Entity Search: Building Bridges between Two Worlds

Krisztian Balog
k.balog@uva.nl

Edgar Meij
e.j.meij@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam, Science Park 107
1098 XG Amsterdam

ABSTRACT

We consider the task of entity search and examine to which extent state-of-art information retrieval (IR) and semantic web (SW) technologies are capable of answering information needs that focus on entities. We also explore the potential of combining IR with SW technologies to improve the end-to-end performance on a specific entity search task. We arrive at and motivate a proposal to combine text-based entity models with semantic information from the Linked Open Data cloud.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Measurement, Performance, Experimentation

1. INTRODUCTION

We have come to depend on technological resources to create order and find meaning in the ever-growing amount of online data. One frequently recurring type of query in web search are queries containing named entities (persons, organizations, locations, etc.) [14]: we organize our environments around entities that are meaningful to us. Hence, to support humans in dealing with massive volumes of data, next generation search engines need to organize information in *semantically meaningful* ways, structured around entities. Furthermore, instead of merely finding documents that mention an entity, finding the entity itself is required.

The problem of entity search has been and is being looked at by both the Information Retrieval (IR) and Semantic Web (SW) communities and is, in fact, ranked high on the research agendas of the two communities. The entity search task comes in several flavors. One is known as *entity ranking* (given a query and target category, return a ranked list of relevant entities [10]), another is *list completion* (given a query and example entities, return similar entities [10]), and a third is *related entity finding* (given a source entity, a relation and a target type, identify target entities that enjoy the specified relation with the source entity and that satisfy the target type constraint [5]).

State-of-the-art IR models allow us to address entity search by identifying relevant entities in large volumes of web data. These methods often approach entity-oriented retrieval tasks by establishing associations between topics, documents, and entities or amongst entities themselves [4], where such associations are modeled by observing the language usage around entities [22, 23]. A major challenge with current IR approaches to entity retrieval is that they fail to produce interpretable descriptions of the found entities or of the relationships between entities. The generated models tend to lack human-interpretable semantics and are rarely meaningful for human consumption: interpretable labels are needed (both for entities and for relations) [7, 19]. Linked Open Data (LOD) is a recent contribution of the emerging semantic web that has the potential of providing the required semantic information [2, 3, 6, 24].

From a SW point of view, entity retrieval should be as simple as running SPARQL queries over structured data. However, since a true semantic web still has not been fully realized, the results of such queries are currently not sufficient to answer common information needs. By now, the LOD cloud contains millions of concepts from over one hundred structured data sets. This abundance, however, also introduces novel issues such as "cheap semantics" (e.g. `wikilink` relations in DBpedia) and the need for ranking potentially very large amounts of results [1]. Furthermore, given the fact that most web users are not proficient users of semantic web languages such as SPARQL or standards such as RDF and OWL, the free-form text input used by most IR systems is more appealing to end users.

These concurrent developments give rise to the following general question: to which extent are state-of-art IR and SW technologies capable of answering information needs related to entity finding? In this paper we focus on the task of *related entity finding* (REF). E.g., for a source entity ("Michael Schumacher"), a relation ("Michael's teammates while he was racing in Formula 1") and a target type ("people"), a REF system should return entities such as "Eddie Irvine" and "Felipe Massa." REF aims at making *arbitrary* relations between entities searchable. We focus on an adaptation of the official task as it was run at TREC 2009 and restrict the target entities to those having a primary Wikipedia article: this modification provides an elegant way of making the IR and SW results comparable.

From an IR perspective, a natural way of capturing the relation between a source and target entity is based on their co-occurrence in suitable contexts [5]. Later, we use an aggregate of methods all of which are based on this approach. In contrast, a SW perspective on the same task is to search for entities through links such as the ones in LOD and for

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

this we apply both standard SPARQL queries and an exhaustive graph search algorithm.

Below, we analyze and discuss to which extent REF can be solved by IR and SW methods. It is important to note that our goal is not to perform a quantitative comparison, and make claims about one approach being better than the other or vice versa. Rather, we investigate results returned by either approach and perform a more qualitative evaluation. We find that IR and SW methods discover different sets of entities, although these sets are overlapping. Based on the results of our evaluation, we demonstrate that the two approaches are complementary in nature and we discuss how each field could potentially benefit from the other.

