Overview of the INEX 2007 Entity Ranking Track

Arjen P. de Vries[1,2], Anne-Marie Vercoustre[3], James A. Thom[4], Nick Craswell[5], and Mounia Lalmas[6]

[1] CWI, Amsterdam, The Netherlands ⋆
[2] Technical University Delft, Delft, The Netherlands
[3] INRIA-Rocquencourt, Le Chesnay Cedex, France
[4] RMIT University, Melbourne, Australia
[5] Microsoft Research Cambridge, Cambridge, UK
[6] Queen Mary, University of London, London, UK ⋆⋆

Abstract. Many realistic user tasks involve the retrieval of specific entities instead of just any type of documents. Examples of information needs include 'Countries where one can pay with the euro' or 'Impressionist art museums in The Netherlands'. The Initiative for Evaluation of XML Retrieval (INEX) started the XML Entity Ranking track (INEX-XER) to create a test collection for entity retrieval in Wikipedia. Entities are assumed to correspond to Wikipedia entries. The goal of the track is to evaluate how well systems can rank entities in response to a query; the set of entities to be ranked is assumed to be loosely defined either by a generic category (entity ranking) or by some example entities (list completion). This track overview introduces the track setup, and discusses the implications of the new relevance notion for entity ranking in comparison to ad hoc retrieval.

1 Introduction

Information retrieval evaluation assesses how well systems identify information objects relevant to the user's information need. TREC has used the following working definition of relevance: 'If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant.' Here, a document is judged relevant if any piece of it is relevant (regardless of how small that piece is in relation to the rest of the document).

Many realistic user tasks seem however better characterised by a different notion of relevance. Often, users search for specific entities instead of just any type of documents. Examples of information needs include 'Countries where one can pay with the euro' or 'Impressionist art museums in The Netherlands', where the

⋆ This work is partially supported by the Dutch ICES/KIS III. BSIK project MultimediaN.

⋆⋆ This work is partially supported by DELOS, the European Network of Excellence in Digital Libraries.

entities to be retrieved are countries and museums; articles discussing the euro currency itself are not relevant, nor are articles discussing Dutch impressionist art.

To evaluate retrieval systems handling these *typed* information needs, the Initiative for Evaluation of XML Retrieval (INEX) started the XML Entity Ranking track (INEX-XER), with the aim to create a test collection for entity retrieval in Wikipedia. Section 2 provides details about the collection and assumptions underlying the track. Section 3 summarizes the results of the participants. Section 4 presents some findings related to the modified working definition of relevance, comparing entity ranking to ad hoc retrieval.

2 INEX-XER Setup

The main objective in the INEX-XER track is to return *entities* instead of 'just' web pages. The track therefore concerns triples of type `<category, query, entity>`. The category (that is *entity type*), specifies the type of 'things' to be retrieved. The query is a free text description that attempts to capture the information need. Entity specifies a (possibly empty) list of example instances of the entity type.

The usual information retrieval tasks of document and element retrieval can be viewed as special instances of this more general retrieval problem, where the category membership relates to a syntactic (layout) notion of 'text document', or, in the case of INEX ad hoc retrieval, 'XML element' or 'passage'. Expert finding uses the semantic notion of 'people' as its category, where the query would specify 'expertise on $\mathcal{T}$' for expert finding topic $\mathcal{T}$.

2.1 Data

The general case of retrieving entities (such as countries, people and dates) requires the estimation of relevance of items (i.e., instances of entities) that are not necessarily represented by text content other than their descriptive label [2]. INEX-XER 2007 approached a slightly easier sub-problem, where we restricted candidate items to those entities that have their own Wikipedia article. This decision simplifies not only the problem of implementing an entity ranking system (ignoring the natural language processing requirement of the general case), but, importantly, it also simplifies evaluation – as every retrieved result will have a proper description (its Wikipedia entry) to base the relevance judgement on.

The Wikipedia category metadata about entries has been exploited to loosely define entity sets. This category metadata is contained in the following files:

- `categories_name.csv` which maps category ids to category names
- `categories_hcategories.csv` which defines the category graph (which is not a strict hierarchy!)
- `categories_categories.csv` which maps article ids (that is pages that correspond to entities) to category ids

The entities in such a set are assumed to loosely correspond to those Wikipedia pages that are labeled with this category (or perhaps a sub-category of the given category). For example, considering the category 'art museums and galleries' (10855), an article about a particular museum such as the 'Van Gogh Museum' (155508) may be mapped to a sub-category like 'art museums and galleries in the Netherlands' (36697). Obviously, the correspondence between category metadata and the entity sets is far from perfect, as Wikipedia articles are often assigned to categories inconsistently. Since the human assessor of retrieval results is not constrained by the category assignments made in the corpus when making his or her relevance assessments, track participants have to handle the situation that the category assignments to Wikipedia pages are not always consistent, and also far from complete. correct answers may belong to other categories *close to* the provided one in the Wikipedia category graph, or may not have been categorized at all by the Wikipedia contributors. The challenge is to exploit a rich combination of information from text, structure and links for this purpose.

