Ranking Distributed Knowledge Repositories

Robert Neumayer, Krisztian Balog, and Kjetil Nørvåg

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

Abstract. Increasingly many knowledge bases are published as Linked Data, driving the need for effective and efficient techniques for information access. Knowledge repositories are naturally organised around objects or entities and constitute a promising data source for entity-oriented search. There is a growing body of research on the subject, however, it is almost always (implicitly) assumed that a centralised index of all data is available. In this paper, we address the task of ranking distributed knowledge repositories—a vital component of federated search systems—and present two probabilistic methods based on generative language modeling techniques. We present a benchmarking testbed based on the test suites of the Semantic Search Challenge series to evaluate our approaches. In our experiments, we show that both our ranking approaches provide competitive performance and offer a viable alternative to centralised retrieval.

1 Introduction

In recent years the number of knowledge bases published as Linked Data has significantly increased. These range from general-purpose encyclopaediae like DBpedia or Freebase, to domain-specific databases, such as GeoNames for geographical entities. These knowledge bases are inherently organised around "entities" or "objects," such as persons, places, organisations, artifacts, etc. This, coupled with the fact that the most frequent types of queries in web search revolve around entities [10], lends significance to the *ad-hoc entity retrieval* task, defined as follows: "answering arbitrary information needs related to particular aspects of objects [entities], expressed in unconstrained natural language and resolved using a collection of structured data" [10]. The importance of search focused on entities is also witnessed by numerous tasks that have been featured at the TREC [15, 2, 3] and INEX [7] evaluation benchmarking campaigns.

Knowledge repositories are typically both heterogeneous and inherently distributed as they are located on disparate servers. Only in few cases it is possible to maintain a central index encompassing the contents of all individual data sources. Many sources can be covered only partially (for example, specific parts might be overlooked by spiders), while others may not be crawleable at all (due to authorisation settings prohibiting access). Instead of expending effort to crawl all data from these sources, one might pass the query to the search interface of multiple, suitable collections (usually distributed across several locations)—an approach known as *distributed information retrieval (DIR)*, also referred to as *federated search* [11]. For example, the query "painters of the gothic era" may be passed to a related collection, such as a digital library of a

museum, while for the query "San Antonio" collections containing information about the city, such as GeoNames or DBpedia, might be more appropriate. Of course, there are also queries for which multiple databases can contain answers.

When querying distributed knowledge repositories (from now on *collections*) it is desirable to choose only those that (are likely to) contain relevant results. That is, we need to be able to rank individual collections with respect to a given query. This task, *collection ranking*, has received a lot of attention in the past in the DIR literature, but only in the context of traditional document search. We target the more general concept of entities, i.e., any digital object described in terms of ontology-based metadata (from now *entity*). We focus on collection representation and ranking for ad-hoc entity search in a distributed environment. Building on prior DIR research we formulate two collection ranking strategies using a unified probabilistic retrieval framework based on language modeling techniques. According to one model (Collection-centric), each collection is represented as a term distribution computed over its contents. Our second model (Entity-centric) estimates the relevance of each individual entity within the collection and then aggregates these scores to determine the collection's relevance.

We introduce an experimental platform based on the data set and topics from the Semantic Search Challenge [9, 4]. We assume a cooperative environment, where collections provide information about their contents, such as their term statistics, and can implement the same retrieval function for ranking entities. We find that it is indeed necessary for collections to provide information about their term statistics; with these available, our models achieve very competitive performance. Assigning higher prior importance to larger collections, a reasonable heuristic, brings in further improvements.

In summary, this work makes the following contributions: (1) a unified generative modeling framework and two particular models for collection ranking, (2) a test set for evaluating the collection ranking task in Linked Data, and (3) an experimental comparison of our models using this data set.

2 Related Work

Distributed information retrieval (DIR) or *federated search*, is ad-hoc search in environments containing multiple, text databases [5]. DIR targets cases when documents cannot be copied into a centralised database. It involves three important sub-problems: (I) *acquiring resource descriptions*, representing the content of each collection in some suitable form, (II) *resource selection*, selecting the most relevant collections, and, finally, (III) *result merging*. We restrict our attention to (I) and focus on both the representation and ranking of resources (collections).

Federated search techniques have recently been picked up by the digital library community too. MinervaDL, a digital library architecture for information retrieval in peer-to-peer (P2P) networks is presented in [16]. It differs from our work in three important aspects: they use a different architecture, do not consider retrieval effectiveness, and work with much smaller collections. Linked Data (LD) also bears increasing importance to digital libraries, as it can be beneficial to enriching metadata. For example, [14] suggest to automatically link FRBR works to the corresponding entity in LD.