In Section 2 we discuss related work from both fields. We then zoom in on our experimental environment (Section 3) and approaches (Section 4). In Section 5 we discuss our results and provide suggestions as to how IR could benefit from SW and vice versa. We end with a concluding section.

2. RELATED WORK

Until the mid 1990s, research in IR was mostly aimed at document retrieval but, since then, interest in IR tasks that go beyond document retrieval has steadily increased. Research into identifying entities with a certain property or engaging in a certain relation received a boost with the launch of the TREC Question Answering track [25] in 1999. The track ran for a decade and had a strong focus on entities (e.g., "Who invented the paperclip?" or "List subway stations in Washington."). In 2005, a dedicated expert finding track was launched at TREC, where a list of experts had to be returned for a given topic [9]. In 2007, the Initiative for the Evaluation of XML Retrieval (INEX) launched an Entity Ranking track, to evaluate entity retrieval in Wikipedia [10]. Another key development is the recent introduction of an Entity track at TREC, which aims at evaluating entity-related search tasks on the Web [5].

In IR, entity search, in any of the flavors listed above, is typically addressed using statistical methods, possibly complemented with more knowledge-intensive components. At the heart of these methods lies a mechanism for computing co-occurrences: between question words and answer entities [25], between topics and experts [9], or between source and target entities [5].

Within the SW community there are various approaches relevant to the entity finding task [18]. Given our task, i.e., finding related entities given a source entity and relation, Lehmann et al. [17] describe a highly relevant method which looks for any path between a source and target object. They use a variant of Dijkstra's shortest path algorithm [11] to find the shortest path between any two objects in DBpedia (their most recent demo also includes more sources from the LOD cloud.) We have implemented an approach that is very similar in nature; see Section 4.2. Other related work from the SW community includes linking free-text queries to objects in the LOD cloud [20], using variants of PageRank for ranking items in a result set [13, 16], and using information-theoretic measures to rank associations [1]. However, without any clear evaluation methodology or standard test collections, the results of many of these approaches are hard to compare. Moreover, most experimental results seem to be presented as small-scale case-studies, by their success inside a target application, or by extension, i.e., merely providing examples for which the method in question works or fails.

3. EXPERIMENTAL SETUP

We consider the following research questions: (i) To which extent are state-of-art IR and SW technologies capable of answering entity search types of information needs? And (ii) What is the potential of combining IR with SW technologies to improve the end-to-end performance? Next, we detail the experimental setup that we use for answering those questions.

3.1 Task definition

The TREC Entity track defines the REF task as follows: given an input entity (with its name and homepage), a type of target entity, and the nature of their relation (described in free text), find related entities that are of the target type and that stand in the required relation to the input entity [5]. For our experiments we use a modified version of the original TREC Entity task, where target entities are represented by their Wikipedia page instead of their homepage, if any. As Wikipedia articles are easily mappable to DBpedia concepts, this modification allows us to compare result sets generated by IR and SW methods.

3.2 Topics and ground truth

We base our test set on the TREC 2009 Entity topics. A topic consists of a source entity (E), a target entity type (T) and the desired relation (R) described in free text, where T is limited to either a person (PER), organization (ORG), or product (PROD). A set of 20 test topics was made available, but we use only 17 of them as 3 topics (#2, #3, and #16) have source entities that do not have a Wikipedia page. We manually mapped the source entity to a Wikipedia page. In addition, we also mapped target categories to the most specific class possible within the DBpedia ontology.[1] We establish ground truth by extracting all Wikipedia pages from the TREC 2009 Entity relevance assessments where any Wikipedia redirects and duplicates are replaced by the target page. Table 1 lists each topic, along with the mapping and the total number of relevant Wikipedia entities.

