Entity Ranking Using Wikipedia as a Pivot

Rianne Kaptein[1] kaptein@uva.nl
Pavel Serdyukov[2] p.serdyukov@tudelft.nl
Arjen de Vries[2,3] arjen@acm.org
Jaap Kamps[1,4] kamps@uva.nl

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam, The Netherlands
[2] Delft University of Technology, The Netherlands
[3] Centrum Wiskunde & Informatica, The Netherlands
[4] ISLA, Informatics Institute, University of Amsterdam, The Netherlands

ABSTRACT

In this paper we investigate the task of Entity Ranking on the Web. Searchers looking for entities are arguably better served by presenting a ranked list of entities directly, rather than a list of web pages with relevant but also potentially redundant information about these entities. Since entities are represented by their web homepages, a naive approach to entity ranking is to use standard text retrieval. Our experimental results clearly demonstrate that text retrieval is effective at finding relevant pages, but performs poorly at finding entities. Our proposal is to use Wikipedia as a pivot for finding entities on the Web, allowing us to reduce the hard web entity ranking problem to easier problem of Wikipedia entity ranking. Wikipedia allows us to properly identify entities and some of their characteristics, and Wikipedia's elaborate category structure allows us to get a handle on the entity's type.

Our main findings are the following. Our first finding is that, in principle, the problem of web entity ranking can be reduced to Wikipedia entity ranking. We found that the majority of entity ranking topics in our test collections can be answered using Wikipedia, and that with high precision relevant web entities corresponding to the Wikipedia entities can be found using Wikipedia's 'external links'. Our second finding is that we can exploit the structure of Wikipedia to improve entity ranking effectiveness. Entity types are valuable retrieval cues in Wikipedia. Automatically assigned entity types are effective, and almost as good as manually assigned types. Our third finding is that web entity retrieval can be significantly improved by using Wikipedia as a pivot. Both Wikipedia's external links and the enriched Wikipedia entities with additional links to homepages are significantly better at finding primary web homepages than anchor text retrieval, which in turn significantly improved over standard text retrieval.

Categories and Subject Descriptors:
H.3.4 **[Information Storage and Retrieval]**: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms:Experimentation, Measurement, Performance

Keywords: Web Entity Ranking, Wikipedia

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

Entity ranking is the task of finding documents representing entities of a correct type that are relevant to a query. Searchers looking for entities are arguably better served by presenting a ranked list of entities directly, rather than a list of web pages with relevant but also potentially redundant information about these entities. Search engines have started to develop special services for entity retrieval, e.g., Google Squared[1] and the Yahoo Correlator[2]. It is difficult to quantify which part of web searches are actually entity ranking queries. It is known however that a considerable fraction of web searches contains named entities [e.g., 21].

Just like in document retrieval, in entity ranking the document should contain topically relevant information. However, it differs from document retrieval on at least three points: i) returned documents have to represent an entity, ii) this entity should belong to a specified entity type, and iii) to create a diverse result list an entity should only be returned once. The main goal of this paper is to demonstrate how the difficult problem of web entity ranking can often be reduced to the easier task of entity ranking in Wikipedia.

To be able to do web entity ranking, we need to extract structured information, i.e. does this page represent an entity, and of what type is this entity, from the unstructured web. One approach to use structure is to add structure to unstructured web pages, for example by tagging named entities. Another approach would derive implicit structure from the link structure on the web, using links and anchor text. On the web, it is however not so easy to define, identify and represent entities. Just returning the name of an entity will not satisfy users, they need to see some kind of proof that this entity is indeed relevant, and secondly, they may want to know more of the entity than just its name. Depending on the type of entity that we are looking for these problems can be more or less significant. Entities can be represented by many webpages, e.g. an "official" homepage, a fan page, a page in an online encyclopedia or database like Wikipedia, Amazon or IMDB, or the entry in a social network such as Facebook, Twitter, MySpace. A complete representation or profile of a web entity would consist of many pages. The goal of entity ranking however is not to find all pages related to one result entity, but to find all relevant entities which can then be represented by one well-chosen page.

What type of page can be considered representative depends on the entity type, or even the entity itself – in the absence of an "official" homepage for example, alternatives might need to be considered. What would for example be the homepage of a historical person, or a chemical element? The major search engines can give

[1] http://www.google.com/squared
[2] http://sandbox.yahoo.com/correlator

us some clues which pages are appropriate; for movies and actors IMDB pages are among the top results, for well-known people it is often a Wikipedia page, and for companies their official website. Following the TREC 2009 entity ranking track, we will represent entities by their "official" homepage or their Wikipedia page. The latter is useful for entity types where no "official" homepage exists.

Instead of structuring the web ourselves, we propose to exploit the part of the web that is *manually* structured: in this paper, we limit our scope to Wikipedia. Wikipedia is an excellent structured resource of entities; its structure can be used as follows. Entities are Wikipedia pages, where the name of the entity is the title of the page, the content of the page is the representation of the entity. Each Wikipedia page is assigned to a number of categories. These categories can be divided into topical, type, and administrative categories. Administrative categories such as "Pages needing to be revised" are merely there for administrative purposes. Topical categories such as "Barack Obama" indicate that the page is related to this topic. The categories we are interested in are the type categories such as "People from Westminster" or "Museums in Michigan". These categories give us information about the entity type. Since Wikipedia is an encyclopedia each entity is only represented once and we do not have to worry about returning duplicate entities.

Our proposal is to exploit Wikipedia as a pivot for entity ranking. For entity types with a clear representation on the web, like living persons, organisations, products, movies, we will show that Wikipedia pages contain enough evidence to reliably find the corresponding web page of the entity. For entity types that do not have a clear representation on the web, returning Wikipedia pages is in itself a good alternative. So, to rank (web) entities given a query we take the following steps:

1. Associate target entity types with the query
2. Rank Wikipedia pages according to their similarity with the query and target entity types
3. Find web entities corresponding to the Wikipedia entities

First of all, we investigate whether the Web entity ranking task can indeed be effectively reduced to the Wikipedia entity ranking task. Therefore, we have to answer the following two research questions:

- What is the range of entity ranking topics which can be answered using Wikipedia?
- When we find relevant Wikipedia entities, can we find the relevant web entities that correspond to the Wikipedia entities?