3 Representing and Ranking Distributed Collections

In this section we present our approach for representing and ranking collections. We formulate this task in a generative probabilistic framework and rank collections based on their likelihood of containing entities relevant to an input query, $P(C|Q)$. Instead of estimating this probability directly, we apply Bayes' rule and rewrite it to $P(C|Q) \propto P(Q|C)P(C)$. Thus, the score of a collection is made up of two components: (1) *query generator* ($P(Q|C)$), that is, the probability of a query being generated by collection C; this can be interpreted as the collection's relevance to the query; (2) *collection prior* ($P(C)$), that is, the *a priori* probability of selecting collection C; this tells us how likely the collection is to contain the answer to any arbitrary query.

We propose two models for estimating the query generator by drawing upon existing strategies to collection ranking and formalise them within a language modeling framework. Our two approaches bear resemblance to the expert finding models of Balog et al. [1] and to the blog feed search models of Elsas et al. [8], but differ in the estimation of specific components. Let us remind ourselves that we assume a cooperative environment, in which collections can provide general term statistics and can implement the same retrieval function. Further, we assume that for each entity E, a document representing that entity, D_E, has already been created.

Collection-centric model. One of the simplest approaches to resource selection is to treat each collection as a single, large document [6, 13]. Once such a pseudo-document is generated for each collection, we can rank collections much like documents. In a language modeling setting this ranking is based on the probability of the collection generating the query. Formally:

$$P(Q|C) = \prod_{t \in Q} \Big\{ (1-\lambda_G)\big(\sum_{E \in C} P(t|D_E)P(E|C) \big) + \lambda_G P(t|G) \Big\}^{n(t,Q)}, \quad (1)$$

where $n(t,Q)$ is the number of times term t is present in the query Q, $P(t|D_E)$ and $P(t|G)$ are maximum-likelihood estimates of the probability of observing term t given the entity and global (cross-collection) language models, respectively, and λ_G is the (global) smoothing parameter. We assume that all entities are equally important within a given collection, thus set $P(E|C) = 1/|C|$.

Entity-centric model. Instead of creating a direct term-based representation of collections, our second approach models and queries individual entities, then aggregates their relevance estimates:

$$P(Q|C) = \sum_{E \in C} P(E|C) \prod_{t \in Q} \big((1-\lambda_G)P(t|\theta_E) + \lambda_G P(t|G) \big)^{n(t,Q)}, \quad (2)$$

where $P(t|\theta_E)$ is the probability of term t given the entity's language model and $P(t|G)$ is the global background language model. It is worth noting that this model employs smoothing on two levels: (1) on the entity's level, by smoothing the entity document with the collection (in estimating $P(t|\theta_E)$), and (2) on the collection level, by mixing with the global background model using coefficient λ_G.

This model resembles the ReDDE collection selection algorithm by Si and Callan [12]. The main difference is that we do not incorporate the collection size directly into the scoring formula, but accommodate it through the collection prior.

Collection priors. To estimate the *a priori* probability of a collection, $P(C)$, we consider two alternatives. The simplest choice is to assume that all collections are equally important: $P(C) \propto 1$. We refer to this as the *uniform* prior. Intuitively, larger collections are more likely to contain relevant entities to any information need. According to the *collection size* prior, we set the $P(C) \propto |C|$.

4 Experimental setup

Our testbed is based on the 2010 and 2011 editions of the Semantic Search Challenge [9, 4]. Queries there request specific named entities (e.g., "american embassy nairobi" or "martin luther king") from a collection of structured data, described as RDF. The data collection we use is the Billion Triple Challenge 2009 dataset.[1] Crawled during February/March 2009, it comprises about 1.14 billion RDF statements. For our experiments, we consider the top 100 second-level domains (measured in terms of the number of entities contained) as distributed knowledge repositories (i.e., our set of collections).

Queries originate from the Yahoo! Search Query Tiny Sample dataset.[2] Relevance judgments were obtained using Amazon's Mechanical Turk. We used all queries from 2010 and 2011, but filtered out those that did not have any relevant results from the top 100 domains that we considered as our collections; this left us with 136 queries in total. In our setting, a collection is considered relevant if it contains at least one entity that was judged relevant. For graded evaluation metrics, we set the relevance level of a collection to the number of relevant documents it contains.

We use standard IR evaluation metrics: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Recall at 5 (R@5), and Normalised Discounted Cumulative Gain (NDCG). Significance testing is done using a two-tailed paired t-test.

5 Experimental evaluation

The main research question we seek to answer is the following: How to represent collections for distributed entity retrieval? We address the following specific sub-questions: (1) What is the effect of taking global (cross-collection) term statistics into account? (2) Which of the Collection-centric (CC) and Entity-centric (EC) collection ranking approaches perform better? (3) What is the impact of collection priors?

