Exploiting Locality of Wikipedia Links in Entity Ranking

Jovan Pehcevski[1] Anne-Marie Vercoustre[1] James Thom[2]

[1]INRIA
Rocquencourt, France
{jovan.pehcevski, anne-marie.vercoustre}@inria.fr

[2]RMIT University
Melbourne, Australia
james.thom@rmit.edu.au

Overview

- Entity extraction and ranking
- Wikipedia XML document collection
- INEX 2007 XML Entity Ranking (XER) track
- Our approach to entity ranking
- Experimental results
- Comparison to other XER systems
- Conclusions and future work

Named entities

What are (named) entities?

- Names of people / organisations
- Locations
- Dates
- Events
- Drug names / symptoms
- Titles of movies / books
- ...

Entity-related research

A very active research domain:

- Entity extraction/tagging from texts
- Entity reference solving (“The president of the Republic”)
- Entity disambiguation (which Michael Jackson)
- Question-answering (who/when/where/what)
- Expert search (names of experts in information retrieval)
- Entity ranking (entity retrieval)

INRIA

Entity extraction

- Main goal: extracting or tagging entities in texts
- Two main approaches:
  - Grammar-based: efficient but requires many rules written by hand (experts)
  - Statistical model-based: more flexible, but requires large collections for training

Entity ranking

- Main goal: return a list of relevant entities for a query
- Example:
  - Query: “European countries where I can pay with Euros”
  - Results: a **list of entities** representing relevant countries

Ranking people

- Expert search task in the TREC Enterprise Track
- Collection:
  - Corpus: crawl of *.w3.org sites
  - People: names of 1092 people who may be experts
- Query: "information retrieval experts"
- Results: a **list of people** who are experts in information retrieval

Ranking famous actors

- Results are **lists of famous actors**. For example:
  - Query: “Actors in the 1930s”
  - Results: Fred Astaire, Charlie Chaplin, W.C. Fields, Errol Flynn, Clark Gable, Greta Garbo, etc.
  - Query: “Actors in action movies”
  - Results: Arnold Schwarzenegger, Jean-Claude Van Damme, etc.

Ranking ...

- People
- Actors
- {... insert your favourite entity type here ...}

Entity Ranking!!!

Some entity ranking scenarios

- Impressionist art museums in The Netherlands
- Countries with the Euro currency
- German car manufacturers
- Artists related to Pablo Picasso
- Actors who played Hamlet
- English monarchs who married French women

INEX

- INEX: INitiative for the Evaluation of XML retrieval
- The INEX test collection comprises an XML document collection, a set of topics and a set of relevance assessments that correspond to those topics
- INEX XML document collections:
  - 2002–2004: 12,107 IEEE Computer Society articles, covering the period of 1995–2002, totalling 494 MB in size with 8 million XML elements
  - 2005: 16,819 IEEE Computer Society articles, covering the period of 1995–2004, totalling 764 MB in size with 11 million XML elements
  - 2006–2007: 659,388 Wikipedia articles (distributed across 113,483 categories), covering the period until 2006, totalling 4.6 GB in size with 52 million elements

Wikipedia XML document collection

- An XML-based corpus based on a snapshot of the Wikipedia in 2006
- Used by various INEX tracks in 2006 and 2007
- Structural features:
  - Entity pages
  - Links to entity pages from entity occurrences
  - Lists of entity co-occurrences in pages
- Semantic features:
  - Categories attached to entity pages

Entities in Wikipedia

- Examples of different types of entities: Art museums and galleries, Countries, Famous people (actors, writers, explorers), Magicians, Diseases, Movies, Songs, Books, etc. (nearly everything)
- Extract from the Euro page:

"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."