3.3 Collections

For the IR experiments, we use the official document collection of the TREC Entity track: the ClueWeb09 Category B subset [8], with about 50 million documents (including the English Wikipedia). As to the SW experiments, we query two SPARQL endpoints. The first is the one provided by the DBpedia project[2] which contains a knowledge repository of extracted facts from Wikipedia. DBpedia itself is also linked to more knowledge sources in the LOD cloud. Ontotext provides the other SPARQL endpoint[3] we use and which contains a subset of the LOD cloud, including DBpedia, Freebase, Geonames, UMBEL, Wordnet, and more. We use Jena's ARQ toolkit to query the SPARQL endpoints [15].

4. APPROACH

In this section we describe our approaches to the REF task, using IR and SW techniques. Our goal is to find all relevant entities, but it is not our focus to actually rank them. In other words, we aim to find a set of entities for each

[1] http://dbpedia.org/ontology
[2] http://dbpedia.org/sparql
[3] http://ldsr.ontotext.com/sparql

ID	Source entity (E)	Relation (R)	Type (T)	dbpedia-owl	#rel
1	Blackberry	Carriers that Blackberry makes phones for.	ORG	Company	11
4	Philadelphia, PA	Professional sports teams in Philadelphia.	ORG	SportsTeam	9
5	Medimmune, Inc.	Products of Medimmune, Inc.	PROD	Drug	5
6	Nobel Prize	Organizations that award Nobel prizes.	ORG	Organisation	8
7	Boeing 747	Airlines that currently use Boeing 747 planes.	ORG	Airline	25
8	The King's Singers	CDs released by the King's Singers.	PROD	MusicalWork	-
9	The Beaux Arts Trio	Members of The Beaux Arts Trio.	PER	MusicalArtist	7
10	Indiana University	Campuses of Indiana University.	ORG	EducationalInstitution	9
11	Home Depot Foundation	Donors to the Home Depot Foundation.	ORG	Organisation	4
12	Air Canada	Airlines that Air Canada has code share flights with.	ORG	Airline	11
13	American Veterinary Medical Association	Journals published by the AVMA.	PROD	Magazine	-
14	Bouchercon 2007	Authors awarded an Anthony Award at Bouchercon in 2007.	PER	Writer	3
15	SEC conference	Universities that are members of the SEC conference for football.	ORG	University	10
17	The Food Network	Chefs with a show on the Food Network.	PER	Person	28
18	Jefferson Airplane	Members of the band Jefferson Airplane.	PER	MusicalArtist	16
19	John L. Hennessy	Companies that John Hennessy serves on the board of.	ORG	Company	2
20	Isle of Islay	Scotch whisky distilleries on the island of Islay.	ORG	Company	9

Table 1: Description of the test topics. See Section 3 for further details.

topic that could be considered for ranking in a subsequent processing step and we would like this set to be as complete as possible, with respect to the candidate target entities.

4.1 Information Retrieval

To get a reasonably accurate estimate of what IR methods can achieve on the REF task, we use an aggregation of IR approaches employed at the TREC Entity track. These exhibit a great variety in how they recognize entities in text and calculate their ranking. Yet, at the heart of all scoring methods lies a mechanism for capturing the co-occurrence between source and target entities. A common take on the task was to first gather snippets for the input entity and then extract co-occurring entities from these snippets using a named entity recognizer. Several submissions built heavily on Wikipedia, for example by exploiting outgoing links from the entity's Wikipedia page, by using it to improve named entity recognition, or by making use of Wikipedia categories for entity type detection [5].

The number of entities with a Wikipedia page that are found by any of the 41 TREC runs submitted by participating groups, is shown in Table 1 (#rel). This result set may not be complete, as only the top 10 entities per topic per submission were pooled for assessment, and some Wikipedia pages were not included in the ClueWeb crawl.

4.2 Semantic Web

In order to answer the information needs using semantic web technologies, we follow two approaches. The first is straightforward and transforms each query into a SPARQL query, by instantiating E and T in a template query. For example, for topic #5 "Products of Medimmune, Inc.," the following SPARQL query is issued (the namespaces have been removed to improve readability):

```
SELECT DISTINCT ?m ?r
WHERE {
  ?m rdf:type dbpedia-owl:Drug .
  { ?m ?r dbpedia:MedImmune }
  UNION
  { dbpedia:MedImmune ?r ?m }
}
```

This query returns all items that are of type T and that appear as either the predicate or object of a relation with E. Table 2 shows the results of this example query using the LOD SPARQL endpoint. There is no support within SPARQL for querying structures such as trees or lists or to query transitive relations. Inference rules may be used to specify transitive closures or hierarchal membership relations that can then be queried with SPARQL, but we do not make use of this in our setup. This means that we potentially miss instances that are of a subtype of T.