We use the results of the TREC 2009 Entity Ranking Track (based on the Web including Wikipedia) and the INEX 2009 Entity Ranking Track (based on Wikipedia). We extend the INEX topics to the Web to answer these research questions.

The second step of our approach corresponds directly to the setup of the INEX entity ranking track, so we adopt a competitive approach from this track for our experiments. However, the INEX setup assumed detailed knowledge of entity target type. Although users might be able and/or willing to indicate a general target entity type along with their query, e.g., choosing from people, organisations, products, we cannot realistically expect users to provide accurate Wikipedia categories; thousands of these categories exist, and they are only very loosely organised. So, we investigate the following two issues related to the second step of our approach:

- Can we exploit category information to improve entity ranking queries?
- Can we automatically assign entity types to natural language queries?

Finally, we evaluate our complete entity ranking approach and compare it to alternative approaches that do not use Wikipedia to answer the questions:

- Can we improve web entity ranking by using Wikipedia as a pivot?
- Can we automatically enrich Wikipedia with additional links to homepages of found entities?

The paper is structured as follows. The next section discusses related work on entity ranking. Section 3 analyzes the relations between entities in Wikipedia and entities on the web. Section 4 then examines entity ranking on Wikipedia, seeking to exploit different levels of entity types. In Section 5 we focus on Web entity ranking proper with and without the use of Wikipedia as a pivot. Finally, in Section 6 we draw our conclusions.

2. RELATED WORK

Entity ranking has recently become a popular new task. It started out with ranking entities of a specific type, for example persons in expert search [2]. The more general problem is to rank all kinds of entities, e.g. persons, locations, organizations etc.

When working with different types of entities, often some mechanism is needed to recognize and classify entities. A framework to identify persons and organizations is introduced in [7]. Besides extracting entities they also try to determine relationships between them. Named entity taggers such as [14, 15] have been developed to extract entities of different types from documents and are publicly available.

Little work has been done on classifying entity types of queries automatically. Instead of finding the category of the query, the approach described in [27] seeks to find the most important general entity types such as locations, persons and organizations. Their approach executes a query and extracts entities from the top ranked result passages. The entity type that can be associated with most of these extracted entities is assigned to the query. The majority of queries can be classified correctly into three top entity types.

Besides ranking entities, entities can be used to support many other tasks as well. Entity models of entities are built and clustered in [23]. A semantic approach to suggesting query completions, which leverages entity and entity type information is proposed in [20]. A formal method for explicitly modeling the dependency between the named entities and terms which appear in a document is proposed in [22], and applied to an expert search task.

Several search engines provide the possibility of ranking entities of different types. The semantic search engine NAGA for example builds on a knowledge base that consists of millions of entities and relationships extracted from Web-based corpora [18]. A graph-based query language enables the formulation of queries with additional semantic information such as entity types. The search engine ESTER combines full-text on Wikipedia with ontology search in YAGO [6]. The interactive search interface suggests to the user possible semantic interpretations of his/her query, thereby blending entity ranking and ad hoc retrieval.

Wikipedia is used as a resource to identify a number of candidate entities in [30]. A statistical entity extractor identified 5,5 million entities in Wikipedia and a retrieval index was created containing both text and the identified entities. Different graph centrality measures are used to rank entities in an entity containment graph. Also

a web search based method is used to rank entities. Here, query-to-entity correlation measures are computed using page counts returned by search engines for the entity, query and their conjunction. Their approaches are evaluated on a self-constructed test collection. Both their approaches outperform methods based on passage retrieval.

A lot of entity ranking research has recently been done in context of the INEX and TREC evaluation fora. INEX has run an entity ranking track since 2006, using Wikipedia as the test collection [8, 11]. The INEX entity ranking track is set up as follows. The document collection is a snapshot of the English Wikipedia. For the tracks from 2006 to 2008 a snapshot from Wikipedia from early 2006 containing 659,338 articles was used [12]. Since then Wikipedia has significantly grown, and for the 2009 track a new snapshot of the collection is used. It is extracted in October 2008 and consists of 2.7 million articles [24]. A query topic consists of a keyword query and one or a few target categories which are the desired entity types. A description and narrative are added to clarify the query intent. A topic looks as follows:

```
<inex_topic_id="9999">
<title> Impressionist art in the Netherlands
</title>
<description> I want a list of art galleries and
museums in the Netherlands that have impressionist
art.
</description>
<narrative> Each answer should be the article about
a specific art gallery or museum that contain
impressionist or post-impressionist art works.
</narrative>
<categories>
<category> art museums and galleries
</category>
</categories>
```

Because the Wikipedia category structure is hierarchical and not applied consistently, relevant result entities do not always belong to one of the specified target categories. A result entity is only considered relevant if it belongs to a category similar or equal to one of the target categories.

Several approaches have been quite successful in exploiting category information. Wikipedia categories are used by defining similarity functions between the categories of retrieved entities and the target categories. The similarity scores are estimated based on the ratio of common categories between the set of categories associated with the target categories and the union of the categories associated with the candidate entities [29] or by using lexical similarity of category names [28]. Random walks to model multi-step relevance propagation from the articles describing entities to all related entities and further are used in [26]. After relevance propagation, the entities that do not belong to a set of allowed categories are filtered out the result list. The allowed category set leading to the best results included the target categories with their child categories up to the third level. A probabilistic framework to rank entities based on the language modelling approach is presented in [3]. Their model takes into account for example the probability of a category occurrence and allows for category-based feedback. Finally, in addition to exploiting Wikipedia structure i.e. page links and categories, [9] applies natural language processing techniques to improve entity retrieval. Lexical expressions, key concepts, and named entities are extracted from the query, and terms are expanded by means of synonyms or related words to entities corresponding to spelling variants of their attributes.

