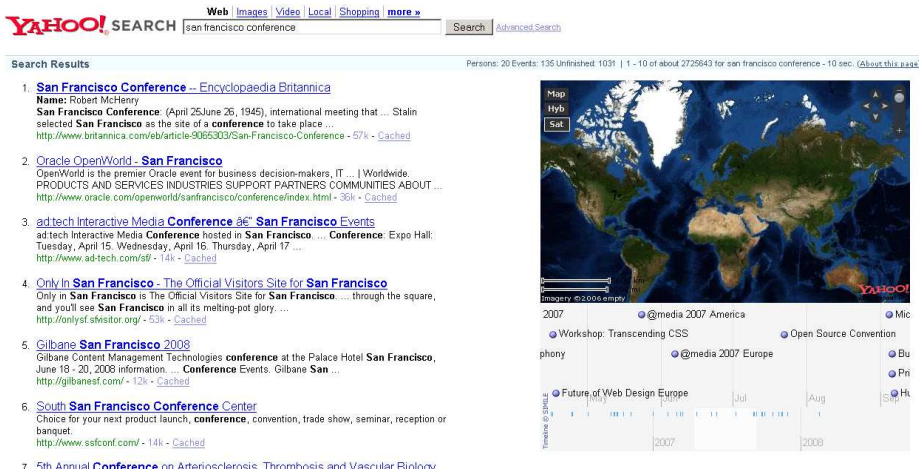

Fig. 3. Result display for the query *san francisco conference*.

Next, the metadata is aggregated and stored in a temporary Sesame[3] repository as well as cached to speed up further queries. We perform entity reconciliation on the aggregated data although this is not used in the current version of the system. Next, the result display is generated by using the Elmo API to populate a Java object model from the RDF data. The Fresnel API[4] developed by the SIMILE project is used to generate snippets from metadata. Transformations in Fresnel are described in declarative manner, providing among others what properties to display for certain classes of objects, which properties should be visualized as links or images etc. These descriptions known as Fresnel lenses are written in RDF using the Fresnel vocabulary. Using RDF provides the flexibility to create visualizations by inheriting from existing descriptions. Further, in principle the system could discover and reuse Fresnel lenses created by external developers to visualize resource types unknown to the current system. However, this possibility is not yet exploited.

3 Discussion

The microsearch demo has been made available online only recently and therefore long term statistics are not available yet. Although the prototype was not widely advertized, in the first week of its availability 7848 queries have been issued from 1037 unique IP addresses.

Figure 5 shows the distribution of unique queries according to the number of displayed results that contained metadata and thus resulted in metadata-based snippets. These statistics show that in 53.6% per cent of the unique queries at