Categories in Wikipedia

- Wikipedia categories are associated to entity pages (on average, 2.28 categories are associated to a page)
- They have unique names (e.g. “france”, “european countries”, “countries”), and can have multiple sub-categories and parent-categories
- New categories can also be created by authors, although they have to follow Wikipedia recommendations
- For example, the Spain page is associated with the following categories: “spain”, “european union member states”, “spanish-speaking countries” and “constitutional monarchies” (among others)

INEX and entity ranking

- New track at INEX 2007 on XML Entity Ranking
- Using the XML version of Wikipedia as a document collection
- Two entity ranking tasks:
  - Task 1: *Entity Ranking*, which aims at retrieving entities of a given category that satisfy a topic described in natural language text
  - Task 2: *List Completion*, where given a topic text and a small number of entity examples, the aim is to complete this partial list of answers

Example INEX 2007 entity ranking topic

```
<inex_topic>
<title>
European countries where I can pay with Euros
</title>
<description>
I want a list of European countries where I can pay with Euros.
</description>
<narrative>
Each answer should be the article about a specific European country that uses the Euro
as currency.
</narrative>
<entities>
   <entity ID="10581">France</entity>
   <entity ID="11867">Germany</entity>
   <entity ID="26667">Spain</entity>
</entities>
<categories>
   <category ID="185">european countries</category>
</categories>
</inex_topic>
```

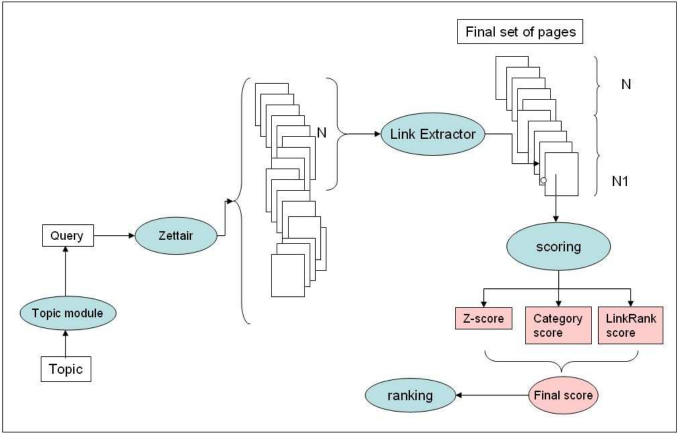
Our approach to entity ranking

Our approach to identifying and ranking entities combines:

1. The full-text similarity of the entity page with the query;
2. The similarity of the page's categories with the categories of entity examples; and
3. The link contexts found around entity examples in the top ranked pages returned by a search engine

We focus on Task 2, although our approach can be used (in a simplified form) in Task 1.

Architecture



Score functions and parameters

- The global score of an entity page is derived by combining three normalised scores:
  - a **Z score**, based on the initial Zettair score assigned to the page
  - a **Category score**, based on the ratio of common categories found for the page and the entity examples
  - a **Linkrank score**, based on the number of links to the page, from the first N pages returned by the search engine
- The global score is then calculated as a linear combination of the three normalised scores

Z score

- The Z score assigns the initial Zettair score to a target entity page t:

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

Category score

- The Category score reflects the ratio of common categories between the set of categories associated to the target entity page cat(t) and the set of the union of the categories associated to the entity examples cat(E):

$$S_C(t) = \frac{|\text{cat}(t) \cap \text{cat}(E)|}{|\text{cat}(E)|}$$

Linkrank score

- The Linkrank score 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 number of links that point to the target page from the referring page $\#links(p, t)$:

$$S_L(t) = \sum_{r=1}^{N} z(p_r) \cdot g(\#ent(p_r)) \cdot f(\#links(p_r, t))$$

where $g(x) = x + 0.5$, and $f(x) = x$

Preliminary results

	Search Engine Query only (Q)	Search Engine Query + examples (QE)	QE + Linkrank score	QE + Category score
1	Eurobilltracker	2000 European Football …	Euro	France
2	Euro	Euro banknotes	France	Germany
3	Economic and Monetary Union	2004 European Football …	Italy	Italy
4	Euro banknotes	Eurochart Hot 100 Singles	Spain	Spain
5	Euro coins	Euro	Germany	Netherlands
6	European Monetary…	Currency	Netherlands	Finland
7	Eurozone	Eurozone	Austria	United Kingdom
8	Eonia	Special member state territories	Finland	Belgium
9	Eurojargon	Monetary union	United Kingdom	Portugal
10	Currency bill tracking	Latin Europe	Belgium	Denmark

Global score

- The global score for a target entity page t is calculated as a linear combination of the Z score, the Category score and the Linkrank score:

$$S(t) = \alpha S_L(t) + \beta S_C(t) + (1 - \alpha - \beta) S_Z(t)$$

- Special cases:
  - $\alpha = 0, \beta = 0$, which uses only the Z score
  - $\alpha = 0, \beta = 1$, which uses only the Category score
  - $\alpha = 1, \beta = 0$ which uses only the Linkrank score

Preliminary results ...

