# Using Wikipedia Categories and Links in Entity Ranking

Anne-Marie Vercoustre[1], Jovan Pehcevski[1], and James A. Thom[2]

[1] INRIA Rocquencourt, France
{anne-marie.vercoustre,jovan.pehcevski}@inria.fr
[2] RMIT University, Melbourne, Australia
james.thom@rmit.edu.au

**Abstract.** This paper describes the participation of the INRIA group in the INEX 2007 XML entity ranking and ad hoc tracks. We developed a system for ranking Wikipedia entities in answer to a query. Our approach utilises the known categories, the link structure of Wikipedia, as well as the link co-occurrences with the examples (when provided) to improve the effectiveness of entity ranking. Our experiments on both the training and the testing data sets demonstrate that the use of categories and the link structure of Wikipedia can significantly improve entity retrieval effectiveness. We also use our system for the ad hoc tasks by inferring target categories from the title of the query. The results were worse than when using a full-text search engine, which confirms our hypothesis that ad hoc retrieval and entity retrieval are two different tasks.

## 1 Introduction

*Entity ranking* has recently emerged as a research field that aims at retrieving entities as answers to a query [7, 10, 12, 14]. Here, unlike in the related field of entity extraction, the goal is not to tag the names of the entities in documents but rather to get back a list of the relevant entity names. It is a generalisation of the expert search task explored by the TREC Enterprise track [11], except that instead of ranking people who are experts in the given topic, other types of entities such as organisations, countries, or locations can also be ranked.

The Initiative for the Evaluation of XML retrieval (INEX) ran a new track on entity ranking in 2007, using Wikipedia as its document collection [5]. There were two tasks in the INEX 2007 XML entity ranking (XER) track: task 1 (*entity ranking*), with the aim of retrieving entities of a given category that satisfy a topic described in natural language text; and task 2 (*list completion*), where given a topic text and a small number of entity examples, the aim was to complete this partial list of answers. Two data sets were used by the participants of the XER track: a *training* data set, comprising 28 XER topics which were adapted from the INEX 2006 ad hoc topics; and a *testing* data set, comprising 46 XER topics that were proposed and assessed by the track participants.

In the XER track, the expected entities correspond to Wikipedia articles that are likely to be referred to by links in other articles. As an example, the query

"European countries where I can pay with Euros" [5] should only return a list of entities (or pages) representing relevant countries, and not include entities representing non-relevant countries nor other entities found in pages about the Euro and similar currencies.

In this paper, we describe our approach to ranking entities from the Wikipedia XML document collection. Our approach is based on the following hypotheses:

1. A good entity page is a page that answers the query, or a query extended with names of target categories (task 1) or entity examples (task 2).
2. A good entity page is a page associated with a category close to the target category (task 1) or to the categories of the entity examples (task 2).
3. A good entity page is referred to by a page answering the query; this is an adaptation of the HITS [9] algorithm to the problem of entity ranking.
4. A good entity page is referred to by contexts with many occurrences of the entity examples (task 2). A broad context could be the full page that contains the entity examples, while smaller and more narrow contexts could be elements such as paragraphs, lists, or tables.

This paper is organised as follows. After a short review of the related work and a brief presentation of the INEX Wikipedia XML collection used for entity ranking, we provide a detailed description of our entity ranking approach and the runs we submitted for evaluation to the INEX 2007 XER track. We also report on our run submissions to the INEX 2007 ad hoc track that are based on our entity ranking approach. For both tracks we submitted a run based on a full-text retrieval approach. By analysing and comparing the performances of runs based on these two approaches, we address the following research question: Are ad hoc retrieval and entity retrieval two different tasks?

## 2 Related work

Entity ranking has attracted a lot of research recently. It can be seen as a generalisation of expert search where the entities of interest are not only people. For example, Craswell et al. [4] use the co-occurrence of people's names and query words in documents as evidence to rank experts. Zhu et al. [15] have extended their expert search system to allow for entity search. Their approach involves an association model based on co-occurrence of entities with query terms in documents mentioning the entity. The association can be made at multiple levels: phrase, sentence, paragraph and up to a document level, with associated weights that decrease for larger contexts. Entities are filtered by comparing their categories with the target category and its child and parent categories.

ESTER [2] was recently proposed as a system for searching text, entities and relations. ESTER relies on the Wikipedia links to identify the entities and on the context of the links for disambiguation (using 20 words around the anchor text instead of just the anchor text). Hu et al. [8] propose a linear model that uses a number of features to weight passages containing entity names. They first determine top $k$ passages and extract the top $n$ entities from these passages.

"The **euro** ... is the official currency of the Eurozone (also known as the Euro Area), which consists of the European states of Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Slovenia and Spain, and will extend to include Cyprus and Malta from 1 January 2008."