For the second approach we look for all paths between E and T in the knowledge base. We have implemented an exhaustive search algorithm which recursively traverses all incoming and outgoing relations starting from E, looking for T [11]. When T is found, the path from E is recorded. We limit the depth of the search by setting a maximum step limit, n. Note that such an exhaustive search is more general than the template SPARQL query and will include the same results when $n \geq 2$. Results for example topic #5 using this approach are shown in Figure 1.

5. RESULTS AND DISCUSSION

The issues we set out to explore are (i) to which extent state-of-art semantic web technologies are sufficient to answer real-life information needs and (ii) determining the potential of combining IR with SW technologies to improve the end-to-end performance.

Recall that we consider the task of finding relevant entities given a source entity, target type, and a constraining relation. To start, we limit ourselves to Wiki-/DBpedia; here, IR and SW methods should basically find the same set of entities, since DBpedia is directly extracted from Wikipedia and most SW approaches essentially perform the same functions as co-occurrence based IR methods. One area in which the SW methods might have the upper hand, however, is the explicit labeling of some relations in DBpedia which we cannot obtain using IR methods.

When only considering DBpedia, we observe that for most queries the relations that are returned by both SW methods consist mainly of `wikilink` relations. We find proper (explicit) relations for 7 queries, but only 6 of those correspond to the actual information need; for topic #12 we see non-`wikilink` relations like "foundationPerson" and "parentCompany", but none of these characterize the desired association "shares flight with." A positive example is topic #9, where members stand in "associatedMusicalArtist" relation with the band.

In a way, DBpedia provides "cheap semantics", e.g., in the form of `wikilink` relations (which merely indicate that

?m	?r
dbpedia:Amifostine	dbp-prop:wikilink
dbpedia:Motavizumab	dbp-prop:wikilink
dbpedia:Palivizumab	dbp-prop:wikilink
dbpedia:Blinatumomab	dbp-prop:wikilink
dbpedia:Motavizumab	fb:base.bioventurist.product.developed_by
dbpedia:Palivizumab	fb:base.bioventurist.product.developed_by
dbpedia:Motavizumab	fb:base.bioventurist.science_or_technology_company.products
dbpedia:Palivizumab	fb:base.bioventurist.science_or_technology_company.products

Table 2: SPARQL results for query #5 "Products of Medimmune, Inc." and the LOD SPARQL endpoint.