QE + combining 2 scores: α=0.5; β=0.5; (1-α-β)=0	QE + combining 2 scores: α=0.2; β=0.8; (1-α-β)=0 (more on categories)	QE + combining 2 scores: α=0.8; β=0.2; (1-α-β)=0 (more on links)	QE + combining 3 scores: α=0.33; β=0.33; (1-α-β)=0.33
France	France	France	France
Germany	Germany	Euro	Euro
Spain	Spain	Italy	Germany
Italy	Netherlands	Germany	Spain
Netherlands	Finland	Spain	Andorra
Euro	United Kingdom	Netherlands	Austria
Finland	Belgium	Finland	Denmark
United Kingdom	Portugal	Austria	Eurozone
Belgium	Denmark	United Kingdom	Italy
Portugal	Italy	Belgium	Netherlands

Identifying link contexts

- Assumption: target entities located in close proximity to the entity examples are more likely to represent relevant entities than those that appear in other parts of the page
- Three types of contexts:
    - Full page context
    - Element contexts:
        - *Static* contexts, which are predefined types of elements such as paragraphs, lists and tables
        - *Dynamic* contexts, which are determined dynamically by utilising the underlying XML document structure

Extract from the Euro page

"The **euro** . . . is the official currency of the Eurozone (also known as the Euro Area or the Euro Land), which consists of 13 European states (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."

"All nations that have joined the EU since the 1993 implementation of the Maastricht Treaty have pledged to adopt the euro in due course. Maastricht obliged current members to join the euro; however, the United Kingdom and Denmark negotiated exemptions from that requirement for themselves."

Static contexts

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

- Three types of list-like elements: paragraphs (tag `p`); lists (tags `normallist`, `numberlist` and `definitionlist`); and tables (tag `table`)
- Two algorithms: `StatL` (shown in bold) and `StatR` (shown in underline)
- Advantage: list-like contexts are easy to identify
- Drawback: collection-dependent

Dynamic contexts

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

- We adapted the concept of a Coherent Retrieval Element (CRE): the *lowest common ancestor* (LCA) of at least two entity examples
- One algorithm: `DynCRE` (shown in bold)
- Advantage: collection-independent
- Drawback: list-like contexts containing exactly one entity example are never identified

New Linkrank score

- The new Linkrank score 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 set of link contexts identified around entity examples in the referring page $C(p)$:

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

where $g(x) = x + 0.5$, and

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

Preliminary results

QE + Linkrank (Full page context)	QE + Linkrank (Static context)	QE + Linkrank (Dynamic context)
Euro	France	France
France	Italy	Italy
Italy	Spain	Spain
Spain	Germany	Germany
Germany	Netherlands	Netherlands
Netherlands	Finland	Finland
Austria	Austria	Austria
Finland	Portugal	Portugal
United Kingdom	Belgium	Belgium
Belgium	Greece (UK: rank 14)	Greece (UK: rank 15)

Experimental results

Results that investigate the effectiveness of our entity ranking approach using the INEX Wikipedia document collection

- Test collection
- Full page context
- Static and dynamic contexts

Test collection

- There was no existing set of topics with relevance assessments for entity ranking prior to INEX 2007
- We developed a test collection based on a selection of topics from the INEX 2006 ad hoc track
- We use 28 topics considered to be of an "entity ranking" nature (including the Euro topic example)
- We use mean average precision (MAP) as our primary method of evaluation, but also report results using several alternative measures (P[5], P[10] and R-prec)

Full page context

Table: Mean average precision scores for runs using 66 possible α–β combinations, obtained on the 28 topics of our test collection. Queries sent to Zettair include only terms from the topic title (Q). The MAP score of the plain Zettair run is 0.172.

