When Simple is (more than) Good Enough: Effective Semantic Search with (almost) no Semantics

Robert Neumayer, Krisztian Balog, and Kjetil Nørvåg

Norwegian University of Science and Technology, Trondheim, Norway
{robert.neumayer,krisztian.balog,kjetil.norvag}@idi.ntnu.no

Abstract. Using keyword queries to find entities has emerged as one of the major search types on the Web. In this paper, we study the task of ad-hoc entity retrieval: keyword search in a collection of structured data. We start with a baseline retrieval system that constructs pseudo documents from RDF triples and introduce three extensions: preprocessing of URIs, using two-fielded retrieval models, and boosting popular domains. Using the query sets of the 2010 and 2011 Semantic Search Challenge, we show that our straightforward approach outperforms all previously reported results, some generated by far more complex systems.

1 Introduction

A considerable amount of all Web search queries target entities or objects such as persons, locations, or services [7]. At the same time, there is an increased amount of information published as Linked Data that is inherently organised around entities; each entity is identified by a unique URI and is described using a set of subject-predicate-object RDF triples. Querying these structured data sources by the means of simple keyword search (as opposed to SPARQL-like languages) emerged as a genuine user need and has recently become an active topic of research [1–3, 6, 7]. The task we are studying in this paper is *ad-hoc entity retrieval* (often referred to as *semantic search*): "answering arbitrary information needs related to particular aspects of objects [entities], expressed in unconstrained natural language and resolved using a collection of structured data" [7].

The Semantic Search Challenge, organised in 2010 and 2011, has provided a platform for researchers to empirically evaluate systems developed for this task. Approaches range from plain text-based retrieval on the one end of the spectrum to "semantic search," taking relations and links between entities into account, on the other. For a full account, we refer to the challenge overviews [1, 4].

Commonly, standard document retrieval methods are adapted to the ad-hoc entity search task, by building pseudo-documents from RDF triples associated with entities. The challenge is how to capture semantics in this document-like representation. While assigning different relations (predicates) to different fields seems like the most natural option, this solution becomes computationally prohibitive because of the large number of possible fields. Therefore, a viable alternative is to group predicates together into a small set of predefined categories, for example based on their type (attributes, incoming/outgoing relations, etc.) [6] or based on their manually assigned importance [2]. These representations can then be ranked using fielded extensions of standard document retrieval models, such as the Mixture of Language Models [5] or BM25F [8].

R. Baeza-Yates et al. (Eds.): ECIR 2012, LNCS 7224, pp. 540–543, 2012.

Starting from a standard document retrieval approach, we consider the following extensions: (1) extended preprocessing, a heuristic for extracting textual content from URI descriptors, (2) a two-field representation, distinguishing between title and content, and (3) entity importance, assigning more weight to entities from trusted, high-quality sources. We show that these extensions lead to improvements and that they add up. In fact, our approach outperforms all previously reported results, despite that those were generated by far more complex systems.

2 Approach

We address the ad-hoc entity search task in RDF data: given a keyword query, targeting a particular entity, return a ranked list of relevant entities identified by their URIs. Each entity is described by a set of subject-predicate-object RDF triples. For each entity, we build a textual representation by considering all triples where it stands as the subject; we use only the object's (string) value from the triple and refer to it as *object value*.

Baseline Retrieval. All object values are concatenated together into a flat text representation. We perform standard tokenization and stopword removal; no stemming is applied. We use standard retrieval models: BM25 and Language Models (LM).

Fielded Representation. We use a simple heuristic to identify predicates that hold title values: these end with "name," "label," or "title". Object values belonging to title-type predicates are concatenated into an additional *title* field. This is the only part where we have some (limited) semantics captured in our approach. Given this title+content representation, we use fielded versions of BM25 and LM, specifically, BM25F [8] and the Mixture of Language Models [5], referred to as LMF.

Entity Importance. Entities from trusted, high-quality sources are considered more important and receive an extra query-independent weight in their retrieval score. In case of BM25, this is incorporated as a multiplication factor; for LM, we use the document (entity) priors for this is purpose. We illustrate the effects of this component by boosting DBpedia, which is a central hub in the Linked Data cloud.

Extended Preprocessing. For all settings we introduced before, we apply a heuristic to extract (additional) textual content from URIs. We do so by using the string part of the URI after the last slash as the object value. Additionally, we make sure that characters like underscores, dashes, brackets, etc. are all treated as whitespaces.

3 Experiments

Our experimental evaluation is based on the Semantic Search Challenge (SemSearch) 2010 and 2011 evaluation campaigns.[1] The data collection used there is the Billion Triple Challenge 2009 corpus; it comprises about 1.14 billion RDF statements collected by a Semantic Web crawler.[2] The two entity search query sets contain 92 and 50 keyword queries for 2010 and 2011, respectively, and come with relevance judgments (in standard TREC format). We report on Mean Average Precision (MAP), the main metric used at SemSearch. Significance testing is performed using a two-tailed paired t-test.

[1] `http://km.aifb.kit.edu/ws/semsearch{10|11}`
[2] `http://vmlion25.deri.ie`

Table 1. Retrieval results. (Rows 1-12): results from this paper; (Rows 13-14): best results from the literature. Best scores for each column are in boldface.