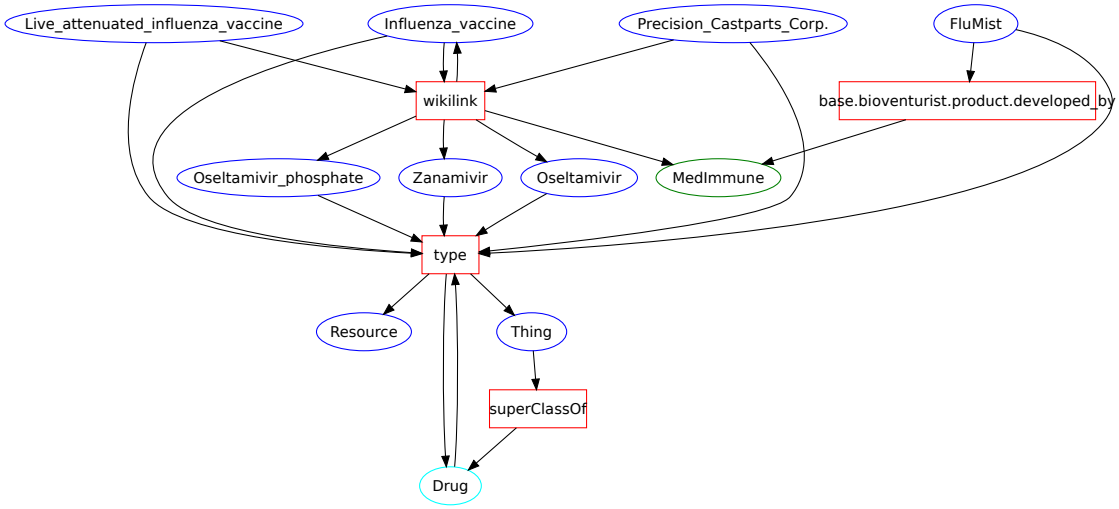

Figure 1: Example output of the graph search algorithm for query #5 on the LOD cloud using $n = 3$. An ellipse indicates an instance, class, or type and a rectangle indicates a relation (duplicate vertices and namespaces have removed to improve readability.) E is indicated using green and T is colored light blue.

there exists a hyperlink from one Wikipedia article to another) and we believe there is much to be gained by automatically generating a classification or label for such links. Furthermore, Wikipedia is essentially user-generated (although heavily moderated). This is witnessed by an uneven distribution of articles in various categories, the fact that not all templates/infoboxes are used consistently, the inconsistent assignment of categories, and the fact that even the number of intra-wiki links can vary greatly from article to article. All of these factors propagate into the knowledge base that DBpedia provides.

When we turn to the LOD cloud, we find that we obtain more entities as well as more diverse relations, which are, indeed, more explicit. For example, Table 2 shows the results for the SPARQL template for query #5. Having more data does not automatically improve results, however. Figure 1 shows the results of running our graph search algorithm on the LOD cloud using $n = 3$, from which we observe that some of the identified entities are now too general.

Another important finding is that most of the retrieved entities are the same as one would obtain by solely looking at the Wikipedia articles. A reason for this is that Wikipedia is a very rich source of entities, with the clear exception of topic #8 ("CDs released by the King's Singers"). While all target entities (CDs) are listed in Wikipedia, none of them have their own Wikipedia page. SW methods can still find all of these when searching in LOD, since Freebase contains a link to these albums. IR methods, if not restricted to Wikipedia, find 13 out of the 42 CDs; these have a primary homepage in the web crawl.

Another interesting topic is #19 ("Companies that John Hennessy serves on the board of."). For this topic, the SW approaches identify 4 relevant companies, whereas the gold standard (generated by aggregating all IR based submissions to the TREC Entity track) only lists 2.

Based on the above cases, we make the following observations. With some manual intervention of mapping queries to entities and classes, the SW has the potential of generating a large number of candidate entities and relations. As such, SW can provide both the data and methods to address entity search. When both entities and relations are present in LOD, answering related entity queries can be as simple as instantiating and executing a SPARQL query (see the King's singers example). However, for many of the queries, we find LOD to be very sparse w.r.t. semantically meaningful links between entities. Exploiting intra-wiki links (which constitute a large portion of all relations) allows us to identify the set of candidate entities, yet, without proper filtering and/or ranking, such large candidate sets are meaningless; IR may provide just those functions.

IR, on the other hand, has excellent ways of finding associations between topics, documents, and entities and one could easily imagine IR models being trained using SW data, e.g., to learn how to recognize entities or relations [12]. Also, IR approaches tend to perform better for less popular entities, which are not represented or connected in LOD, but do occur on the Web. What is lacking in IR, however, is a clear semantics of the found associations and of the obtained entities. Most IR methods merely return most probable or frequent entities and it is here that SW can provide the nec-

essary tools and technology with which to help disambiguate and add semantic anchors to the candidate entities. Interestingly, such semantic contributions to IR were already witnessed at some submissions of the TREC Entity track, where several teams used Wikipedia as a semantic backbone. More specifically, when providing a repository of entity names and name variants for entity recognition and normalization, type detection and filtering, and for finding official homepages of entities. Nevertheless, robust approaches capable of bridging the semantic gap between query terms and terms observed around co-occurring entities are yet to come [21].

6. CONCLUSIONS

Entity retrieval, the task of finding objects related to a particular information need is an emerging research topic. It is being addressed by both the Information Retrieval (IR) and Semantic Web (SW) communities, although they use different instruments and resources. We focus on a specific task, related entity finding, as defined at the recently launched TREC Entity track. We explore to which extent this task can be solved using state-of-the-art IR and SW approaches.