**Fig. 1.** Extract from the Euro Wikipedia page

Features include term frequency, distance to the entity name and co-occurrences in the same section as the entity. Tsikrika et al. [13] build a graph of the initial set of documents returned in answer to the query, but do not extend the graph to linked documents outside the initial set; this graph is then used to propagate relevance in entity retrieval based on $k$-step or infinite random walk. Entities are filtered by using the target category and its child categories (up to a third level).

## 3 INEX Wikipedia XML collection

Wikipedia is a well known web-based, multilingual, free content encyclopedia written collaboratively by contributors from around the world. Denoyer and Gallinari [6] have developed an XML-based corpus based on a snapshot of the Wikipedia, which has been used by various INEX tracks in 2006 and 2007. It differs from the real Wikipedia in some respects (size, document format, category tables), but it is a very realistic approximation.

### 3.1 Entities in Wikipedia

The entities have a name (the name of the corresponding page) and a unique ID in the collection. When mentioning such an entity in a new Wikipedia article, authors are encouraged to link occurrences of the entity name to the page describing this entity. This is an important feature as it makes it easy to locate potential entities, which is a major issue in entity extraction from plain text.

However in this collection, not all potential entities have been associated with corresponding pages. The INEX 2007 XER topics have been carefully designed to make sure there is a sufficient number of answer entities. For example, in the Euro page (see Fig. 1), all the underlined hypertext links can be seen as occurrences of entities that are each linked to their corresponding pages. In this figure, there are 18 entity references of which 15 are country names; specifically, these countries are all "European Union member states", which brings us to the notion of categories in Wikipedia.

### 3.2 Categories in Wikipedia

Wikipedia also offers categories that authors can associate with Wikipedia pages. There are 113,483 categories in the INEX Wikipedia XML collection, which are organised in a graph of categories. Each page can be associated with many categories (2.28 as an average).

Wikipedia categories have unique names (e.g. "France", "European Countries", "Countries"). New categories can also be created by authors, although they have to follow Wikipedia recommendations in both creating new categories and associating them with pages. For example, the Spain page is associated with the following categories: "Spain", "European Union member states", "Spanish-speaking countries", "Constitutional monarchies" (and some other Wikipedia administrative categories).

When searching for entities it is natural to take advantage of the Wikipedia categories since they would give a hint on whether the retrieved entities are of the expected type. For example, when looking for entities of type "authors", pages associated with the category "Novelist" are more likely to be relevant than pages associated with the category "Book".

## 4 Our entity ranking approach

Our approach to identifying and ranking entities combines: (i) the full-text similarity of the answer entity page with the query; (ii) the similarity of the page's categories to the target categories (task 1) or to the categories attached to the entity examples (task 2); and (iii) the contexts around entity examples found in the top ranked pages returned by a search engine for the query.

We have built a system based on the above ideas, and a framework to tune and evaluate a set of different entity ranking algorithms.

### 4.1 Architecture

The system involves several modules and functions that are used for processing a query, submitting it to the search engine, applying our entity ranking algorithms, and finally returning a ranked list of entities. We use Zettair[3] as our choice for a full-text search engine. Zettair is a full-text information retrieval (IR) system developed by RMIT University, which returns pages ranked by their similarity score to the query. We used the Okapi BM25 similarity measure that has proved to work well on the INEX 2006 Wikipedia test collection [1].

Our system involves the following modules and functions:

- the topic module takes an INEX topic as input and generates the corresponding Zettair query and the list of target categories and entity examples; as an option, the names of target categories (task 1) or example entities (task 2) may be added to the query;
- the search module sends the query to Zettair and returns a list of ranked Wikipedia pages (typically 1500);
- the link extraction module extracts the links from a selected number of highly ranked pages,[4] together with the information concerning the paths of the links (using an XPath notation);

[3] http://www.seg.rmit.edu.au/zettair/

[4] We discarded external links and some internal collection links that do not refer to existing pages in the INEX Wikipedia collection.

– the category similarity module calculates a weight for a page based on the similarity of the page categories to target categories or to those attached to entity examples (see sub-section 4.2);
– the linkrank module calculates a weight for a page based (among other things) on the number of links to this page (see sub-section 4.3); and
– the full-text IR module calculates a weight for a page based on its initial Zettair score.

The global score for a page is calculated as a linear combination of three normalised scores coming out of the last three modules (see sub-section 4.4).

The architecture provides a general framework for evaluating entity ranking which allows for some modules to be replaced by more advanced modules, or by providing a more efficient implementation of a module. It also uses an evaluation module to assist in tuning the system by varying the parameters and to globally evaluate our entity ranking approach.