	Beta										
Alpha	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	*0.178*	0.204	0.234	0.267	0.302	0.313	0.326	0.334	0.344	0.349	*0.309*
0.1	0.188	0.212	0.254	0.285	0.312	0.327	0.341	0.351	**0.357**	0.316	
0.2	0.194	0.218	0.251	0.291	0.324	0.331	0.345	0.348	0.309		
0.3	0.201	0.227	0.248	0.273	0.303	0.321	0.324	0.291			
0.4	0.209	0.233	0.253	0.275	0.285	0.287	0.262				
0.5	0.217	0.236	0.249	0.261	0.265	0.244					
0.6	0.212	0.227	0.243	0.240	0.227						
0.7	0.205	0.215	0.218	0.202							
0.8	0.188	0.186	0.175								
0.9	0.164	0.153									
1.0	*0.131*										

Full page context ...

Table: Performance scores for runs using the context of the full page.

Run	P[r] 5	P[r] 10	R-prec	MAP
Zettair	0.229	0.232	0.208	0.172
α0.0–β0.0	0.229	0.232	0.213	0.178
α0.0–β1.0	0.364	0.307	0.315	0.309
α1.0–β0.0	0.157	0.157	0.138	0.131
α0.1–β0.8	**0.471**	**0.386**	**0.390**	**0.357**
α0.2–β0.6	0.436	0.377	0.375	0.345

(Q) Topic title

Run	P[r] 5	P[r] 10	R-prec	MAP
Zettair	0.200	0.171	0.157	0.143
α0.0–β0.0	0.200	0.171	0.177	0.153
α0.0–β1.0	0.336	0.282	0.275	0.267
α1.0–β0.0	0.186	0.175	0.159	0.152
α0.1–β0.8	0.336	0.329	0.311	0.314
α0.2–β0.6	**0.343**	**0.336**	**0.336**	**0.324**

(QE) Topic title and entity examples

- Adding names of the entity examples to the query generally performs worse for all but the linkrank module
- Different optimal values are observed for the two parameters in the two tables
- The best entity ranking approaches are those that combine the ranking evidence from the three modules

Static and dynamic contexts

Table: Performance scores for runs using different types of contexts in the linkrank module (α1.0–β0.0).

Run	P[r] 5	P[r] 10	R-prec	MAP
`FullPage`	0.157	0.157	0.138	0.131
`StatL`	0.214	**0.225**	**0.228**	**0.190**
`StatR`	**0.221**	0.214	0.219	0.185
`DynCRE`	**0.221**	0.211	0.215	0.183

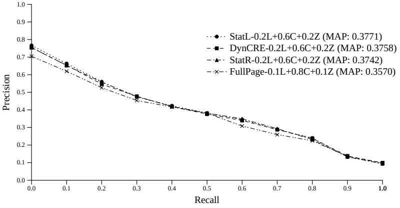
(Q) Topic title

Run	P[r] 5	P[r] 10	R-prec	MAP
`FullPage`	0.186	0.175	0.159	0.152
`StatL`	0.243	0.218	**0.226**	0.203
`StatR`	0.243	**0.221**	0.225	**0.204**
`DynCRE`	**0.257**	0.211	0.221	0.194

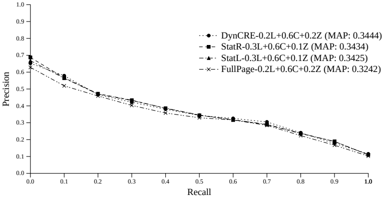
(QE) Topic title and entity examples

- There is a significant performance improvement ($p < 0.05$) for the three runs that use element (static and dynamic) contexts, irrespective of the type of query used
- The `StatL` run performs best in the case of Q, while the `StatR` run performs best in the case of QE
- The `DynCRE` run achieves the best early precision but overall performs worst among the three runs

Static and dynamic contexts ...