TREC introduced the Entity Ranking track in 2009 [5]. It makes use of the Clueweb collection Category B, which consists of about 50 million English-language web pages including the complete Wikipedia. The task in this track was an entity relationship search task: given an entity (name and document id) and a narrative, find the related relevant entities. A query topic looks as follows:

```
<query>
<num>1</num>
<entity_name>Blackberry</entity_name>
<entity_URL>clueweb09-en0004-50-39593
</entity_URL>
<target_entity>organization</target_entity>
<narrative>Carriers that Blackberry makes phones
 for.</narrative>
</query>
```

Three entity types are used in 20 topics, 6 topics are looking for persons, 11 topics for organizations, and 3 topics for products. Although this test collection is relatively small, it is the best data available to study our research questions. We will supplement the results with results on other data wherever possible, in particular we have extended the INEX 2009 Entity Ranking track data to the web [10].

TREC participants have approached the task in two main steps. First, candidate entity names are extracted, using entity repositories such as Wikipedia, or using named entity recognizers. Link information of the given entity can be used to make a first selection of documents. In a second step, candidate entity names are ranked, and primary homepages retrieved for the top ranked entity names. The University of Glasgow method builds entity profiles for a large dictionary of entity names using DBPedia and common proper names derived from US Census data [19]. At query time, a voting model considers the co-occurrences of query terms and entities within a document as a vote for the relationship between these entities. Purdue University expands the query with acronyms or the full name of the source entity [13]. Candidate entities are selected from top retrieved documents, heuristic rules are applied to refine the ranking of entities.

3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

In this section, we investigate our first group of research questions. What is the range of entity ranking topics which can be answered using Wikipedia? When we find relevant Wikipedia entities, can we find the relevant web entities that correspond to the Wikipedia entities?

3.1 From Web to Wikipedia

While the advantages of using Wikipedia or any other encyclopedic repository for finding entities are evident, there are still two open questions: whether these repositories provide enough clues to find the corresponding entities on the Web and whether they contain enough entities that cover the complete range of entities needed to satisfy all kinds of information needs. The answer to the latter question is obviously "no". In spite of the fact that Wikipedia is by far the largest encyclopedia in English—it contains 3,147,000 articles after only 9 years of existence; the second largest, Encyclopaedia Britannica, contains only around 120,000 articles—Wikipedia is still growing, with about 39,000 new articles per month in 2009 [3]). We can therefore only expect that it has not yet reached its limit as a tool for entity ranking. One of the most important factors imped-

[3] en.wikipedia.org/wiki/Size_of_Wikipedia

Table 1: Topic and Entity Coverage in Wikipedia

# Topics	20	
- with entities in Wikipedia	17	(85%)
# Entities	198	
- with Wikipedia pages	160	(81%)

ing the growth of Wikipedia and also interfering with its potential to answer all kinds of queries looking for entities is the criterion of notability used by editors to decide whether a particular entity is worthy of an article. There are general and domain specific notability guidelines[4] for entities such as people, organizations, events, etc. They are based on the principle of significant coverage in reliable secondary sources and help to control the flow of valuable and potentially popular topics into Wikipedia. However, the desire of the Wiki community to have also repositories for the entities of lesser importance led to establishing side projects, like Wikicompany (≈3,200 articles about organizations), Wikispecies (≈150,000 articles about all species of life) or CDWiki (≈500,000 articles about audio CDs).

In order to study how far we can go with Wikipedia only when looking for entities, we analyzed the list of relevant entities for 20 queries used in Entity ranking track at TREC 2009, see Table 1. We found that 160 out of 198 relevant entities have a Wikipedia page among their primary pages, while only 108 of them have a primary web page (70 entities have both). As not all primary Wikipedia pages are returned by participants and judged, or Wikipedia pages might have not existed yet when the ClueWeb collection was crawled (January/February 2009), we manually searched online Wikipedia (accessed in December 2009) for primary Wikipedia pages for the 38 entites that had only primary web pages. As a result, we discovered primary Wikipedia pages for a further 22 entities. Those 16 entities that are not represented in Wikipedia are seemingly not notable enough. However, they include all answers for 3 of 20 queries (looking for audio cds, phd students and journals). Although the numbers of topics is small, the percentage of pages and topics that are covered by Wikipedia is promising. Topics can have no primary Wikipedia entities because no participant found relevant entities, or they were not judged. For some topics however, no primary entities will exist in Wikipedia, due to its encyclopedic nature. For example no relevant entities for the topic 'Students of Claire Cardie' will appear in Wikipedia, unless one of these students becomes famous in some way, and meets the requirements to be included in Wikipedia. To cover this gap, other databases can be used; e.g., it has already been shown that US Census data can be used to derive common variants of proper names to improve web entity ranking [19].

3.2 From Wikipedia to Web

After we found that there is a strong link from entities represented on the Web (so, notable to a certain extent) to Wikipedia, it was further important to find out whether the opposite relation also exists. If it does, it would prove that Wikipedia has the potential to safely guide a user searching for entities through the Web and serve as a viable alternative to a purely web-based search, considering the immense size of the Web and the amount of spam it contains. Again, thanks to the Wikipedia community, those articles that follow the official guidelines are supposed to have an "External links" section, where the web pages relevant to the entity should be enlisted. Moreover, it is stated that "articles about any organization, person, website, or other entity should link to the subject's official site" and "by convention are listed first"[5]. In our case, 141 primary Wikipedia pages out of 160 (≈88%) describing relevant entities had the "External links" section. Actually, only 4 out of 19 entities described by Wikipedia pages with no "External links" section had also the corresponding primary Web pages, what can be explained by the fact that Wikipedia pages often serve as the only "official" pages for many entities (e.g. historical objects or non-living people).

In order to be sure that it is easy to discover a primary Web page by looking at these external links, we also analyzed how many of these links point to primary Web pages for the same entities.

In addition to the TREC entity ranking topics, we use INEX 2009 Entity Ranking topics. The topic set consists of 55 entity ranking topics, and each topic has at least 7 relevant entities. We have mapped the relevant wikipedia pages from the INEX Wikipedia collection to the Clueweb collection by matching on the page title and found matches for 1,381 out of the 1,665 relevant pages. Differences occur because the INEX Wikipedia collection is extracted from a dump in October 2008, while the TREC Wikipedia collection is crawled in January and February 2009. All links from relevant Wikipedia pages to pages in Clueweb (Category B) are judged by the authors of this paper. The difference between the TREC topics and the INEX topics is that the TREC topics are restricted to the entity types person, organization and product, while the INEX topics can be virtually any entity type. The TREC guidelines define a primary homepage as devoted to and in control of the entity. For the entity types that cannot control a homepage, e.g. deceased persons or concepts like chemical elements, we take the second best thing: an authorative homepage devoted to the entity. For some of these entity types the Wikipedia page could in fact be considered the best primary page.