The major cost in running our system lies in extracting the links from the selected number of pages retrieved by the search engine. Although we only extract links once by topic and store them in a database for reuse in later runs, an online system would require extracting and storing all the links at indexing time.

### 4.2 Using Wikipedia categories

To make use of the Wikipedia categories in entity ranking, we define similarity functions between the categories of answer entities and the target categories (task 1), or between the categories of answer entities and a set of categories attached to the entity examples (task 2).

Similarity measures between concepts of the same ontology, such as tree-based similarities [3], cannot be applied directly to Wikipedia categories, mostly because the notion of sub-categories in Wikipedia is not a pure subsumption relationship. Another reason is that categories in Wikipedia do not form a hierarchy (or a set of hierarchies) but form a graph with potential cycles.

**Task 1** We first define a similarity function that computes the ratio of common categories between the set of categories, $\mathsf{cat}(t)$, associated to an answer entity page $t$, and the set $\mathsf{cat}(C) = C$, where $C$ is the set of provided target categories:

$$S_C(t) = \frac{|\mathsf{cat}(t) \cap \mathsf{cat}(C)|}{|\mathsf{cat}(C)|} \tag{1}$$

The target categories will be generally very broad, so it is to be expected that the answer entities would not be directly attached to these broad categories. Accordingly, we experimented with several extensions of the set of categories, both for the target categories and the categories attached to answer entities.

We first experimented with extensions based on using sub-categories and parent categories in the graph of Wikipedia categories. However, on the training

data set, we found that these category extensions overall do not result in an improved performance [12], and so they were not used in our INEX 2007 runs.

Another approach is to use lexical similarity between the category names. For example, "european countries" is lexically similar to "countries" since they both contain the word "countries" in their names. We use an information retrieval approach to retrieve similar categories, by constructing a separate index with Zettair of all the category names (using the names as documents). By sending both the title of the topic T and the category names C as a query to Zettair, we then retrieve all the categories that are lexically similar to C. We keep the top M ranked categories and add them to C to form the set $\mathsf{TCcat}(C)$. On the training data set, we found that the value M=5 is the optimal parameter value to retrieve the likely relevant categories for this task [12]. We then use the same similarity function as before, but where $\mathsf{cat}(C) = \mathsf{TCcat}(C)$.

We also experimented with two alternative approaches: by sending the category names C as a query to Zettair (denoted as $\mathsf{Ccat}(C)$); and by sending the title of the topic T as a query to Zettair (denoted as $\mathsf{Tcat}(C)$). On the training data set, we found that these two approaches were less effective than the $\mathsf{TCcat}(C)$ approach [12]. However, we use $\mathsf{cat}(C) = \mathsf{Tcat}(C)$ in our ad-hoc runs since no target categories are provided.

**Task 2** Here, the categories attached to entity examples are likely to correspond to very specific categories, just like those attached to the answer entities. We define a similarity function that computes the ratio of common categories between the set of categories attached to an answer entity page $\mathsf{cat}(t)$ and the set of the union of the categories attached to entity examples $\mathsf{cat}(E)$:

$$S_C(t) = \frac{|\mathsf{cat}(t) \cap \mathsf{cat}(E)|}{|\mathsf{cat}(E)|} \tag{2}$$

### 4.3 Exploiting locality of links

For task 2, exploiting locality of links around entity examples can significantly improve the effectiveness of entity ranking [10]. The idea is that entity references (links) that are located in close proximity to the entity examples, especially in list-like elements, are likely to refer to more relevant entities than those referred to by links in other parts of the page.

Consider the example of the Euro page shown in Fig. 1, where France, Germany and Spain are the three entity examples. We see that the 15 countries that are members of the Eurozone are all listed in the same paragraph with the three entity examples. In fact, there are other contexts in this page where those 15 countries also co-occur together. By contrast, although there are a few references to the United Kingdom in the Euro page, it does not occur in the same context as the three examples (except for the page itself).

We have identified in the Wikipedia collections three types of elements that correspond to the notion of lists: paragraphs (tag `p`); lists (tags `normallist`,

**Table 1.** List of links referring to entity examples (France, Germany, and Spain), extracted from the Wikipedia page 9272.xml.