[3] `http://www.openrdf.org`
[4] `http://simile.mit.edu/wiki/Fresnel`

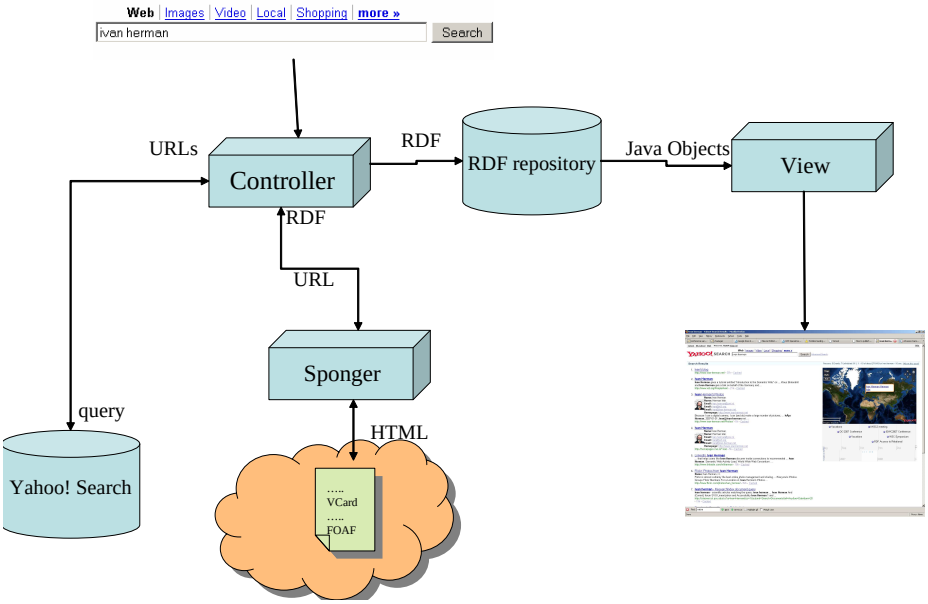

Fig. 4. The architecture of the microsearch system.

least one of the top 10 displayed results contained some metadata.(Note that the map and timeline may show metadata extracted from results below rank ten.)

The population of those who have tried the demo is hardly indicative of the general web population (mostly Semantic Web researchers and developers) and the queries issued are also a-typical (mostly person names). Thus the only observation we can make for now is that a metadata-enriched search engine can bring benefits to this particular community and the kind of queries issued, with no extra cost on the user's side. (When no metadata is present, microsearch simply behaves as the main search engine except for latency). We plan to investigate the shape of this distribution using a query log from Yahoo!'s main search engine. The advantage of using a live search engine or a query log for this analysis is that one is able to measure the metadata content of the pages that are likely to be useful for users. (While the Web is large, only a fragment of it is ever accessed through search.)

Based on the feedback we received the experience was also positive for the users with the obvious drawback of the increased query time. (However, by extracting and storing metadata as part of an offline process this delay can be significantly reduced.) Some of the expected benefits of exposing metadata were immediately visible: the present author, for example, discovered that his FOAF profile links to his old geographic address in the Netherlands. After being exposed to the interface, some users have also asked for ways in which they could metadata to their own pages. To help them, we have created a simple FAQ with short descriptions of how to add common types of metadata to HTML using

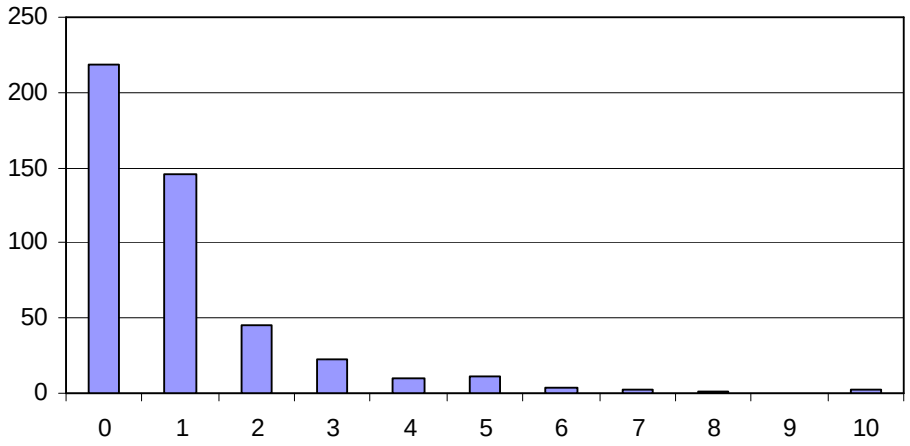

Fig. 5. Histogram showing the number of queries (y-axis) with 0, 1, ... 10 metadata-enabled pages (x-axis) within the top ten results.

microformats or RDFa. We have also included an "Update metadata" button next to each search result so that users can immediate see the results after adding or updating metadata to a particular page. Semantic Web developers have also asked for ways in which they could build other kinds of interfaces using the aggregated metadata produced, which prompted us to expose the metadata as a feed. Their reaction also confirmed our expectation that on the long run semantic search is likely to impact both query input and results presentation, reshaping the ways users interact with search engines.