Unfortunately, not all web-sites linked from Wikipedia are included in the TREC ClueWeb collection (Category B). For the TREC topics 98 out of 141 primary Wikipedia pages had at least one linked web-site in the collection and only 60 of them described entities for which a primary Web page was found as well. At the same time, in 52 of these cases (≈87%) at least one primary Web page was linked from the corresponding Wikipedia page. Moreover, in 4 out of the 8 unsuccessful cases another page from the primary web page's domain was linked. In the case, when we considered only the first external link in the list, 43 of 46 links pointing to an existing page in the collection actually pointed to the primary Web page of the respective entity.

Looking at the INEX topics we find comparable numbers, but on a larger scale. Most relevant Wikipedia pages have external links (72%), but only a relatively small number of these external links point to pages in the Clueweb category B collection, i.e for 289 pages a total of 517 external links are found. Compared to the TREC topics, for INEX topics a smaller percentage of the external links are indeed relevant primary pages, of all external links 37% are relevant, of the first external links a respectable 77% of the pages is relevant. Comparing the TREC and the INEX topics, we see that the relevance of all external links is much higher for the TREC topics than for the INEX topics, and the relevance of the first links is also lower for the INEX topics. The TREC topics contain only 14 links below rank one that are judged, so we cannot really say much here about the relevance of links below rank one. The INEX topics however are more substantial, and present a clear difference between the first external link, and the lower ranked links. Out of the 361 links below rank one, only 69 are deemed relevant.

[4] en.wikipedia.org/wiki/Wikipedia:Notability

[5] en.wikipedia.org/wiki/Wikipedia:External_links

Table 2: External Link and Assessment Statistics

Topic Set	TREC 2009		INEX 2009	
# Rel. Wiki. pages	160		1381	
- with external links	141	(88%)	994	(72%)
- with external Clueweb links	88	(55%)	289	(21%)
# Judged ext. links	60		517	
- relevant links	52	(87%)	189	(37%)
# Judged first ext. links	46		156	
- relevant first links	43	(93%)	120	(77%)

Most of these relevant links are found for entities which have indeed more than one primary homepage, for example organisations that link to several corporate homepages for different regions.

Furthermore, the TREC topics are designed to have at least some primary homepages in the Clueweb Category B collection, otherwise the topic wouldn't have made it into the test set. Also the entity types restriction to products, persons and organisations is making these topics more likely to have easily identifiable primary homepages. For the less restricted INEX topics primary homepages are harder to find, moreover these pages might not be considered entities by the Wikipedia editors, which alleviates their need to link to a primary homepage.

To validate that primary web pages would not be so easily discovered without the Wikipedia "External links" section, we first measured Mean Reciprocal Rank (MRR) of the first primary web page which we find using the ranking naturally provided in the "External links" section. We also measured MRR for the ranking which we get by using entity names as queries to search anchor text index built for ClueWeb collection (category B). We experimented with 60 entities from the TREC topics that have a Wikipedia page, at least one primary Web page and at least one linked web-site existing in the ClueWeb collection. Indeed, using "External links" is much more effective for primary web page finding ($MRR = 0.768$) than using an anchor text index ($MRR = 0.442$).

In this section, we investigated whether the hard problem of web entity ranking can be in principle reduced to the easier problem of Wikipedia entity ranking. We found that the overwelming majority of relevant entities of the TREC 2009 Entity ranking track are represented in Wikipedia, and that 85% of the topics have at least one Wikipedia primary page.

We also found that with high precision and coverage relevant web entities corresponding to the Wikipedia entities can be found using Wikipedia's "external links", and that especially the first external link is a strong indicator for primary homepages.

4. ENTITY RANKING ON WIKIPEDIA

In this section we investigate our second group of research questions. Can we exploit category information to improve entity ranking queries? Can we automatically assign entity types to natural language queries? We conduct a range of experiments with entity ranking in Wikipedia, seeking to exploit entity type information.

4.1 Entity Types

4.1.1 Entity Type Assignment

In the TREC and INEX entity ranking tracks entity types are assigned by the topic creators. In practice however, it has proven difficult to convince users to submit more than a few keywords as a query. Common web users hardly ever use even simple structured queries. It is therefore an unlikely user scenario that a user will come up with a keyword query and a specific targeted entity type. So, we will examine whether we can assign an entity type to a query automatically.

There are many ways to automatically categorize topics, for example by building language models of each entity type and calculating KL-divergence between the query and/or top retrieved results. Here, we keep it simple and exploit the existing Wikipedia categorization of documents. Pseudo-relevance feedback of the top retrieved documents is used, but instead of extracting the most frequently occurring terms from the top ranked documents as is done in standard pseudo-relevance feedback, we extract the categories that are most frequently assigned. From our baseline run, we take the top 10 results, and look at the 2 most frequently occurring categories belonging to these documents. Categories that occur only once are excluded. These parameter settings lead to good results in previous similar experiments [17]. The categories are assigned as target entity types to the query topic. This entity type assignment method will lead to specific entity types, since these are the categories that are assigned to pages. More general categories are more loosely connected to the pages. Due to the category structure of Wikipedia, which is an undirected graph, rather than a tree, it is difficult to use the hierarchical structure to assign general entity types.

4.1.2 Scoring Entities

To exploit entity type information we calculate for the top ranked documents in our initial ranking an entity type score, using a language modeling approach [16]. The initial ranking is based solely on the likelihood of the query terms occurring in the document. This probability is calculated using a language model with Jelinek-Mercer smoothing with uniformly distributed prior document probabilities:

$$P(q_1, ..., q_n|d) = \sum_{i=1}^{n} \lambda P(q_i|d) + (1-\lambda)P(q_i|D) \quad (1)$$

where $q_1, ..., q_n$ are the query terms, d is the document, and D is the entire Wikipedia document collection, which is used to estimate background probabilities.