| Page | | Links | | |
|---|---|---|---|---|
| ID | Name | XPath | ID | Name |
| 9472 | Euro | /article[1]/body[1]/p[1]/collectionlink[7] | 10581 | France |
| 9472 | Euro | /article[1]/body[1]/p[1]/collectionlink[8] | 11867 | Germany |
| 9472 | Euro | /article[1]/body[1]/p[1]/collectionlink[15] | 26667 | Spain |
| 9472 | Euro | /article[1]/body[1]/p[3]/p[5]/collectionlink[6] | 11867 | Germany |
| 9472 | Euro | /article[1]/body[1]/normallist[1]/item[4]/collectionlink[1] | 10581 | France |
| 9472 | Euro | /article[1]/body[1]/normallist[1]/item[5]/collectionlink[2] | 11867 | Germany |
| 9472 | Euro | /article[1]/body[1]/normallist[1]/item[7]/collectionlink[1] | 26667 | Spain |
| 9472 | Euro | /article[1]/body[1]/normallist[1]/item[8]/collectionlink[1] | 26667 | Spain |

`numberlist`, and `definitionlist`); and tables (tag `table`). We use an algorithm for identifying the (static) element contexts on the basis of the leftmost occurrence of any of the pre-defined tags in the absolute XPaths of the entity examples. The resulting list of element contexts is sorted in a descending order according to the number of distinct entity examples contained by the element. If two elements contain the same number of distinct entity examples, the one that has a longer XPath length is ranked higher. Finally, starting from the highest ranked element, we filter all the elements in the list that either contain or are contained by that element. We end up with a final list of (one or more) non-overlapping elements that represent the statically defined contexts for the page.[5]

Consider Table 1, where the links to entity examples are identified by their absolute XPath notations. The three static contexts that will be identified by the above algorithm are the elements `p[1]`, `normallist[1]` and `p[3]`. The first two element contexts contain three (distinct) examples, while the last one contains only one entity example.

The drawback of this approach is that it requires a predefined list of static elements that is dependent on the collection. The advantage is that the contexts are fast to identify. We have also experimented with an alternative algorithm that dynamically identifies the link contexts. On the training data set, we found that this algorithm does not significantly improve the entity ranking performance over the algorithm that uses the static contexts [10].

### 4.4 Score functions and parameters

The core of our entity ranking approach is based on combining different scoring functions for an answer entity page, which we now describe in more detail.

[5] In the case when there are no occurrences of the pre-defined tags in the XPath of an entity example, the document element (`article[1]`) is chosen to represent the element context.

**LinkRank score** The linkrank function calculates a score for a page, based on the number of links to this page, from the first N pages returned by the search engine in response to the query. The number N has been kept to a relatively small value mainly for performance purposes, since Wikipedia pages contain many links that would need to be extracted. We carried out some experiments with different values of N and found that N=20 was a good compromise between achieving efficient performance and discovering more potentially good entities [14].

The linkrank function can be implemented in a variety of ways. We have implemented a linkrank function that, for an answer entity page $t$, takes into account the Zettair score of the referring page $z(p)$, the number of distinct entity examples in the referring page $\#ent(p)$, and the locality of links around the entity examples:

$$S_L(t) = \sum_{r=1}^{N} \left( z(p_r) \cdot g(\#ent(p_r)) \cdot \sum_{l_t \in L(p_r,t)} f(l_t, c_r | c_r \in C(p_r)) \right) \quad (3)$$

where $g(x) = x + 0.5$ (we use 0.5 to allow for cases where there are no entity examples in the referring page); $l_t$ is a link that belongs to the set of links $L(p_r, t)$ that point from the page $p_r$ to the answer entity $t$; $c_r$ belongs to the set of contexts $C(p_r)$ around entity examples found for the page $p_r$; and $f(l_t, c_r)$ represents the weight associated to the link $l_t$ that belongs to the context $c_r$.

The weighting function $f(l_t, c_r)$ is represented as follows:

$$f(l_t, c_r) = \begin{cases} 1 & \text{if } c_r = p_r \text{ (the context is the full page)} \\ \\ 1 + \#ent(c_r) & \text{if } c_r = e_r \text{ (the context is an XML element)} \end{cases}$$

A simple way of defining the context of a link is to use its full embedding page [14]. In this work we use smaller contexts using predefined types of elements such as paragraphs, lists and tables (as described in sub-section 4.3).

**Category similarity score** The category score $S_C(t)$ is calculated using equation (1) for task 1 and equation (2) for task 2 (as described in sub-section 4.2).

For task 1, we consider variations on the category score $S_C(t)$ based on lexical similarities of category names (see sub-section 4.2), by replacing $\mathsf{cat}(C)$ with $\mathsf{TCcat}(C)$.

For task 2 we do not use any category extensions since, on the training data set, we found that extending the set of categories attached to both entity examples and answer entities did not increase the entity ranking performance [12].

**Z score** The Z score assigns the initial Zettair score to an answer entity page. If the answer page does not appear among the initial ranked list of pages returned by Zettair, then its Z score is zero:

$$S_Z(t) = \begin{cases} z(t) & \text{if page } t \text{ was returned by Zettair} \\ \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The Z score is not the same as the plain Zettair score, since our system extracts new entities (pages) from the links contained in the highest N pages returned by Zettair; these new pages may or may not be included in the initial 1500 pages retrieved by Zettair.

