Semantic Search for Entities in Structured Web Data: NTNU at the Yahoo Semantic Search Challenge 2011

Robert Neumayer, Krisztian Balog, Marek Ciglan,
Wei Wei, Kjetil Nørvåg
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
{neumayer,krisztib,ciglan,wwei,noervaag}@idi.ntnu.no

Abstract

Search for entities has become the most popular type of Web search, second to navigational queries. As such, the search for entities has attracted considerable amount of research interest. In this paper we describe our participation in the Semantic Search Challenge of the Semantic Search 2011 Workshop at WWW'2011, where we achieved first and third place in the list and entity search tracks.

1 Introduction

Entity search denotes to searches targeting entities instead of documents. Entities cover a wide range of categories such as products, persons, or other machine-readable units of search and have become the most common type of Web search after navigational searches [7]. Opposed to common information retrieval search, which operates on a document level, entity search often operates on RDF (Resource Description Framework) or other types of structured data representation. Such structured representations, which provide the directions and types of links between entities, are often referred to as "Semantic Web" as envisioned by Tim Berners Lee[3].

In the context of the semantic search this means that the classical information retrieval keyword search is extended by using RDF input data in the form of *(subject, predicate, object)*, where each component is described by a URI (Uniform Resource Identifier). Entities are represented by subjects and occur together with predicates and objects closer identifying this entity. For example the RDF triple *example.org/NTNU, example.org/hasLocation, example.org/Trondheim* implies that NTNU is located in Trondheim.

In this paper, we give an overview of the Semantic Search Challenge of the Semantic Search 2011 Workshop, summarize our participation and present the results from the participating research teams.

This paper was presented at the NIK-2011 conference; see http://www.nik.no/.

2 Semantic search challenge

The dataset used in the challenge is the Billion Triple Challenge 2009 (BTC) collection. This data set was created for the 2009 Semantic Web Challenge and comprises of 1.4B RDF statements describing 114 million entities. The dataset is publicly available for download[1].

The 2011 search challenge consisted of two separate tasks: 1) entity search and 2) list search. The entity search task is a continuation of the previous year's task where participating teams are given keyword queries and the goal being to find the most relevant entities with respect to one particular entity (e.g. 'YMCA Tampa') [6]. The list search task, on the other hand, contains more complex queries matching multiple entities (e.g. 'Arab states of the Persian Gulf'). This is a task similar in spirit to the List Completion problem at the INEX Entity Ranking track [5] and to the Entity List Completion task of the TREC Entity track [2]). More details on the search challenge can be found in [4].

Each team was allowed to submit three runs, i.e. different setups, to the challenge. The best out of these would then be used to determine the teams' rankings.

Our main emphasis for the entity search was on combining evidence from multiple knowledge sources, i.e., BTC and DBPedia[2], where each source is queried using a retrieval method tailored to its specific properties. In our participation we focused on integrating evidence from multiple sources for the entity search task: we employed a fielded Language Modeling approach to rank entities in the BTC collection and in DBPedia.

With respects to list search, our goal was to mimic the behaviour of humans searching in Wikipedia for we believe much of the answers to list queries is available there, albeit not directly accessible. Our approach includes a query analysis step to identify the principal entity in the query. In the ranking phase we utilize the Wikipedia link graph and semantically related article sets, defined by Wikipedia categories and templates.

For a more detailed description of our approaches and the theoretical background we refer to the system description in [1].

3 Results of the challenge

A total of four teams participated in the entity search track. An overview of the results is given in Table 1 (left). Sindice's submission achieved the highest scores in terms of MAP, second to the university of Delaware (Udel). All three of NTNU's setups were competitive and got the respective third, fourth, and fifth rank.

For the list search track, a total of five teams submitted (all teams participating in the entity search track plus an additonal team, the Ambani Institute of Information and Communication Technology (DAA-IICT)). NTNU's approach clearly outperformed the other submissions as shown in Table 1 (right). Sindice and Delaware came second and third, however, with a clearly lower MAP. We attribute the clear win to our strategy of dividing our resources between both tasks early in the process and targeting strong submissions in both categories. The other teams used only slight modifications with respect to query processing over their submissions to the entity search track. Also all other teams used the BTC data whereas we used mainly Wikipedia.

[1] `http://vmlioni25.deri.ie`

[2] `http://wiki.dbpedia.org/Downloads36`

Entity search track				List search track			
Rank	Participant	Run	MAP	Rank	Participant	Run	MAP
1	9-Sindice	2	0.2346	1	3-NTNU-Godfrid	3	0.2790
2	13-UDel	2	0.2167	2	3-NTNU-Harald	2	0.2594
3	3-NTNU	1	0.2072	3	3-NTNU-Olav	1	0.1625
4	3-NTNU	2	0.2063	4	9-Sindice	1	0.1591
5	3-NTNU	3	0.2050	5	9-Sindice	3	0.1526
6	13-UDel	1	0.1858	6	9-Sindice	2	0.1505
7	9-Sindice	1	0.1835	7	13-UDel	1	0.1079
8	9-Sindice	3	0.1635	8	13-UDel	2	0.0999
9	5-IIIT Hyd	1	0.0876	9	5-IIIT Hyd	1	0.0328
10	5-IIIT Hyd	2	0.0870	10	5-IIIT Hyd	2	0.0328
				11	15-Daiict	1	0.0050

Table 1: Competition results.

4 Conclusions

With our best entity search run ranked third among all submissions and our list search runs were placed first, second, and third, of all runs, we consider our approaches competitive and intend to further improve on them.

In future work we plan to perform an exhaustive success and failure analysis based on the full system evaluations. As our approaches employ a number of parameters, most of which were set intuitively in the lack of time and—in case of the list search task—in the lack of training data, we believe that there is much to gain by adjusting these settings.

Acknowledgements

This work was carried out as part of the COMIDOR (Cooperative Mining of Independent Document Repositories) project, supported by grant #183337/S10 from the Norwegian Research Council.

References

[1] K. Balog, M. Ciglan, R. Neumayer, W. Wei, and K. Nørvåg. NTNU at SemSearch 2011. In *SemSearch Challenge System Descriptions*, `http://semsearch.yahoo.com/results.php`, 2011.

[2] K. Balog, P. Serdyukov, and A. P. de Vries. Overview of the TREC 2010 entity track. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010)*. NIST, February 2011.

[3] T. Berners-Lee and M. Fischetti. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper San Francisco, September 1999.

[4] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and T. T. Duc. Entity search evaluation over structured web data. In *Proceedings of the First International Workshop on Entity-Oriented Search (EOS)*, 2011.

[5] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the INEX 2009 entity ranking track. In *Proceedings of the Focused retrieval and evaluation, and 8th international conference on Initiative for the evaluation of XML retrieval*, INEX'09, pages 254–264. Springer-Verlag, 2010.

[6] H. Halpin, D. M. Herzig, P. Mika, R. Blanco, J. Pound, H. S. Thompson, and D. T. Tran. Evaluating ad-hoc object retrieval. In *Proceedings of the International Workshop on Evaluation of Semantic Technologies (IWEST 2010)*, 2010.

[7] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web (WWW'10)*, pages 771–780, New York, NY, USA, 2010. ACM.