The entity type score corresponds to the similarity between the target entity types and the document entity type. Entity types are represented by the names of Wikipedia categories. To calculate entity type scores, first of all we make a maximum likelihood estimation of the probability of a term occurring in a category name. To avoid a division by zero, we smooth the probabilities of a term occurring in a category name with the background collection:

$$P(t_1, ..., t_n|C) = \sum_{i=1}^{n} \lambda P(t_i|C) + (1-\lambda)P(t_i|D) \quad (2)$$

where $t_1, ..., t_n$ are the terms in C, the name of the category.

To calculate the similarity between two categories we use KL-divergence as follows:

$$S_{cat}(C_t|C_d) = -D_{KL}(C_t|C_d) \quad (3)$$

$$= -\sum_{t \in C_t} \left(P(t|C_t) * \log \left(\frac{P(t|C_t)}{P(t|C_d)} \right) \right) \quad (4)$$

where d is a document, i.e. an answer entity, C_t is a target category and C_d a category assigned to a document . The entity type score for a document in relation to a query topic ($S_{cat}(d|QT)$) is the maximum of the scores of all target and document categories:

$$S_{cat}(d|QT) = \arg\max_{C_t \in QT} \arg\max_{C_d \in d} S_{cat}(C_t|C_d) \quad (5)$$

We use a standard method for score normalization that takes the standard deviation of score into account, the Z-score. Scores are normalized to the number of standard deviations that are higher (or lower) than the mean score. The mean and standard deviation depend on the length of the ranking. It was shown in [1] that this is a simple, yet effective method to normalize retrieval scores.

Finally, a linear combination of the normalized scores is made to calculate the final score:

$$S(d|QT) = \mu P(q|d) + (1-\mu) S_{cat}(d|QT) \quad (6)$$

4.2 Experimental Setup

In this section we investigate the effects of using manually assigned versus automatically assigned entity types. We use the Indri search engine [25] for our experiments. We have created index of the Wikipedia test collection of INEX applying the Krovetz stemmer. Our baseline model is a language model using Jelinek-Mercer smoothing with a collection λ of 0.15.

We use topic sets from TREC and INEX. Entity types can be defined on many levels, from general types such as 'person' or 'organisation' to more specific types such as 'Olympic medalists' or 'shoe shops'. When entity ranking is restricted to few general entity types, specific rankers for entity types could be designed. To rank people for instance, people-specific attributes and models could be used [4]. We would however prefer a generic approach that is effective for all types of entities. The entity types of the INEX entity ranking track are quite specific. Some examples of entity types are countries, national parks, baseball players, and science fiction books. The TREC entity ranking track uses only three general entity types, i.e. people, organisations, and products. The advantages of these entity types are that they are clear, there are few options and could be easily selected by users. The disadvantage is that they only cover a small part of all possible entity ranking queries. To make our test set more consistent we manually assigned more specific entity types to the TREC entity ranking topics so that they are on the same level as the INEX entity types.

Another difference between the tracks is that the TREC entity ranking task was entity relationship search, i.e. answer entities should be related to a given entity. For this paper we do not use the given website of the entity, but we add the entity name to the narrative. Together the entity name and the narrative serve as our keyword query. By not using the given entity, we can consider this task as an entity ranking task.

Entity type information is used as described in section 4.1.2. We rerank the top 2,500 results of the baseline run using two sets of entity type information:

- Manually assigned: assigned manually by the authors for the TREC topics, and assigned by the topic creators for the INEX topics
- Automatically assigned: assigned by pseudo-relevance feedback on entity types of the baseline run

For evaluation we look at P10 and NDCG for the 20 2009 TREC topics and P10 and MAP for INEX topics. We use INEX topics 2006-2008 consisting of 79 topics and INEX 2009 topics, consisting of a selection of 55 topics from the 2006-2008 topics. The main difference between the INEX runs is the version of the Wikipedia collection. For the evaluation, we count only the pages that are relevant and of the correct entity type. For the TREC topics, this means we only count the so-called 'primary' pages, i.e. authoritative or official homepages of an entity. Pages that contain relevant information, but are not primary homepages are judged as 'relevant' pages in the official qrels. We are only interested in the primary pages, which represent an answer entity, and not in the relevant pages, so we only give credit to the primary pages. We remove redirected Wikipedia pages from our runs and from the assessments, and replace them where possible with the correct, non-redirected page.

Table 3: Wikipedia retrieval results on TREC topics

Cats	μ	$\#Rel$	P10	NDCG
Baseline	1	78	0.1200	0.0797
Auto.	0.7	74	0.1500	0.0980
Auto.	0.8	75	0.1500	0.0971
Auto.	0.9	79	0.1500	0.0969
Man.	0.4	89	**0.1750**	0.1123°
Man.	0.5	91	**0.1750**	**0.1193**$^{\circ}$
Man.	0.8	**96**$^{\circledcirc}$	0.1700°	0.1132°

Significance of increase or decrease over baseline according to t-test, one-tailed, at significance levels 0.05($^{\circ}$), 0.01($^{\circledcirc}$), and 0.001($^{\bullet}$).

Table 4: Wikipedia retrieval results on INEX 2006-2008 topics

Cats	μ	$\#Rel$	P10	MAP
Baseline	1	1142	0.2405	0.1948
Auto.	0.7	1239°	0.2949$^{\circledcirc}$	0.2602$^{\bullet}$
Auto.	0.8	1279$^{\circledcirc}$	0.2987$^{\circledcirc}$	0.2686$^{\bullet}$
Auto.	0.9	1289$^{\bullet}$	0.2937$^{\circledcirc}$	0.2561$^{\bullet}$
Man.	0.7	1346$^{\bullet}$	**0.3797**$^{\bullet}$	**0.3245**$^{\bullet}$
Man.	0.8	**1361**$^{\bullet}$	0.3620$^{\bullet}$	0.3048$^{\bullet}$
Man.	0.9	1327$^{\bullet}$	0.3241$^{\bullet}$	0.2711$^{\bullet}$

Table 5: Wikipedia retrieval results on INEX 2009 topics

Cats	μ	$\#Rel$	P10	MAP
Baseline	1	1042	0.2164	0.1674
Auto.	0.8	911°	0.2382	0.1993°
Auto.	0.9	982	0.2509	0.2014$^{\circledcirc}$
Man.	0.6	1171$^{\circledcirc}$	**0.3145**$^{\bullet}$	0.2376$^{\bullet}$
Man.	0.7	1178$^{\circledcirc}$	0.3127$^{\bullet}$	**0.2396**$^{\bullet}$
Man.	0.9	**1180**$^{\bullet}$	0.2982$^{\bullet}$	0.2350$^{\bullet}$

4.3 Experimental Results

The results of our experiments expressed in the number of retrieved relevant pages, P10 and NDCG@R or MAP are summarized in Tables 3, 4 and 5. These show the results of our baseline run and different entity type sets over different values of μ, where μ is the weight of the query score and $(1-\mu)$ is the weight of the KL-divergence category score.