Using global term statistics. To investigate the potential benefits of having knowledge about the global (cross-collection) importance of query terms, we consider two settings: (1) not making use of using global term statistics; within our language modeling framework this is implemented by setting the λ_G (global) smoothing parameter to 0 in Eqs. 1

[1] http://vmlion25.deri.ie/

[2] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

Table 1. Collection ranking results. †/‡denote significant differences at the 0.05/0.01 levels, respectively. Significance is tested for rows 2 vs. 1, 4 vs. 3, 5 vs. 3, and 6 vs. 4. Best scores within each block are typeset boldface.

Global term stat.	Coll. priors	Model	MAP	MRR	R@5	NDCG
No	No	Collection-centric	0.3048	0.4304	0.3552	**0.4873**
No	No	Entity-centric	**0.3710**‡	**0.4980**†	**0.4289**‡	0.4766
Yes	No	Collection-centric	**0.5149**	**0.8901**	0.5208	**0.8280**
Yes	No	Entity-centric	0.5134‡	0.8404†	**0.5262**	0.7967‡
Yes	Yes	Collection-centric	0.5368‡	0.9282†	0.4645†	0.8506‡
Yes	Yes	Entity-centric	**0.5494**‡	**0.9283**‡	**0.4817**†	**0.8564**‡

and 2, and (2) using global term statistics; we use Dirichlet smoothing and set λ_G proportional to the average collection length. Table 1 displays the results without and with global term statistics (rows 1-2 vs. rows 3-4). It is clear that using this information leads to substantial and significant improvements for both methods.

Collection-centric vs. Entity-centric models. Based on the numbers in Table 1, if global term statistics are omitted, the Entity-centric model clearly outperforms the Collection-centric one (except for an insignificant degradation for NDCG). With global term statistics, however, their performance is much closer to each other, with CC actually performing better on MRR and NDCG. Interestingly, the two models have almost the same performance when averaged over all topics, but they generate significantly different results, i.e., on the level of individual topics, there are sometimes substantial differences between the two models, but these differences equal out on average.

Collection priors. Our last set of experiments focuses on collection priors; the last two rows of Table 1 report the results. Priors improve for both collection selection methods on the precision-oriented metrics (MAP, MRR, NDCG), and especially help precision at the top rank (MRR). With MRR scores in the 0.9 range, there is not much room left for improvement at rank 1. Nevertheless, this is done at the expense of recall. All differences are significant.

6 Conclusions

To the best of our knowledge, ours is the first work to apply federated IR techniques in the context of entity search. In this paper, we presented two methods for collection ranking of distributed knowledge repositories. One approach scores an individual collection by collapsing all text associated with its entities into one pseudo-document. The other considers individual entities and aggregates their relevance scores on the collection level. Our experimental comparison of these two approaches showed that both deliver excellent performance. For both cases, we have shown the importance of having

access to global term statistics and the benefits of incorporating collection priors. In future work, we plan to expand our work to non-cooperative environments.

Bibliography

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR*, pages 43–50, 2006.

[2] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. P. de Vries, and P. Bailey. Overview of the TREC 2008 enterprise track. In *The 17th Text Retrieval Conference Proceedings (TREC 2008)*. NIST, 2009.

[3] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *Proceedings of the 18th Text REtrieval Conference (TREC 2009)*. NIST, February 2010.

[4] R. Blanco, H. Halpin, D. Herzig, P. Mika, J. Pound, H. Thompson, and T. Duc. Entity search evaluation over structured web data. In *1st International Workshop on Entity-Oriented Search (EOS)*, pages 65–71, 2011.

[5] J. Callan. Distributed information retrieval. In *In: Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.

[6] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of SIGIR*, pages 21–28, 1995.

[7] G. Demartini, A. de Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 entity ranking track. volume 5631, pages 243–252. Springer, 2009.

[8] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR*, pages 347–354, 2008.

[9] 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.

[10] 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*, pages 771–780, 2010.

[11] M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5:1–102, 2011.

[12] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of SIGIR*, pages 298–305, 2003.

[13] L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 391–397, 2002.

[14] N. Takhirov, F. Duchateau, and T. Aalberg. Linking FRBR entities to LOD through semantic matching. In *TPDL*, volume 6966, pages 284–295. Springer, 2011.

[15] E. Voorhees. Overview of the TREC 2004 question answering track. In *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, Gaithersburg, 2005. NIST. Special Publication: SP 500-261.

[16] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: an architecture for information retrieval and filtering in distributed digital libraries. In *ECDL*, volume 4675, pages 148–160. Springer, 2007.