(Q) Topic title



(QE) Topic title and entity examples

- For the three runs using element contexts the optimal value for α has shifted from 0.1 to 0.2 (in the case of Q), while for the two static runs the optimal α value was 0.3 (in the case of QE)
- The optimal value for β is 0.6 (a shift from 0.8 when using the full page context)
- The three optimal runs using element contexts substantially outperform the optimal run using the full page context

Comparison to other XER systems

INEX 2007 XER

Entity Ranking | **List Completion**

Run	MAP (no ex)	#ex	MAP (submitted)
inria_LC_comb-Q-a2-b6.txt	**0.3092**	0	0.3092
utwente_qolckrwlinfeedb	**0.2805**	115	0.2463
utwente_qolcrwfeedb	0.2791	115	0.2449
utampere_lc_1	**0.2465**	1	0.2463
single_EntityCats_d0_u0.txt	**0.2206**	102	0.1975
utampere_lc_2	0.2194	1	0.2188
au_ceglc.txt	**0.2168**	1	0.2167
ou_lc01	**0.2072**	101	0.1678
combined_EntityCats_d0_u0.txt	0.1913	93	0.1682
inria_LC_comb-Q-a0-b10.txt	0.1909	0	0.1909

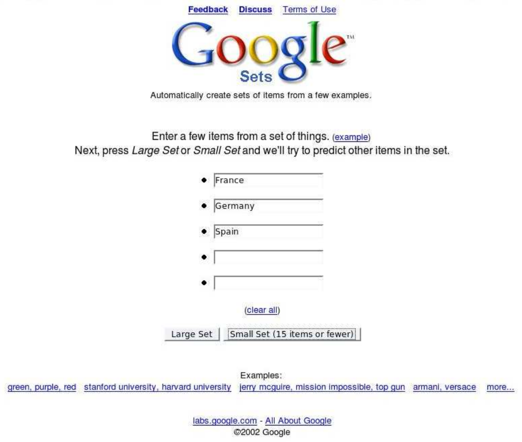
Figure: INEX07 Task 2 (List Completion) scores

INRIA

Conclusions and future work

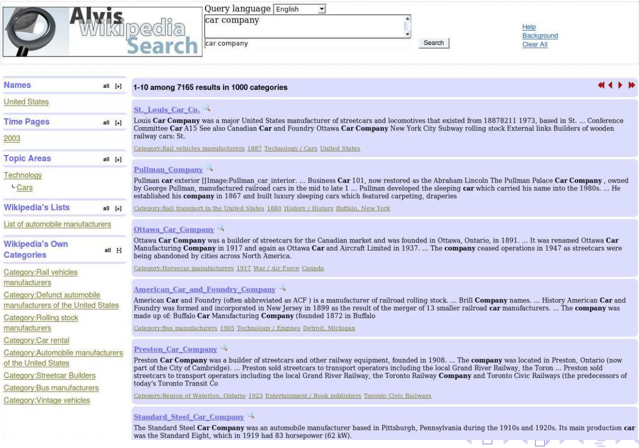
- We have presented our entity ranking approach for the INEX Wikipedia XML document collection, which exploits the interesting structural and semantic properties of the collection
- The use of categories and locality of Wikipedia links around entity examples results in an improved effectiveness of entity ranking
- In the future, we plan to further improve our linkrank algorithm by incorporating relevance feedback with natural language processing techniques that we expect would reveal more potentially relevant entities
- To improve the category similarity function, we plan to introduce different category weighting rules
- We will also carry out a detailed per-topic error analysis, which should reveal the effect of topic type on entity ranking

Appendix A: Applications - Google Sets



Appendix A: Applications - Google Sets ...

Appendix A: Applications - Alvis



Appendix B: Task 1

- Motivation: Wikipedia is not an ontology!
- Some properties of Wikipedia categories include:
  - a category may have many sub-categories and parent-categories;
  - some categories have many associated pages (i.e. large *extension*), while others have smaller extension;
  - a page that belongs to a given category extension generally does not belong to its ancestors' extension;
  - the sub-category relation is not always a subsumption (*is-a*) relationship; and
  - there are cycles in the category graph.

Appendix B: Task 1 functions

- We define a similarity function between target categories cat(C) and categories attached to answer entities cat(t):

$$S_C(t) = \frac{|\mathrm{cat}(t) \cap \mathrm{cat}(C)|}{|\mathrm{cat}(C)|}$$

- We use extensions based on sub-categories and parent-categories in the graph of Wikipedia categories: $\mathrm{cat}_d(C)$ and $\mathrm{cat}_u(t)$
- We also use lexical similarity between categories, by indexing all the categories using either their names or their names and the names of their attached entities as corresponding documents
- We retrieve the top M ranked categories: Ccat(C), Tcat(C), and TCcat(C)

Appendix B: Task 1 results

Table: Performance scores for runs using different strategies in our category similarity module (α0.0–β1.0). The number of Zettair category answers is M=10.