**Global score** The global score $S(t)$ for an answer entity page is calculated as a linear combination of three normalised scores: the normalised linkrank score $nS_L(t)$, the category similarity score $nS_C(t)$ and the Z score $nS_Z(t)$:

$$S(t) = \alpha \cdot nS_L(t) + \beta \cdot nS_C(t) + (1 - \alpha - \beta) \cdot nS_Z(t) \tag{5}$$

where $\alpha$ and $\beta$ are two parameters that can be tuned differently depending on the entity retrieval task.

We consider some special cases that allow us to evaluate the effectiveness of each module in our system: $\alpha = 1, \beta = 0$, which uses only the linkrank score; $\alpha = 0, \beta = 1$, which uses only the category score; and $\alpha = 0, \beta = 0$, which uses only the Z score. More combinations for the two parameters are explored in the training phase of our system. The optimal combination is then used on the testing data set.

## 5 Experimental results

In this section, we present results that investigate the effectiveness of our entity ranking approach when applied to both the INEX 2007 XER and ad hoc tracks.

For the XER track, we submitted three runs for task 1 (entity ranking) and three runs for task 2 (list completion). For this track, we aim at investigating the impact of using various category and linkrank similarity techniques on the entity ranking performance; we also compare the run performances with a full-text retrieval run as a baseline. For the ad hoc track, we submitted three entity ranking runs that correspond to the three individual modules of our system and compare their performances to the performance of the full-text Zettair run submitted by RMIT.

### 5.1 Runs description

Table 2 lists the six XER and four ad hoc runs that we submitted to INEX 2007. With the exception of the plain Zettair run, all the runs were created by using

**Table 2.** List of six XER and four ad hoc runs submitted for evaluation. "Cat-sim" stands for category similarity, "Ctx" for context, "Cat" for categories, "Ent" for entities, "T" for title, "TC" for title and categories, "C" for category names, "CE" for category and entity names, "FC" for full page context, and "EC" for element context.

| | | | | **Category index** | | | | **Topic** | |
|---|---|---|---|---|---|---|---|---|---|
| **Run ID** | cat-sim | $\alpha$ | $\beta$ | Query | Type | M | Ctx | Cat | Ent |
| Zettair | | – | – | – | – | – | – | – | – |
| **XER task 1** | | | | | | | | | |
| run 1 | $\mathsf{cat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | – | – | – | FC | Yes | No |
| run 2 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | TC | C | 5 | FC | Yes | No |
| run 3 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.1 | 0.8 | TC | C | 5 | FC | Yes | No |
| **XER task 2** | | | | | | | | | |
| run 1 | $\mathsf{cat}(E)$-$\mathsf{cat}(t)$ | 1.0 | 0.0 | – | – | – | EC | No | Yes |
| run 2 | $\mathsf{cat}(E)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | – | – | – | EC | No | Yes |
| run 3 | $\mathsf{cat}(E)$-$\mathsf{cat}(t)$ | 0.2 | 0.6 | – | – | – | EC | No | Yes |
| **Ad hoc retrieval task** | | | | | | | | | |
| run 1 | $\mathsf{Tcat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 0.0 | T | CE | 10 | FC | No | No |
| run 2 | $\mathsf{Tcat}(C)$-$\mathsf{cat}(t)$ | 1.0 | 0.0 | T | CE | 10 | FC | No | No |
| run 3 | $\mathsf{Tcat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | T | CE | 10 | FC | No | No |

our entity ranking system. However, as seen in the table the runs use various parameters whose values are mainly dependent on the task. Specifically, runs differ depending on whether (or which) Zettair category index is used, which of the two types of link contexts is used, whether categories or example entities are used from the topic, and which combination of values is assigned to the $\alpha$ and $\beta$ parameters.

For example, the run "run 3" for XER task 1 can be interpreted as follows: the Zettair index of category names is used to extract the top five ranked categories, using both the title and the category names (TC) from the INEX topic as a query. This set of five categories is used as an input in the category similarity function ($\mathsf{TCcat}(C)$). The full page context (FC) is used to calculate the scores in the linkrank module. The final scores for answer entities are calculated by combining the scores coming out of the three modules ($\alpha = 0.1$, $\beta = 0.8$).

### 5.2 XER track

Two data sets were used by the participants of the XER track: a training data set and a testing data set. The training data set is based on a selection of topics from the INEX 2006 ad hoc track, resulting in total of 28 topics with corresponding relevance assessments. The testing data set consists of two subsets: a subset of topics based on a selection of topics from the INEX 2007 ad hoc track, and a subset of topics specifically developed by participants for the purposes of the XER track. The complete testing data set results in total of 46 topics with corresponding relevance assessments.

