The Web as a Semantic Source

Ricardo Baeza-Yates

Yahoo! Research
Ocata 1, 08003 Barcelona, Spain
rbaeza@acm.org

Abstract

In this extended abstract we describe several semantic sources that can be found in the Web that are either explicit, e.g. Wikipedia, or implicit, e.g. derived from Web data. We also show how we are using them to improve search and to generate new semantic resources, as our final goal is to produce a virtuous feedback circuit for semantic enhancement based in machine learning.

1. Introduction

The Semantic Web dream would effectively help many applications, in particular search. However, the Semantic Web is more a social rather than a technological problem. Hence, we need to help the process of adding semantics and generating more semantic resources by using automatic techniques. This must be based in semantic sources already available in the Web as well as other sources. However, the advantage of the Web is that is several order of magnitude larger than well known sources. On the other hand, the Web is noisy and still incomplete.

We can distinguish two different types of semantic sources in the Web: explicit and implicit. Well known explicit sources are mostly based in collaborative work such as the Wikipedia. Implicit sources are raw Web content and structure as well as human interaction in the Web, what is called nowadays the Wisdom of Crowds (Surowiecki, 2004).

We first characterize and mention the main semantic sources in the Web, describing at the same time several results of Yahoo! Research to exploit and enhance these sources, as well as to improve search. We finish with some thoughts and examples that will have impact in the future.

2. Explicit Semantic Sources

Explicit sources can be subdivided further in three cases:

- Metadata, appearing in many ways and formats, such as microformats, Dublin Core, etc.
- Sources available in the early semantic Web such as RDF.
- User generated content or Web 2.0. In this case we have resources categorized using formal taxonomies such as Wikipedia or the Open Directory Project, and resources categorized and labeled using folksonomies such as Flickr.

These resources are the main baseline to evaluate semantic resources generated automatically. They are also easier to enhance. For example, Atserias et al. (2008) have shared a semantically tagged version of the Wikipedia based on results that we mention in the next section.

In the case of the Flickr folksonomy, Sigurbjornsson et al. (2008) have shown how to use collective knowledge (or the wisdom of crowds) to enhance image tags, and also they prove that almost 80% of the tags can be semantically classified by using Wordnet and Wikipedia (Overell et al., 2008). This effectively improves image search.

3. Implicit Semantic Sources

The main sources of implicit semantics are basically three:

- Natural language text: the Web comprises hundreds of terabytes of text in several languages, sometimes with parallel translations. This source has been well described in (Kilgarriff et al., 2003).
- Link structure of the Web: links and anchor text encode semantic information and due to its large number is frequently used.
- Usage data in the Web: human actions recorded in weblogs also encode semantic information and is the largest source available.

The semantics behind these sources must be extracted by different techniques, being Web data mining the main one.

In the case of text, our initial efforts to improve search are based in shallow semantics. Ciaramita et al. (2007) have shown that it is advantageous to combine syntactic parsing and semantic tagging in state-of-the-art frameworks. The next step is to rank information units of varying complexity and structure; e.g., entities (Zaragoza et al., 2007) or answers (Surdeanu et al., 2008), based on semantic annotations. One additional semantic information to exploit in the future is time (Alonso et al., 2007).

The main usage source are queries and the actions after them. In (Baeza-Yates et al., 2007) we present a first step to infer semantic relations from query logs by defining equivalent, more specific, and related queries, which may represent an implicit folksonomy. To evaluate the quality of the results we used the Open Directory Project, showing that equivalence or specificity had precision of over 70% and 60%, respectively. For the cases that were not found in the ODP, a manually verified sample showed that the real precision was close to 100%. What happened was that the ODP was not specific enough to contain those relations. So one main challenge is how to prove the quality of semantic resources if what we can generate is larger than any other available semantic resource and every day the problem gets worse as we have more data. This shows the real power of

the wisdom of crowds, as queries involve almost all Internet users.

4. Epilogue

By being able to generate semantic resources automatically, even with noise, and coupling that with the open semantic resources we have described, we create a virtuous feedback circuit. In fact, taxonomies as well as explicit and implicit folksonomies can be used to do supervised machine learning without the need of manual intervention (or at least by drastically reducing it) to improve semantic tagging. After, we can feedback the results on itself, and repeat the process. Using the right conditions, every iteration should improve the output, obtaining a virtuous cycle.

In particular, SearchMonkey is a strong initiative by Yahoo! to feed this virtuous circle by allowing people to mash up based on result metadata. Microsearch (Mika, 2007) is an early example of this: you can see the metadata in the search and therefore you are encouraged to add to it. There is a button next to every result called "Update metadata" which gives you instant feedback of what your metadata looks like.

Acknowledgments

We thanks the comments of Massi Ciaramita, Peter Mika, and Hugo Zaragoza.

5. References

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2007. On the Value of Temporal Information in Information Retrieval, *ACM SIGIR Forum* 41(2), 35–41.

Jordi Atserias, Hugo Zaragoza, Massimiliano Ciaramita and Giuseppe Attardi. 2008. Semantically Annotated Snapshot of the English Wikipedia. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).

Ricardo Baeza-Yates, Peter Mika, and Hugo Zaragoza. 2008. Search, Web 2.0, and the Semantic Web. In Trends and Controversies: Near-Term Prospects for Semantic Technologies, R. Benjamins, editor. *IEEE Intelligent Systems* 23 (1), 80–82.

Ricardo Baeza-Yates and Alessandro Tiberi. 2007. Extracting Semantic Relations from Query Logs. In *ACM KDD 2007*, San Jose, California, USA, 76–85.

Massimiliano Ciaramita and Giuseppe Attardi. 2007. Dependency Parsing with Second- Order Feature Maps and Annotated Semantic Information. In Proceedings of the 10th International Conference on Parsing Technology.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. Computational Linguistics 29 (3), 333-347.

Peter Mika. 2008. Microsearch demo: http://www.yr-bcn.es/demos/microsearch/.

Simon Overell, Borkur Sigurbjornsson, and Roelof Van Zwol. 2008. Classifying Tags using Open Content Resources. Submitted for publication.

Borkur Sigurbjornsson, and Roelof Van Zwol. 2008. Flickr Tag Recommendation based on Collective Knowledge. In *WWW 2008*, Beijing, China.

Mihai Surdeanu, Massimiliano Ciaramita and Hugo Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT).

James Surowiecki. 2004. *The Wisdom of Crowds*. Random House, New York.

Hugo Zaragoza, Henning Rode, Peter Mika, Jordi Atserias, Massimiliano Ciaramita and Giuseppe Attardi. 2007. Ranking Very Many Typed Entities on Wikipedia. In CIKM '07: Proceedings of the sixteenth ACM international conference on Information and Knowledge Management, Lisbon, Portugal.