Run	P[r] 5	P[r] 10	R-prec	MAP	Run	P[r] 5	P[r] 10	R-prec	MAP
cat(C)-cat(t)	0.229	0.250	0.215	0.196	cat(C)-cat(t)	0.229	**0.250**	**0.215**	**0.196**
cat_d(C)-cat_u(t)	0.243	0.246	0.209	0.185	cat_d(C)-cat_u(t)	**0.243**	0.246	0.209	0.185
Ccat(C)-cat(t)	0.214	0.250	0.214	0.197	Ccat(C)-cat(t)	0.157	0.171	0.149	0.148
Tcat(C)-cat(t)	**0.264**	0.261	0.239	0.216	Tcat(C)-cat(t)	0.171	0.182	0.170	0.157
TCcat(C)-cat(t)	**0.264**	**0.286**	**0.247**	**0.226**	TCcat(C)-cat(t)	0.207	0.214	0.175	0.173

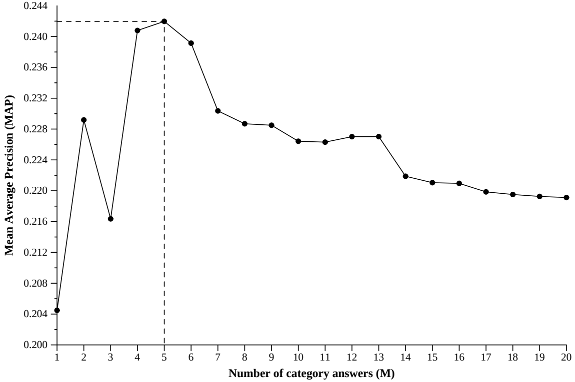
(C) Index of category names — (CE) Index of category and entity names

- Overall, using category extensions does not result in an improved performance
- The choice of using the Zettair category index can dramatically influence the entity ranking performance
- The run that uses the query that combines the terms from the title and the category fields of an INEX topic (TCcat(C)-cat(t)) performs the best among the three runs using lexical similarity

INRIA

Appendix B: Task 1 results ...

- Investigating the parameter M using the run TCcat(C)-cat(t)



Appendix B: Task 1 results ...

- We also investigated the optimal values for combining parameters α and β
- We used the run TCcat(C)-cat(t) with the optimal value M=5 and the Zettair index of category names
- We calculated MAP over the 28 topics in our test collection, as we varied α from 0 to 1 in steps of 0.1; for each value of α, we also varied β from 0 to $(1 - \alpha)$ in steps of 0.1
- We found that the highest MAP score (0.287) is achieved for $\alpha = 0.1$ and $\beta = 0.8$, which is a 19% relative performance improvement over the best score (0.242) achieved by using only the category module (α0.0–β1.0)
- This performance improvement is statistically significant ($p < 0.05$)

Appendix B: Comparison to other XER systems

INEX 2007 XER

Entity Ranking | List Completion

Run	MAP
utwente_qokrwlin	**0.3061**
utwente_qoirw	0.3015
inria_ER_comb-Q-TC-n5-a1-b8.txt	**0.2934**
ou_er01	**0.2582**
ou_er02	0.2306
ou_er03	0.2306
ukobe_qlm50_wwswitchlda800_fixed.txt	**0.2273**
ukobe_qlm50_wwswitchlda400_fixed.txt	0.219
utampere_er_2v2	**0.2098**
inria_ER_comb-Q-TC-n5-a0-b10.txt	0.2046
au_ceger.txt	**0.1909**
utampere_er_1v2	0.1803
ukobe_qlm50_wwswitchlda800.txt	**0.1724**
ukobe_qlm50_wwswitchlda400.txt	0.1664
ukobe_qlm50.txt	0.1562
single_nofilter.txt	**0.1303**
l3s_qcs.txt	**0.1225**
ukobe_wwswitchlda800_fixed.txt	0.0614
combined_nofilter.txt	0.053
ukobe_wwswitchlda800.txt	0.0233

Figure: INEX07 Task 1 (Entity Ranking) scores