**Table 3.** Performance scores for Zettair and our three XER submitted runs on the training data set (28 topics) and testing data set (46 topics), obtained with different evaluation measures for INEX 2007 XER task 1: entity ranking. For each data set, the best performing score under each measure is shown in bold.

| | | | | P[r] | | | |
|---|---|---|---|---|---|---|---|
| Run ID | cat-sim | $\alpha$ | $\beta$ | 5 | 10 | R-prec | MAP |
| **Training data set** | | | | | | | |
| Zettair | | – | – | 0.229 | 0.232 | 0.208 | 0.172 |
| run 1 | $\mathsf{cat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | 0.229 | 0.250 | 0.215 | 0.196 |
| run 2 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | 0.307 | 0.318 | 0.263 | 0.242 |
| run 3 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.1 | 0.8 | **0.379** | **0.361** | **0.338** | **0.287** |
| **Testing data set** | | | | | | | |
| Zettair | | – | – | 0.230 | 0.211 | 0.208 | 0.186 |
| run 1 | $\mathsf{cat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | 0.283 | 0.243 | 0.235 | 0.199 |
| run 2 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.0 | 1.0 | 0.322 | 0.296 | 0.300 | 0.243 |
| run 3 | $\mathsf{TCcat}(C)$-$\mathsf{cat}(t)$ | 0.1 | 0.8 | **0.378** | **0.339** | **0.346** | **0.294** |

We use mean average precision (MAP) as our primary method of evaluation, but also report results using several alternative measures that are typically used to evaluate the retrieval performance: mean of P[5] and P[10] (mean precision at top 5 or 10 entities returned), and mean R-precision (R-precision for a topic is the P[R], where R is the number of entities that have been judged relevant for the topic). For task 1 all the relevant entities in the relevance assessments are used to generate the scores, while for task 2 we remove the entity examples both from the list of returned answers and from the relevance assessments, as the task is to find entities other than the provided examples.

**Task 1: Entity ranking** Table 3 shows the performance scores on both the training and the testing data sets for task 1, obtained for Zettair and our three submitted XER runs. Runs 1 and 2 use scores coming out from the category module only ($\alpha = 0.0$, $\beta = 1.0$) while run 3 uses a combination of linkrank, category, and Z scores ($\alpha = 0.1$, $\beta = 0.8$). Runs 2 and 3 use lexical similarity for extending the set of target categories.

When comparing the performances of runs that use the category module only (runs 1 and 2), we observe that run 2 that uses lexical similarity between category names ($\mathsf{TCcat}(C)$) is more effective than the run that uses the topic-provided target categories ($\mathsf{cat}(C)$). With MAP, the difference in performance between the two runs is statistically significant ($p < 0.05$). We also observe that the third run, which uses combined scores from the three modules, performs the best among the three. To find the optimal values for the two combining parameters for this run, we calculated MAP over the 28 topics in the training data set as we varied $\alpha$ from 0 to 1 in steps of 0.1. For each value of $\alpha$, we also varied $\beta$ from 0 to $(1 - \alpha)$ in steps of 0.1. We found that the highest MAP score

**Table 4.** Performance scores for Zettair and our three XER submitted runs on the training data set (28 topics) and testing data set (46 topics), obtained with different evaluation measures for INEX 2007 XER task 2: list completion. For each data set, the best performing score under each measure is shown in bold.

| | | | | P[r] | | | |
|---|---|---|---|---|---|---|---|
| Run | cat-sim | $\alpha$ | $\beta$ | 5 | 10 | R-prec | MAP |
| **Training data set** | | | | | | | |
| Zettair | – | – | – | 0.229 | 0.232 | 0.208 | 0.172 |
| run 1 | cat($E$)-cat($t$) | 1.0 | 0.0 | 0.214 | 0.225 | 0.229 | 0.190 |
| run 2 | cat($E$)-cat($t$) | 0.0 | 1.0 | 0.371 | 0.325 | 0.319 | 0.318 |
| run 3 | cat($E$)-cat($t$) | 0.2 | 0.6 | **0.500** | **0.404** | **0.397** | **0.377** |
| **Testing data set** | | | | | | | |
| Zettair | – | – | – | 0.183 | 0.170 | 0.173 | 0.155 |
| run 1 | cat($E$)-cat($t$) | 1.0 | 0.0 | 0.157 | 0.150 | 0.163 | 0.141 |
| run 2 | cat($E$)-cat($t$) | 0.0 | 1.0 | 0.370 | 0.298 | 0.292 | 0.263 |
| run 3 | cat($E$)-cat($t$) | 0.2 | 0.6 | **0.409** | **0.330** | **0.336** | **0.309** |

(0.287) is achieved for $\alpha = 0.1$ and $\beta = 0.8$ [12]. This is a statistically significant 19% relative performance improvement over the best score achieved by using only the category module ($\alpha$0.0–$\beta$1.0). The same performance behaviour among the three XER runs is also observed on the testing data set.