Run	URI Preproc.	Retrieval Model	2010 MAP	2011 MAP	2011 (opt) MAP
Baseline (content)	-	LM	0.1832	0.1840	0.1840
	-	BM25	0.1888	0.1970	0.2154
	+	LM	0.2388‡	0.2445‡	0.2445‡
	+	BM25	0.2464‡	0.2502‡	0.2702‡
title+content	-	LMF	0.1832	0.1840	0.1840
	-	BM25F	0.1888	0.1970	0.2154
	+	LMF	0.2900‡	0.2618‡	0.2765‡
	+	BM25F	0.2621‡	0.2625‡	0.2937‡
title+content	-	LMF	0.1836‡	0.1846‡	0.1846‡
+ *dbpedia.org* boosting	-	BM25F	0.1909‡	0.2031‡	0.2166‡
	+	LMF	**0.2914**‡	0.2651‡	0.2756‡
	+	BM25F	0.2631‡	**0.2642**‡	**0.2991**‡
Best at SemSearch [1, 4]			0.1919		0.2346
Best reported since [3]			0.2805		

Table 1 reports on a series of experiments we performed using two different retrieval models (LM and BM25) and two different parameter settings. For the default setting, shown in columns 4 and 5, no training material is used; we take values suggested in the literature or values that intuitively seem reasonable. For LM, we use the average document/field length ($avgdl$) as the smoothing parameter μ [5]. For BM25, we use $k1$=1.2 and b=.25; we use the same b value for all fields in the fielded variant BM25F, analogous to [5] and [3]. We use a weighting of $0.2/0.8$ for the title/content fields. The optimised parameter setting, displayed in column 6, is only for the 2011 query set. We use relevance assessments from the previous year as training material; these were also available to SemSearch 2011 participants. The best found parameter settings are: μ=$avgdl$ for LM, μ=$2 \cdot avgdl$ for LMF, and $k1$=0.4 and b=0.4 for BM25/BM25F.

First, in rows 1-4, we use standard retrieval models with flat text representation. For both query sets, we see large differences depending on the URI preprocessing; all results using the advanced preprocessing in rows 3-4 for URIs are significantly different from the baselines without preprocessing in rows 1-2. Next, in rows 5-8, we use fielded variants of these models, with two fields: title and content. The results in rows 5-6 equal rows 1-2; this is because the title field cannot contribute to the entity representation without URI preprocessing. The results in rows 7 and 8, however, outperform their counterparts in rows 3 and 4; assigning higher weight to the title field clearly benefits retrieval when title values are extracted correctly. Finally, in rows 9-12, we boost entities coming from high-quality trusted sources, in our case DBpedia. In columns 4 and 5 we use the boosting value of 1.5, indicating that all scores of DBpedia entities are multiplied by that value. In column 6, we show results for the boosting factor that showed the best results on the 2010 queries (a value of 2.2). However, this leads to a performance decrease in row 11 compared to row 7; we attribute this to the fact that there are more relevant answers from DBpedia for 2010 than for 2011.

We chose to use both BM25 and LM to investigate if both retrieval models display the same behaviour with respect to the techniques we applied. We find that this is indeed the case, but we also discovered two interesting differences. First, with default parameter settings, LM performs better on the 2010 queries while BM25 does slightly better on 2011. Second, BM25 benefits more from parameter optimisation.

In comparison to other approaches, we outperform all published results for both years' queries as shown in the last two rows of Table 1. Campinas et al. [3] report improved results for the 2010 query set and achieve a MAP of .2805; a large fraction of their improvements can be attributed to additional query, attribute, and entity weighting. Blanco et al. [2] report a MAP score of .2705 on the 2010 queries using a manual grouping and weighting of predicates. Both works use BM25F.

4 Conclusions

We addressed the task of entity search in Linked Data using the BTC-2009 collection and the test sets of the 2010 and 2011 Semantic Search Challenges. Starting from a baseline using standard document retrieval techniques, we introduced three expansions: (1) a heuristic for extracting textual content from URI descriptors, (2) a two-field representation, based on title and content, and (3) boosting entities from trusted domains. We showed that our approach is highly competitive and that it outperforms all previously reported results on these data sets. The extent of our improvements is somewhat surprising because our approach is straightforward in terms of transforming RDF triples into a flat structure and applying known IR techniques. We observe that the extended URI preprocessing component accounts for the majority of the improvements.

References

[1] Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S., Duc, T.T.: Entity search evaluation over structured web data. In: 1st Intl. Workshop on Entity-Oriented Search (EOS 2011), pp. 65–71 (2011a)

[2] Blanco, R., Mika, P., Vigna, S.: Effective and Efficient Entity Search in RDF Data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 83–97. Springer, Heidelberg (2011b)

[3] Campinas, S., Delbru, R., Rakhmawati, N.A., Ceccarelli, D., Tummarello, G.: Sindice BM25F at SemSearch 2011. In: 4th Intl. Semantic Search Workshop (2011)

[4] Halpin, H., Herzig, D.M., Mika, P., Blanco, R., Pound, J., Thompson, H.S., Tran, D.T.: Evaluating ad-hoc object retrieval. In: Proc. of the Intl. Workshop on Evaluation of Semantic Technologies (2010)

[5] Ogilvie, P., Callan, J.: Combining document representations for known-item search. In: SIGIR 2003, pp. 143–150 (2003)

[6] Pérez-Agüera, J.R., Arroyo, J., Greenberg, J., Iglesias, J.P., Fresno, V.: Using BM25F for semantic search. In: 3rd Intl. Semantic Search Workshop (2010)

[7] Pound, J., Mika, P., Zaragoza, H.: Ad-hoc object retrieval in the web of data. In: Proc. of the 19th Intl. Conf. on World Wide Web, pp. 771–780 (2010)

[8] Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proc. of the 13th Intl. Conf. on Inf. and Knowledge Man. (2004)