Results of a small-scale study indicate that using SW methods on top of Linked Open Data (LOD) can answer related entity queries. While a large proportion of entities is present in LOD, for most queries links between them consist mainly of `wikilink` relations. The semantics of such a link are not very well defined, and additional filtering/ranking steps are required on top of the result sets, which IR can provide. Also, IR methods were found to perform better when searching for less popular entities, that are not represented in LOD. What is missing here, however, is a semantic labeling of the associations found, which SW can provide.

There is clearly room for improvement on both sides. To get the best of both worlds, we propose to combine text-based entity models with semantic information from the LOD cloud; LOD can provide training material for associating language usage around entities with semantically meaningful types and relations. In turn, IR models can be used to discover links between entities in the LOD cloud.

Acknowledgements This research was supported by the Center for Creation, Content and Technology (CCCT), the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.

7. REFERENCES

[1] K. Anyanwu, A. Maduko, and A. Sheth. Semrank: ranking complex relationship search results on the semantic web. In *14th Intern. Conf. on World Wide Web*, 2005.

[2] S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. In *4th European Conf. on The Semantic Web: Research and Applications*, 2007.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *6th Intern. Semantic Web Conf., 2nd Asian Semantic Web Conf. (ISWC+ASWC 2007)*, 2007.

[4] K. Balog. *People Search in the Enterprise*. PhD thesis, University of Amsterdam, June 2008.

[5] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *18th Text REtrieval Conf.*, 2010.

[6] C. Bizer, R. Cyganiak, S. Auer, and G. Kobilarov. DBpedia–querying Wikipedia like a database. 16th Intern. Conf. on World Wide Web – Developers track presentation, 2007.

[7] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *SIGIR '09*, 2009.

[8] ClueWeb09. The ClueWeb09 dataset, 2009. URL: `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

[9] N. Craswell, A. De Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *14th Text REtrieval Conf. (TREC 2005)*, 2006.

[10] A. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In *Focused access to XML documents: 6th Intern. Workshop of the Initiative for the Evaluation of XML Retrieval*, 2008.

[11] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959.

[12] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. P. Sheth. Context and domain knowledge enhanced entity spotting in informal text. In *Intern. Semantic Web Conf.*, 2009.

[13] A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. In *Intern. Semantic Web Conf.*, 2009.

[14] B. J. Jansen and A. Spink. How are we searching the world wide web? a comparison of nine search engine transaction logs. *Inf. Processing & Management*, 42(1):248 – 263, 2006.

[15] Jena. ARQ - a SPARQL processor for Jena, 2010. URL: `http://jena.sourceforge.net/ARQ/`.

[16] S. Kulkarni and D. Caragea. Towards bridging the web and the semantic web. In *Web Intelligence*, 2009.

[17] J. Lehmann, J. Schüppel, and S. Auer. Discovering unknown connections - the dbpedia relationship finder. In *1st SABRE Conf. on Social Semantic Web*, 2007.

[18] C. Mangold. A survey and classification of semantic search approaches. *Int. J. Metadata Semant. Ontologies*, 2(1):23–34, 2007.

[19] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Conceptual language models for domain-specific retrieval. *Information Processing & Management*, In Press.

[20] E. J. Meij, M. Bron, B. Huurnink, L. Hollink, and M. de Rijke. Learning semantic query suggestions. In *ISWC '09: 8th Intern. Conf. on The Semantic Web*, 2009.

[21] P. Mika, E. Meij, and H. Zaragoza. Investigating the semantic gap through query log analysis. In *Intern. Semantic Web Conf.*, 2009.

[22] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07*, 2007.

[23] H. Raghavan, J. Allan, and A. Mccallum. An exploration of entity models, collective classification and relation description. In *ACM SIGKDD Workshop on Link Analysis and Group Detection (LinkKDD2004)*, 2004.

[24] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *16th Intern. Conf. on World Wide Web*, 2007.

[25] E. Voorhees. Overview of the TREC 2004 question answering track. In *13th Text Retrieval Conf.*, 2005.