From Table 3 we also observe that, irrespective of the data set used, the three entity ranking runs outperform the plain Zettair run. This suggests that using full-text retrieval alone is not an effective retrieval strategy for this task. The differences in performance between each of the three runs and Zettair are statistically significant ($p < 0.05$) only for the two entity ranking runs that use lexical similarity between category names (runs 2 and 3 in Table 3).

When comparing the MAP scores obtained for runs submitted by all XER track participants, our INRIA run 3 was ranked as the third best performing run among the 20 submitted runs for INEX 2007 XER task 1.

**Task 2: List completion** Table 4 shows the performance scores on both the training and testing data sets for task 2, obtained for Zettair and our three submitted XER runs. With the first two runs, we want to compare two entity ranking approaches: the first that uses scores from the linkrank module only (run 1), and the second that uses scores from the category module only (run 2). We observe that using categories is substantially more effective than using the linkrank scores. With MAP, the difference in performance between the two runs is statistically significant ($p < 0.05$) on both data sets.

Run 3 combines the scores from the three modules. To find the optimal values for the two combining parameters for this run, we again used the training data set and varied the values for parameters $\alpha$ and $\beta$. We found that the highest MAP score (0.377) was achieved for $\alpha = 0.2$ and $\beta = 0.6$ [10]. This is

a statistically significant 19% relative performance improvement over the best score achieved by using only the category module. From Table 4 we see that the same performance behaviour among the three XER runs is also observed on the testing data set.

When the three XER runs are compared with the plain Zettair run, we observe a slightly different performance behaviour depending on the data set used. Specifically, on the training data set the three XER runs outperform the plain Zettair run, while on the testing data set only runs 2 and 3 outperform Zettair which in turn outperforms run 1 (the run that uses linkrank scores only). A more detailed per-topic analysis of this behaviour revealed that this is a result of the different "nature" of the two subsets used in the testing data set. Specifically, Zettair outperformed run 1 only on the 21 topics comprising the ad hoc testing topic subset, while run 1 outperformed Zettair on the 25 topics comprising the testing topic subset developed by the XER participants. This indicates that the ad hoc topic subset may need to be further revised and adapted if it is to be reliably used for XER-specific retrieval tasks.

When comparing the MAP scores obtained for runs submitted by all XER track participants, our INRIA run 3 was ranked as the best performing run among the 10 submitted runs for INEX 2007 XER task 2.

### 5.3 Ad hoc track

There are no target categories and example entities provided for the retrieval tasks of the INEX 2007 ad hoc track. However, we wanted to apply our algorithms to test 1) whether some indication of page categories would improve the ad hoc retrieval performance, and 2) whether extracting new entities from the pages returned by Zettair would be beneficial for ad hoc retrieval.

We submitted four runs for the INEX 2007 ad hoc track: Zettair, representing a full-text retrieval run, and three entity ranking runs. As shown in Table 2, run 1 uses only the Z module for ranking the answer entities, run 2 uses only the linkrank module, while run 3 uses only the category module. For each of the 99 topics with relevance assessments used in the INEX 2007 ad hoc track, we created the set of target categories by sending the title T of the query to the Zettair index of categories that has been created by using the names of the categories and the names of all their attached entities as corresponding documents.

Table 5 shows the performance scores on the INEX 2007 ad hoc data set, obtained for Zettair and our three submitted entity ranking runs. Two retrieval scenarios are distinguished in the table: a *document retrieval* scenario (the first four result columns in Table 5), where we compare how well the runs retrieve relevant documents; and a *focused retrieval* scenario (the last three result columns in Table 5), where we compare how well the runs retrieve relevant information within documents.

