From Capturing Semantics to Semantic Search: A Virtuous Cycle

Ricardo Baeza-Yates

Yahoo! Research, Barcelona, Spain

Abstract. Semantic search seems to be an elusive and fuzzy target to IR, SW and NLP researchers. One reason is that this challenge lies in between all those fields, which implies a broad scope of issues and technologies that must be mastered. In this extended abstract we survey the work of Yahoo! Research at Barcelona to approach this problem. Our research is intended to produce a virtuous feedback circuit by using machine learning for capturing semantics, and, ultimately, for better search.

Summary

Nowadays, we do not need to argue that being able to search the Web is crucial. Although people can browse or generate content, they cannot synthesize Web data without machine processing. But synthesizing data is what search engines do, based not only in Web content, but also in the link structure and how people use search engines. However, we are still far from a real semantic search. In [2] we argue that semantic search has not occurred for three main reasons. First, this integration is an extremely hard scientific problem. Second, the Web imposes hard scalability and performance restrictions. Third, there is a cultural divide between the Semantic Web and IR disciplines. We are simultaneously working in these three issues.

Our initial efforts are based in shallow semantics to improve search. For that we first need to develop fast parsers without losing quality in the semantic tagging process. Attardi and Ciaramita [1] have shown that this is possible. Moreover, Zaragoza *et al* [7] have shared a semantically tagged version of the Wikipedia. The next step is to rank sentences based on semantic annotations, and preliminary results in this problem are presented in [8]. Further results are soon expected, where one additional semantic source to exploit in the future is time [3].

The Semantic Web dream would effectively make search easier and hence, semantic search trivial. However, the Semantic Web is more a social rather than a technological problem. Hence, we need to help the process of adding semantics by using automatic techniques. A first source of semantics can be found in the Web 2.0. One example is the Flickr folksonomy. Sigurbjornsson and Van Zwol [5] have shown how to use collective knowledge (or the wisdom of crowds) to extend image tags, and also they prove that almost 80% of the tags can be semantically classified by using Wordnet and Wikipedia [6]. This effectively improves image

search. A second source of implicit semantics are queries and the actions after them. In fact, in [4] we present a first step to infer semantic relations by defining equivalent, more specific, and related queries, which can represent an implicit folksonomy. To evaluate the quality of the results we used the Open Directory Project, showing that equivalence or specificity have 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, and every day the problem gets worse as we have more data. This shows the real power of the wisdom of the crowds, as queries involve almost all Internet users.

By being able to generate semantic resources automatically, even with noise, and coupling that with open content resources, we create a virtuous feedback circuit. In fact, explicit and implicit folksonomies can be used to do supervised machine learning without the need of manual intervention (or at least drastically reduce 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. As a side effect, in each iteration, we can also improve Web search, our main goal.

References

1. Attardi, G., Ciaramita, M.: Tree Revision Learning for Dependency Parsing. In: Proceedings of the HLT-NAACL 2007 Conference, Rochester, USA (2007)
2. Baeza-Yates, R., Mika, P., Zaragoza, H.: Search, Web 2.0, and the Semantic Web. In: Benjamins, R. (ed.) Trends and Controversies: Near-Term Prospects for Semantic Technologies, IEEE Intelligent Systems, January-February 2008, vol. 23(1), pp. 80–82 (2008)
3. Alonso, O., Gertz, M., Baeza-Yates, R.: On the Value of Temporal Information in Information Retrieval. ACM SIGIR Forum 41(2), 35–41 (2007)
4. Baeza-Yates, R., Tiberi, A.: Extracting Semantic Relations from Query Logs. In: ACM KDD 2007, San Jose, California, USA, August 2007, pp. 76–85 (2007)
5. Sigurbjornsson, B., Van Zwol, R.: Flickr Tag Recommendation based on Collective Knowledge. In: WWW 2008, Beijing, China (April 2008)
6. Overell, S., Sigurbjornsson, B., Van Zwol, R.: Classifying Tags using Open Content Resources (submitted for publication, 2008)
7. Zaragoza, H., Atserias, J., Ciaramita, M., Attardi, G.: Semantically Annotated Snapshot of the English Wikipedia v.0 (SW0) (2007), `research.yahoo.com/`
8. Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., Attardi, G.: Ranking Very Many Typed Entities on Wikipedia. In: CIKM 2007: Proceedings of the sixteenth ACM international conference on Information and Knowledge Management, Lisbon, Portugal (2007)