The runs with automatically assigned entity types reach a performance close to the manually assigned topics. Although $P10$ is low in the baseline run, the 10 top ranked documents do provide helpful information on entity types. Most of the automatic assigned categories are very specific, for example 'College athletics conferences' and 'American mystery writers'. For one topic the category exactly fits the query topic, the category 'Jefferson Airplane members' covers exactly query topic 'Members of the band Jefferson Airplane'. Unsurprisingly, using this category boosts performance significantly. The category 'Living people' is assigned to several of the query topics that originally also were assigned entity type 'Persons'. This category is one of the most frequently occurring categories in Wikipedia, and is assigned very consistently to pages about persons. In the collection there are more than 400,000 pages

that belong to this category. This large number of occurrences however does not make it a less useful category.

The relative improvements from using category information are similar in the two INEX runs, as well as in the TREC run, with maximum respective improvements in P10 of 46%, 58% and 45% for Tables 3, 4 and 5 respectively. Our INEX runs with the manual assigned categories achieve performance similar to the state of the art INEX entity ranking approaches.

The score for the TREC topics are much lower than the scores for the INEX topics. There are several explanations for these low scores. First of all, the TREC runs contain a lot of unjudged pages, since none of these runs were official submissions to the track and they are not in the assessment pool. Only around half the pages in the top 10 are judged, over all results around 15 to 20% of the pages are judged. The results on the TREC topics are an underestimation of the performance of our approach. Results closer to the INEX runs should be achievable. More judging, or different evaluation measures are needed to get more reliable estimations of performance. The second reason for the low scores could be the nature of the original task, which was entity relationship search. By not using the given entity, we lose information that could help. For example, the outgoing links of the example entity can be used to generate a set of candidate documents.

Since the Wikipedia collection used in INEX 2009 is very similar to the Clueweb Wikipedia pages, scores comparable to the INEX 2009 topics should be achievable. The new Wikipedia collection is a lot larger than the old collection, and although the quality of the pages improves, the collection becomes less focused.

In this section we examined the value of entity type information for entity retrieval in Wikipedia. We found that entity types are valuable retrieval cues. Automatically assigned entity types are effective, but less so than the manually assigned types. The general conclusion is that we can exploit the structure of Wikipedia to significantly improve entity ranking effectiveness.

5. ENTITY RANKING ON THE WEB

In this section we examine our third group of research questions. Can we improve web entity retrieval by using Wikipedia as a pivot? Can we automatically enrich Wikipedia with additional links to homepages of found entities? We compare our entity ranking approach of using Wikipedia as a pivot to the baseline of full-text retrieval.

5.1 Experimental Setup

This experimental section consists of two parts: in the first part we discuss experiments with the TREC Entity Ranking topics, in the second part we discuss experiments with the INEX topics that we extended to the web.

Again, we use the Indri search engine [25]. We have created separate indexes for the Wikipedia part and the Web part of the Clueweb Category B. Besides a full text index we have also created an anchor text index. On all indexes we applied the Krovetz stemmer, and we generated a length prior. All runs are created with a language model using Jelinek-Mercer smoothing with a collection λ of 0.15.

Our baseline run uses standard document retrieval on a full text index. The result format of the TREC entity ranking runs differs from the general TREC style runs. One result consists of one Wikipedia page, and can contain up to three webpages from the non-Wikipedia part of the collection. The pages in one result are supposed to be pages representing the same entity.

For our baseline runs we do not know which pages are representing the same entity. In these runs we put one homepage and one Wikipedia page in each result according to their ranks, they do not necessarily represent the same entity. The Wikipedia based runs contain up to three homepages, all on the same entity. When a result contains more than one primary page, it is counted as only one primary page, or rather entity found.

We have three approaches for finding webpages associated with Wikipedia pages.

1. External links: Follow the links in the External links section of the Wikipedia page.

2. Anchor text: Take the Wikipedia page title as query, and retrieve pages from the anchor text index. A length prior is used here.

3. Combined: Since not all Wikipedia pages have external links, and not all external links of Wikipedia pages are part of the Clueweb category B collection, we can not retrieve webpages for all Wikipedia pages. In case, less than 3 webpages are found, we fill up the results to 3 pages using the top pages retrieved using anchor text.

Our second part of experiments describes our runs with the INEX topics that we extended to the web. Instead of using the TREC entity ranking style evaluation, with results consisting of multiple pages in one result, we use a simpler evaluation with one page per result. Therefore we can use the standard evaluation scripts to calculate MAP and P10.

5.2 Experimental Results

Recall from the above that the ultimate goal of web entity ranking is to find the home-pages of the entities (called primary homepages). There are 167 primary home-pages in total (an average of 8.35 per topic) with 14 out of the 20 topics having less than 10 primary homepages. In addition, the goal is to find an entity's Wikipedia page (called a primary Wikipedia page). There are in total 172 primary Wikipedia pages (an average of 8.6 per topic) with 13 out of the 20 topics having less than 10 primary Wikipedia entities.

The results for the TREC Entity Ranking track are given in Table 6. Our baseline is full text retrieval, which works well (NDCG 0.2394) for finding relevant pages. It does however not work well for finding primary Wikipedia pages (NDCG 0.1184). More importantly, it fails miserably for finding the primary homepages: only 6 out of 167 are found, resulting in a NDCG of 0.0080 and a P10 of 0.0050. Full text retrieval is excellent at finding relevant information, but it is a poor strategy for finding web entities.