For the document retrieval scenario, we observe that Zettair outperforms the other three XER runs. The differences in performance between Zettair and any of these three runs are statistically significant ($p < 0.05$). Among the three XER runs, the run that only uses the Z scores performs significantly better than

**Table 5.** Performance scores for Zettair and our three XER submitted runs on the ad hoc data set (99 topics), obtained with different evaluation measures for the INEX 2007 ad hoc track. For each measure, the best performing score is shown in bold.

| | | | P[r] | | | | Foc | RiC | BiC |
|---|---|---|---|---|---|---|---|---|---|
| Run | $\alpha$ | $\beta$ | 5 | 10 | R-prec | MAP | iP[0.01R] | MAgP | MAgP |
| Zettair | – | – | **0.513** | **0.469** | **0.326** | **0.292** | **0.483** | **0.136** | **0.192** |
| run 1 | 0.0 | 0.0 | **0.513** | **0.469** | 0.303 | 0.247 | **0.483** | 0.115 | 0.163 |
| run 2 | 1.0 | 0.0 | 0.339 | 0.289 | 0.170 | 0.121 | 0.289 | 0.045 | 0.068 |
| run 3 | 0.0 | 1.0 | 0.406 | 0.368 | 0.208 | 0.157 | 0.380 | 0.078 | 0.113 |

either of the other two runs, followed by the run that only uses the category scores which in turn performs significantly better than the worst performing run that only uses the linkrank scores.

The same trend among the four runs is observed across the three sub-tasks of the focused retrieval scenario, where again Zettair is able to better identify and retrieve the relevant information compared to the other three XER runs.

The obvious conclusion of our ad hoc experiments is that Zettair, which is specifically designed for full-text retrieval, performs better than our entity ranking system specifically designed for entity retrieval.

## 6 Conclusion and future work

We have presented our entity ranking approach for the INEX Wikipedia XML document collection which is based on exploiting the interesting structural and semantic properties of the collection.

On both the training and the testing data sets, we have shown that our entity ranking system outperforms the full-text search engine in the task of ranking entities. On the other hand, using our entity ranking system for ad-hoc retrieval did not result in any improvement over the full-text search engine. This confirms our hypothesis that the tasks of ad hoc retrieval and entity retrieval are two very different tasks.

Our entity ranking system was one of the best performing systems when comparing the entity ranking performances of all the participating systems in the INEX 2007 XER track. In the future, we aim at further developing our entity ranking algorithms by incorporating natural language processing techniques that we expect would reveal more potentially relevant entities.

## Acknowledgements

Part of this work was completed while James Thom was visiting INRIA in 2007.

## References


1. D. Awang Iskandar, J. Pehcevski, J. A. Thom, and S. M. M. Tahaghoghi. Social media retrieval using image features and structured text. In *Comparative Evaluation of XML Information Retrieval Systems: Fifth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2006*, volume 4518 of *Lecture Notes in Computer Science*, pages 358–372, 2007.
2. H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: efficient search on text, entities, and relations. In *Proceedings of the 30th ACM International Conference on Research and Development in Information Retrieval*, pages 671–678, Amsterdam, The Netherlands, 2007.
3. E. Blanchard, P. Kuntz, M. Harzallah, and H. Briand. A tree-based similarity for evaluating concept proximities in an ontology. In *Proceedings of 10th conference of the International Fedederation of Classification Societies*, pages 3–11, Ljubljana, Slovenia, 2006.
4. N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@noptic expert: searching for experts not just for documents. In *Proceedings of the Australasian Web Conference (Ausweb 01)*, Coffs Harbour, Australia, 2001.
5. A. P. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 Entity ranking track. In *INEX 2007 Workshop Proceedings (This volume)*, 2008.
6. L. Denoyer and P. Gallinari. The Wikipedia XML corpus. *SIGIR Forum*, 40(1):64–69, 2006.
7. S. Fissaha Adafre, M. de Rijke, and E. T. K. Sang. Entity retrieval. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP - 2007), September 27-29, Borovets, Bulgaria*, 2007.
8. G. Hu, J. Liu, H. Li, Y. Cao, J.-Y. Nie, and J. Gao. A supervised learning approach to entity search. In *Proceedings of the Asia Information Retrieval Symposium (AIRS 2006)*, volume 4182 of *Lecture Notes in Computer Science*, pages 54–66, 2006.
9. J. M. Kleinberg. Authoritative sources in hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
10. J. Pehcevski, A.-M. Vercoustre, and J. A. Thom. Exploiting locality of Wikipedia links in entity ranking. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR 2008)*, pages 258–269, Glasgow, Scotland, 2008.
11. I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, pages 32–51, 2006.
12. J. A. Thom, J. Pehcevski, and A.-M. Vercoustre. Use of Wikipedia categories in entity ranking. In *Proceedings of the 12th Australasian Document Computing Symposium*, pages 56–63, Melbourne, Australia, 2007.
13. 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 *INEX 2007 Workshop Proceedings (This Volume)*, 2008.
14. A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in Wikipedia. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC08)*, pages 1101–1106, Fortaleza, Brazil, 2008.
15. J. Zhu, D. Song, and S. Rueger. Integrating document features for entity ranking. In *INEX 2007 Workshop Proceedings (This Volume)*, 2008.