2.2 Tasks

In 2007, the track has distinguished two tasks, Entity Ranking and List Completion.

The motivation for the Entity Ranking task is to return entities that satisfy a topic described in natural language text. In other words, in the entity ranking task, the information need includes which category (entity type) is desired as answers. An Entity Ranking topic specifies the category identifier and the free-text query specification.[7] Results consist of a list of Wikipedia pages (our assumption is that all entities have a corresponding page in Wikipedia). For example, with 'Art museums and galleries' as the input category and a topic text 'Impressionist art in the Netherlands', we expect answers like the 'Van Gogh museum' and the 'Kröller-Müller museum'.

In the List Completion task, instead of knowing the desired category (entity type), the topic specifies between one and three correct entities (instances) together with a free-text context description. Results consist again of a list of entities (Wikipedia pages). As an example, when ranking 'Countries' with topic text 'European countries where I can pay with Euros', and entity examples such as 'France', 'Germany', 'Spain', then the 'Netherlands' would be a correct completion, but the 'United Kingdom' would not. Because the problem is to complete the partial list of answers, the given examples are considered non-relevant results in the evaluation of this task.

2.3 Topics

Figure 1 shows an example topic, developed by a sailing enthusiast. The INEX-XER topics can be used for both entity ranking and list completion tasks. When evaluating methods for entity ranking, the example entities given in the topic

[7] Multiple categories are allowed per topic.

```
<inex_topic topic_id="60" query_type="XER">
<title>olympic classes dinghy sailing</title>
<entities>
  <entity id="816578">470 (dinghy)</entity>
  <entity id="1006535">49er (dinghy)</entity>
  <entity id="855087">Europe (dinghy)</entity>
</entities>
<categories>
  <category id="30308">dinghies</category>
</categories>
<description>
The user wants the dinghy classes that are or have been olympic classes,
such as Europe and 470.
</description>
<narrative>
The expected answers are the olympic dinghy classes, both historic and
current. Examples include Europe and 470.
</narrative>
</inex_topic>
```

Fig. 1. Example topic

are of course not to be known by the entity ranking system. Likewise, in the list completion task, the category information would not be provided.

As mentioned before, Wikipedia categories define the entity type only loosely. Relevant entity answers may not belong to the specified category (in the corpus). Looking into the relevance assessments of the 2007 XER topics, we find that only 221 Wikipedia entries out of the total 996 relevant topic-entity pairs have at least one of the categories as given in the topic assigned in their metadata. For example, when ranking explorers in response to the information need 'Pacific navigators Australia explorers' (topic 65), some of the relevant Wikipedia entries have been labelled with categories 'explorers of australia' or 'explorers of the pacific' instead of topic category 'explorers'. Other relevant entities may have no category information at all. The category given in the topic should therefore be considered no more than an indication of what is expected, not a strict constraint (like in the CAS title for the ad hoc track).

2.4 The 2007 test collection

The created INEX-XER test collection provides training topics and testing topics.

A training set of 28 topics, based on a selection of 2006 ad hoc adapted to the entity task, has been kindly made available by INRIA (who developed this data) for participants to develop and train their systems. The relevance assessments have been derived from the articles judged relevant in 2006, limiting the set to the corresponding 'entities'. Of course, this procedure gives no guarantee that all the relevant entities have been assessed; this depends on completeness of the

Table 1. Entity ranking results, the run from each of the groups with the best MAP, sorted by MAP.

Team	Run	MAP
utwente	qokrwlin	0.306
inria	ER_comb-Q-TC-n5-a1-b8	0.293
uopen	er01	0.258
ukobe	qlm50_wwswitchlda800_fixed	0.227
utampere	er_2v2	0.210
uceg	ceger	0.191
unitoronto	single_nofilter	0.130
uhannover	qcs	0.123

ad hoc pool. Also, notice that the original title, description and narrative fields have not been updated to reflect the new entity ranking interpretation of the training topics.

The testing data consists of two parts. Topics 30–59 have been derived from the ad hoc 2007 assessments, similar to the way that the training data have been produced. For these topics, description and narrative may not be perfect, but they should be similar to the training topics. These topics have been assessed by track organizers (i.e., not by the original topic authors), with pools consisting of the articles that contained relevant information in the INEX ad hoc 2007 assessments. Of the originally proposed set of ad hoc derived topics, seven topics have been dropped because the ad hoc pools on which to base the XER assessments did not exist, and two topics have been dropped because their answer sets contained more than 50 relevant entities (and therefore we do not trust the original pools to be sufficiently complete). The final set consists of *21 ad hoc derived entity ranking test topics with assessments.*

Topics 60–100 are the genuine XER topics, created by participants specifically for the track. Almost all topics have been assessed by the original topic authors. From the originally proposed topics, we have dropped topics 93 (because it was too similar to topic 94) and topic 68 (because the underlying information need was identical to that of topic 61). Nine more topics were dropped because their answer sets contained more relevant entities than (or just about as many as) the pool depth (of 50), and two topics have been dropped because their assessments were never finished. The final set consists of *25 genuine entity ranking test topics with assessments* (that could eventually be expanded to 35 should we decide to perform more assessments).