We now look at the effectiveness of our Wikipedia-as-a-pivot runs. The Wikipedia runs in this table use the external links to find homepages. The second column is based on the baseline Wikipedia run, the third column is based on the run that uses the manual categories that proved effective for entity ranking on Wikipedia in Section 4. Let us first look at the primary Wikipedia pages. We see that we find more primary Wikipedia pages, translating into a significant improvement of retrieval effectiveness (up to a P10 of 0.1700, and a NDCG of 0.1604). Will this also translate into finding more primary home pages? The first run is a straighforward run on the Wikipedia part of ClueWeb, using the external links to the Web (if present). Recall that, in Section 3, we already established that primary pages linked from relevant Wikipedia pages have a high precision. This strategy finds 29 primary homepages (so 11 more than the baseline) and improves retrieval effectiveness to an

Table 6: TREC Web Entity Ranking Results

Run	Full Text	Wikipedia Link	Wikipedia Cat+Link
Rel. WP	**73**	**73**$^{-}$	57°
Rel. HP	**244**	69$^{\bullet}$	70$^{\bullet}$
Rel. All	**316**	134$^{\bullet}$	121$^{\bullet}$
NDCG Rel. WP	**0.2119**	**0.2119**$^{-}$	0.1959^{-}
NDCG Rel. HP	**0.1919**	0.0820$^{\bullet}$	0.0830$^{\bullet}$
NDCG Rel. All	**0.2394**	0.1429$^{\bullet}$	0.1542$^{\bullet}$
Primary WP	78	78^{-}	**96**$^{\bullet}$
Primary HP	6	29°	**34**$^{\circ}$
Primary All	86	107°	**130**$^{\bullet}$
P10 pr. WP	0.1200	0.1200^{-}	**0.1700**$^{\circ}$
P10 pr. HP	0.0050	0.0300°	**0.0400**$^{\bullet}$
P10 pr. All	0.1200	0.1300^{-}	**0.1850**$^{\bullet}$
NDCG pr. WP	0.1184	0.1184^{-}	**0.1604**$^{\bullet}$
NDCG pr. HP	0.0080	0.0292^{-}	**0.0445**$^{\circ}$
NDCG pr. All	0.1041	0.1292^{-}	**0.1610**$^{\bullet}$

Table 7: TREC Homepage Finding Results

Run	Cat+Link	Anchor	Comb.
Rel. HP	70	127	**137**
Rel. All	121	178	**188**
NDCG Rel. HP	0.0830	0.0890	**0.1142**
NDCG Rel. All	0.1542	0.1469	**0.1605**
Primary HP	34	29	**56**
Primary All	130	125	**152**
P10 pr. HP	0.0400	0.0450	**0.0550**
P10 pr. All	**0.1850**	0.1750	**0.1850**
NDCG pr. HP	0.0445	0.0293	**0.0477**
NDCG pr. All	0.1041	0.1472	**0.1610**

NDCG of 0.0292, and a P10 of 0.0300.[6] The second run using the Wikipedia category information improves significantly to 34 primary homepages and a NDCG of 0.0445 and a P10 of 0.0400.

Recall again from Section 3 that the external links have high precision but low recall. We try to find additional links between retrieved Wikipedia pages and the homepages by querying the anchor text index with the name of the found Wikipedia entity (i.e., the title of the Wikipedia page). This has no effect on the found Wikipedia entities, so we only discuss the primary homepages as presented in Table 7. Ignoring the existing external links, searching for the Wikipedia entitities in the anchor text leads to 29 primary homepages. The combined run supplementing the existing external links in Wikipedia with the automatically generated links, finds a total of 56 primary homepages. For homepages this improves the P10 over the baseline to 0.0550, and NDCG to 0.0447.

Our second part of the web experiments uses the INEX topics mapped to the Clueweb collection with our additional judgments for the Clueweb web pages not in Wikipedia. Results can be found in Table 8. Although the assessments for the Wikipedia pages are fairly complete, since they are mapped from the official INEX assessments, for the web entities we are restricted to web pages occurring in the Clueweb collection. The INEX topics were not selected to lead to entities with homepages in the particular ClueWeb collection, so many relevant entities in Wikipedia have no known homepage in ClueWeb. On the negative side, this will make our scores on Wikipedia entities higher than on Web homepages. On the positive side, the 15% of Wikipedia entities with known homepages in ClueWeb substantially extend the TREC data.

[6]Unfortunately, we suffer from relatively few primary pages per topic—less than 10 for the majority of topics—and many unjudged pages for these runs. The baseline anchor text run has 100% of primary HPs and 66% of primary WPs judged in the top 10, but the Wikipedia Links run has only 45% and 53%, respectively, judged. For some of the runs discussed below this goes down to 22% of the top 10 results judged. With these fractions of judged pages, all scores of runs not contributing to the pool are underestimates of their performance.

Table 8: INEX Web Entity Ranking Results

Run	Full Text	Wikipedia Link	Wikipedia Cat+Link
Primary WP	763	763^{-}	**780**$^{-}$
Primary HP	4	73$^{\bullet}$	**86**$^{\bullet}$
Primary all	372	686$^{\bullet}$	**775**$^{\bullet}$
P10 pr. WP	0.2018	0.2018^{-}	**0.2673**$^{\bullet}$
P10 pr. HP	0.0000	0.0385°	**0.0538**$^{\bullet}$
P10 pr. All	0.0418	0.1418$^{\bullet}$	**0.2109**$^{\bullet}$
MAP pr. WP	0.1229	0.1229^{-}	**0.1633**$^{\bullet}$
MAP pr. HP	0.0001	0.0628$^{\bullet}$	**0.0754**$^{\bullet}$
MAP pr. All	0.0267	0.0910$^{\bullet}$	**0.1318**$^{\bullet}$

Our full-text baseline run achieves poor results. While a full-text run works fine on the restricted Wikipedia domain, on the Web it does not succeed in finding primary homepages, also relative to the known homepages in ClueWeb. Again we find that exploiting the Wikipedia category information consistently improves the results for finding primary Wikipedia pages as well as primary homepages. Since there are more primary Wikipedia pages than homepages, the Wikipedia scores are the highest overall. In contrast to the TREC entity ranking runs previously discussed in this section, each result consists of only one page. Since we are better at finding primary Wikipedia pages, we could construct better overall runs, by simply ranking the Wikipedia pages higher than the web pages. Depending on your goal, you could choose to show a ranking that is less diverse and shows only or primarily Wikipedia results, but contains more relevant documents.