Some of the ideas behind microsearch are also reflected in the design of Yahoo!'s Open Search Platform, also known as Search Monkey. Search Monkey will enable for any developer to create similar experiences in a highly scalable fashion. Search Monkey divides up the process of developing semantic search applications in two steps: metadata extraction and result presentation. (These are a single step in the microsearch process.) First, developers will have the possibility to create their own extraction modules as well as provided with metadata automatically extracted during the crawling process. The metadata resulting from running such extraction modules on webpages will be stored in the search index and made publicly available. Second, developers can also write visualization modules that create metadata-based snippets using the extracted metadata. Users of the search engine will be able to pick and choose the visualization modules they would like to use to enhance their search results.

4 Conclusions

Current methods of bringing semantics to Web search rely mostly on large editorial efforts, where web pages are classified manually or semi-automatically into

semantic classes. This method, for example, allows to display custom content on both Yahoo! and Google Search: see for example the Yahoo! Shortcut to Yahoo! News for the query 'britney spears'[5] and the similar shortcut to Yahoo! Shopping for the query 'apple ipod touch 8gb'[6]. Once the query intent is identified in terms of a taxonomy, web search engines are also able to provide much better help in breaking down the results, as shown among others by Google for the query 'ritalin'[7] and Hakia for the query 'george bush' [8].

This classification effort runs into two kinds of scaling problems when applied to Web search. First, there are a vast number of pages on the Web, which is fed by an endless production pipeline. This problem is addressed by harnessing the human effort of Web users as it has been done in Google Co-op[9] which lets users tag certain categories of Web sites (e.g. *health*) with predefined labels (e.g. *side effects*, *overdose*, *clinical trials* etc.)

However, there is another, potentially more difficult challenge related to the breadth of the information needs of Web users. The long tail of information needs is longer than most of us realize: Baeza-Yates et al. report that in the one year query log they studied 88% of the unique queries are singleton queries, and 44% are singleton queries out of the whole volume, which means that the vast majority of Web queries are only seen once, even when looking at a full year of query production [1]. This means that systems that rely on a fixed taxonomy of information needs (as all of the Web examples do) will certainly run into limitations when covering more than just the most common classes of objects and their most common aspects.

Microsearch and SearchMonkey bring semantics to long tail queries by relying on Semantic Web technology. Relying on standard semantic technology enables the system to aggregate information provided by users (manually annotating their web pages), and in the case of SearchMonkey, also information submitted to the system in the form of data feeds or extracted from Web pages. The application of semantic technology to vocabulary management (RDF, OWL) also means that the system is not limited to a fixed hierarchy of information types and a limited set of aspects when it comes to understanding query intent.

These systems in their present forms are still far away from exploiting all the possibilities offered by semantic search and tackling many of the challenges described in Section 1. However, by relying on open Semantic Web standards in metadata representation we believe that these systems have the potential to bring semantics to search in a way that scales to both the size and breadth of the Web.

[5] `http://search.yahoo.com/search?p=britney+spears`
[6] `http://search.yahoo.com/search?p=apple+ipod+touch+8gb`
[7] `http://www.google.com/search?q=ritalin`
[8] `http://www.hakia.com/search.aspx?q=george+bush`
[9] `http://www.google.com/coop/`

References

1. Ricardo Baeza-Yates, Aristides Gionis, Flavio Junqueira, Vanessa Murdock, Vassilis Plachouras, and Fabrizio Silvestri. The impact of caching on search engines. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190, New York, NY, USA, 2007. ACM.
2. V. Richard Benjamins, John Davies, Ricardo Baeza-Yates, Peter Mika, Hugo Zaragoza, Mark Greaves, Jose Manuel Gomez-Perez, Jesus Contreras, John Domingue, and Dieter Fensel. Near-term prospects for semantic technologies. *Intelligent Systems*, 23(1):76–88, 2008.