3 Results

The eight participating groups submitted in total eighteen runs for the entity ranking task, and six participants submitted another ten list completion runs. Pools were constructed from the top 50 results for the two highest priority entity ranking and the two highest priority list completion runs. The pools contained on average about 500 entities per topic.

Table 2. List completion results, the run from each of the groups with the best MAP, sorted by MAP. The additional columns detail the number of examples that have been removed from the submitted run, and the MAP of the original submission.

Team	Run	MAP	#ex	MAP (uncorrected)
inria	LC_comb-Q-a2-b6	0.309	0	0.309
utwente	qolckrwlinfeedb	0.281	115	0.246
utampere	lc_1	0.247	1	0.246
unitoronto	single_EntityCats_d0_u0	0.221	102	0.198
uceg	ceglc	0.217	1	0.217
uopen	lc01	0.207	101	0.168

Table 1 presents the best results per group on the entity ranking task, in reverse order of mean average precision. Participants reported that runs exploiting the rich structure of the collection (including category information, associations between entities, and query-dependent link structure) have performed better than the baseline of plain article retrieval (see e.g. [1], as well as participant papers in this volume).

Table 2 summarizes the list completion results. The given topic examples are regarded non-relevant in the evaluation of the list completion task. Because several teams had by mistake included these given topic examples in their submissions, the Table lists the mean average precision of runs after removing the given examples from the submitted ranked lists. The number of topic examples removed and the original scores are given in the two remaining columns. Notice a minor anomaly: because we discovered in the judging phase (after submission) that one of the examples provided for topic 54 had been incorrect (i.e., non-relevant), the runs of some teams included the replacement entity (WP2892991, now a topic example but not at the time of submission).

4 Relevance in entity retrieval

Many ad hoc topics can serve as entity ranking topics, but the articles containing relevant passages must be re-assessed. These relevance assessments for the ad hoc derived training and testing topics provide therefore a basis to compare the notion of relevance for entity ranking to that used in ad hoc retrieval. On the eighteen ad hoc 2007 topics that were re-assessed as XER topics, only about 35% of the originally relevant documents have been assessed relevant. Depending on the topic, often surprisingly many articles that are *on topic* for the ad hoc track are not relevant entities. Also, articles that contain hubs (e.g., the Wikipedia ‘list of ...’ pages) are not entities, and not considered relevant.

Looking at two specific example topics to illustrate, only 6 out of the 129 ‘French president in the 5th republic’ relevant ad hoc results (XER topic 35 and ad hoc topic 448) are actually presidents, and only 32 out of the 267 ‘Bob Dylan songs’ relevant ad hoc results (XER topic 54 and ad hoc topic 509) are actually songs. The latter result set includes many articles related to Bob Dylan himself, The Band, albums, a documentary, a speech given at the March on Washington, singer-songwriters that play covers or tributes, cities where Bob Dylan lived, etc. Even though the ad hoc results do often contain a passage that mentions a song title, the XER model of the information need seems (arguably) closer to the 'real' user need; and, even in the ad hoc retrieval scenario, one would expect the entity results to be of higher relevance value (for this particular information need) than the remaining relevant pages.

While the track has not investigated the pool quality in-depth, the participants (and topic-authors) missed only few entities that they knew about (this could be a bigger issue with the ad hoc derived topics). In some specific cases, we validated the pool completeness using manually identified hubs in the collection, but no missing entities were found. We have found that runs for the list completion task contribute different relevant entities to the pools than the runs for entity ranking; an additional benefit from defining the two tasks from one entity retrieval problem.

5 Conclusions

The INEX 2007 XML Entity Ranking track has build the first test collection for the evaluation of information retrieval systems that support users that search for specific entities rather than just any type of documents. We developed a set of 28 training and 21 testing topics derived from ad hoc 2006 and 2007 topics, as well as a set of 25 genuine XER topics. The differences between the ad hoc relevance assessments and those of the entity ranking interpretation of the same topics demonstrate that the XER 2007 test collection captures (as expected) a different user need than ad hoc search, and a distinct interpretation of relevance. We would like to investigate in more detail whether system evaluation using genuine XER topics differs significantly from using the ad hoc derived topics, but this will require first the acquisition of larger topic sets; the main goal for the 2008 edition of the track. Aside from acquiring a larger number of topics for the current tasks on the Wikipedia collection, we will investigate how to evaluate searching for relations between entities. Another useful extension of the current track setup would be to allow ranking arbitrary passages as result entities, instead of limiting the possible answers to Wikipedia entries only.

References

1. J. Pehcevski, A.-M. Vercoustre, and J. Thom. Exploiting locality of wikipedia links in entity ranking. In *Advances in Information Retrieval. 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956/2008 of *LNCS*, pages 258–269, 2008.
2. H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *Proceedings of the 16th ACM CIKM International Conference on Information and Knowledge Management*, pages 1015–1018, Lisbon, Portugal, 2007.