Summarising the section, we examined whether web entity retrieval can be improved by using Wikipedia as a pivot. We found that full text retrieval fails miserably at finding primary homepages of entities. Full text retrieval on Wikipedia, in contrast, works reasonable, and using Wikipedia as a pivot by mapping found Wikipedia entities to the Web using the external links leads to many more primary homepages of entities being found. We also investigated whether we could supplement the external links with homepages found by searching an anchor text index for the retrieved Wikipedia entities, and found that this leads to a significant improvement over just using Wikipedia's external links for finding primary homepages of entities.

6. CONCLUSIONS

This paper investigates the problem of entity retrieval. A natural baseline for entity retrieval is standard full text retrieval. While this baseline does find a considerable number of relevant pages, it is not able to locate the primary homepages, which is the main goal of our entity ranking task. The text retrieval runs fare much better at finding Wikipedia pages of relevant entities, hence prompting the use of Wikipedia as a pivot to find the primary web homepages of entities. Our experiments show that our wikipedia-as-a-pivot approach outperforms a baselines of full-text search. Both external links on Wikipedia pages, and searching an anchor text index of the web are

effective approaches to find homepages for entities represented by Wikipedia pages.

The approach is based on three assumptions: i) the coverage of entities in Wikipedia is large enough; ii) we are able to find entities in Wikipedia, iii) we can map Wikipedia entities to the appropriate web home pages. We have shown that the coverage of topics in Wikipedia is large, and Wikipedia is constantly growing. The external links on Wikipedia pages are almost always authoritative or official homepages of the entity. We also demonstrated that a large fraction external links in Wikipedia are relevant web homepages. Besides the external links, querying an anchor text index for entity names is also effective. The combination of these two approaches leads to additional improvements. Considering entity types, automatically assigned target entity types are almost as effective as manually assigned entity types. Entity type information improves retrieval scores considerably, up to 50% improvement rates.

Our future work will examine how alternative knowledge sources could complement Wikipedia's role as a pivot for those information needs involving entities not well represented there. Analysis of search log queries is needed to study more extensively the coverage of Wikipedia concerning different types of entities. If we can find relevant entities in other knowledge sources, we can use these as pivots as well, and identify relevant homepages by using an anchor text index. Search log queries and clicks, which are currently unavailable, can be used in a similar way as we use the anchor text—leading to further improvements. Our broad conclusion is that is it viable to exploiting the available structured information in Wikipedia and other resources, to make sense of the great amount of unstructured information on the Web.

Acknowledgments The created Entity Ranking topics test collection is available at http://staff.science.uva.nl/~kamps/effort/data. This research was supported by the Netherlands Organization for Scientific Research (NWO, under project # 612.066.513).

REFERENCES

[1] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 797–806. ACM Press, New York USA, 2009.

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

[3] K. Balog, M. Bron, and M. de Rijke. Category-based query modeling for entity search. In *32nd European Conference on Information Retrieval (ECIR 2010)*, pages 319–331. Springer, 2010.

[4] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *Proceedings of the IJCAI '07*, pages pages 2657–2662, 2007.

[5] K. Balog, A. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Notebook*. National Institute for Standards and Technology, 2009.

[6] H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: efficient search on text, entities, and relations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 671 – 678, 2007.

[7] J. G. Conrad and M. H. Utt. A system for discovering relationships by feature extraction from text databases. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 260–270, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[8] A. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In *INEX 2007*, pages 245–251, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] G. Demartini, C. S. Firan, T. Iofciu, R. Krestel, and W. Nejdl. Why finding entities in wikipedia is difficult, sometimes. *Information Retrieval", Special Issue on Focused Retrieval and Result Aggregation*, 2010.

[10] G. Demartini, T. Iofciu, and A. de Vries. Overview of the inex 2009 entity ranking track. In *INEX 2009 Workshop Pre-Proceedings*, 2009.

[11] G. Demartini, A. P. Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 entity ranking track. In *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008*, pages 243–252, Berlin, Heidelberg, 2009. Springer-Verlag.

[12] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[13] Y. Fang, L. Si, Z. Yu, Y. Xian, and Y. Xu. Entity retrieval with hierarchical relevance model. In *The Eighteenth Text REtrieval Conference (TREC 2009) Notebook*, 2009.

[14] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages pp. 363–370, 2005.

[15] T. Götz and O. Suhre. Design and implementation of the UIMA common analysis system. *IBM Syst. J.*, 43(3):476–489, 2004.

[16] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001.

[17] R. Kaptein, M. Koolen, and J. Kamps. Using Wikipedia categories for ad hoc search. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York NY, USA, 2009.

[18] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and Ranking Knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*. IEEE, 2008.

[19] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. L. T. Santos. University of glasgow at TREC 2009: experiments with terrier. In *The Eighteenth Text REtrieval Conference (TREC 2009) Notebook*, 2009.

[20] E. Meij, P. Mika, and H. Zaragoza. An evaluation of entity and frequency based query completion methods. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 678–679, New York, NY, USA, 2009. ACM.

[21] M. Paşca. Weakly-supervised discovery of named entities using web search queries. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 683–690, New York, NY, USA, 2007. ACM.

[22] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740, New York, NY, USA, 2007. ACM.

[23] H. Raghavan, J. Allan, and A. Mccallum. An exploration of entity models, collective classification and relation description. In *KDD'04*, 2004.

[24] R. Schenkel, F. M. Suchanek, and G. Kasneci. Yawn: A semantically annotated wikipedia xml corpus. In *BTW*, pages 277–291, 2007.

[25] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[26] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In *Focused Access to XML Documents*, pages 306–320, 2007.

[27] D. Vallet and H. Zaragoza. Inferring the most important types of a query: a semantic approach. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 857–858, New York, NY, USA, 2008. ACM.

[28] A.-M. Vercoustre, J. Pehcevski, and J. A. Thom. Using wikipedia categories and links in entity ranking. In *Focused Access to XML Documents*, pages 321–335, 2007.

[29] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1101–1106, New York, NY, USA, 2008. ACM.

[30] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1015–1018, New York, NY, USA, 2007. ACM.