# Focused Search in Books and Wikipedia: Categories, Links and Relevance Feedback

Marijn Koolen[1], Rianne Kaptein[1], and Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

**Abstract.** In this paper we describe our participation in INEX 2009 in the Ad Hoc Track, the Book Track, and the Entity Ranking Track. In the Ad Hoc track we investigate focused link evidence, using only links from retrieved sections. The new collection is not only annotated with Wikipedia categories, but also with YAGO/WordNet categories. We explore how we can use both types of category information, in the Ad Hoc Track as well as in the Entity Ranking Track. Results in the Ad Hoc Track show Wikipedia categories are more effective than WordNet categories, and Wikipedia categories in combination with relevance feedback lead to the best results. Preliminary results of the Book Track show full-text retrieval is effective for high early precision. Relevance feedback further increases early precision. Our findings for the Entity Ranking Track are in direct opposition of our Ad Hoc findings, namely, that the WordNet categories are more effective than the Wikipedia categories. This marks an interesting difference between ad hoc search and entity ranking.

## 1 Introduction

In this paper, we describe our participation in the INEX 2009 Ad Hoc, Book, and Entity Ranking Tracks. Our aims for this year were to familiarise ourselves with the new Wikipedia collection, to continue the work from previous years, and to explore the opportunities of using category information, which can be in the form of Wikipedia's categories, or the enriched YAGO/WordNet categories.

The rest of the paper is organised as follows. First, Section 2 describes the collection and the indexes we use. Then, in Section 3, we report our runs and results for the Ad Hoc Track. Section 4 briefly discusses our Book Track experiments. In Section 5, we present our approach to the Entity Ranking Track. Finally, in Section 6, we discuss our findings and draw preliminary conclusions.

## 2 Indexing the Wikipedia Collection

In this section we describe the index that is used for our runs in the ad hoc and the entity ranking track, as well as the category structure of the collection. The collection is based, again, on the Wikipedia but substantially larger and with longer articles. The original Wiki-syntax is transformed into XML, and each article is annotated using "semantic" categories based on YAGO/Wikipedia. We used Indri [15] for indexing and retrieval.

### 2.1 Indexing

Our indexing approach is based on earlier work [1, 4, 6, 12, 13, 14].

- *Section index*: We used the `<section>` element to cut up each article in sections and indexed each section as a retrievable unit. Some articles have a leading paragraph not contained in any `<section>` element. These leading paragraphs, contained in `<p>` elements are also indexed as retrievable units. The resulting index contains no overlapping elements.
- *Article index*: We also build an index containing all full-text articles (i.e., all wikipages) as is standard in IR.

For all indexes, stop-words were removed, and terms were stemmed using the Krovetz stemmer. Queries are processed similar to the documents. In the ad hoc track we use either the CO query or the CAS query, and remove query operators (if present) from the CO query and the about-functions in the CAS query.

### 2.2 Category Structure

A new feature in the new Wikipedia collection is the assignment of WordNet labels to documents [11]. The WordNet categories are derived from Wikipedia categories, but are designed to be conceptual. Categories for administrative purposes, such as 'Article with unsourced statements', categories yielding non-conceptual information, such as '1979 births' and categories that indicate merely thematic vicinity, such as 'Physics', are not used for the generation of WordNet labels, but are excluded by hand and some shallow linguistic parsing of the category names. WordNet concepts are matched with category names and the category is linked to the most common concept among the WordNet concepts. It is claimed this simple heuristic yields the correct link in the overwhelming majority of cases.

A second method which is used to generate WordNet labels, is based on information in lists. For example, If all links but one in a list point to pages belonging to a certain category, this category is also assigned to the page that was not labelled with this category. This is likely to improve the consistency of annotation, since annotation in Wikipedia is largely a manual effort.

We show the most frequent category labels of the two category structures in Table 1. Many of the largest categories in Wikipedia are administrative categories. The category *Living people* is the only non-administrative label in this list. The largest WordNet categories are more semantic, that is, they describe what an article is about. The list also shows that many Wikipedia articles are about entities such as persons and locations.

Table 1: The most frequent categories of the Wikipedia and WordNet structure

| **Wikipedia** | | **Wordnet** | |
|---|---|---|---|
| Living people | 307,317 | person | 438,003 |
| All disambiguation pages | 143,463 | physical entity | 375,216 |
| Disambiguation pages | 103,954 | causal agent | 373,697 |
| Articles with invalid date parameter in template | 77,659 | entity | 245,049 |
| All orphaned articles | 34,612 | location | 155,304 |
| All articles to be expanded | 33,810 | region | 146,439 |
| Year of birth missing (living people) | 32,503 | artifact | 131,248 |
| All articles lacking sources | 21,084 | player | 109,427 |

Table 2: The distribution of Wikipedia and WordNet categories over articles

| **cats/article** | **N** | **Min** | **Max** | **Med.** | **Mean** | **St.dev** |
|---|---|---|---|---|---|---|
| *Wikipedia* | 2,547,560 | 1 | 72 | 3 | 3.50 | 2.82 |
| *WordNet* | 2,033,848 | 1 | 41 | 3 | 3.98 | 3.18 |

Table 3: The distribution of articles over Wikipedia and WordNet categories

| **articles/cat** | **N** | **Min** | **Max** | **Med.** | **Mean** | **St.dev** |
|---|---|---|---|---|---|---|
| *Wikipedia* | 346,396 | 1 | 307,317 | **5** | **26** | 643 |
| *WordNet* | 5,241 | 1 | 438,003 | **57** | **1,546** | 12,087 |

### 2.3 Comparing the Category Structures

We first analyse the difference between the two category structures by comparing the number of categories assigned to each article in Table 2. In total, over 2.5 million articles have at least one Wikipedia category and just over 2 million articles have at least one WordNet category. We see that most articles have up to 3 or 4 Wikipedia or WordNet categories. The highest number of categories assigned is somewhat higher for Wikipedia (72) than for WordNet (41). There seem to be no big differences between the distributions of the two category structures.

In Table 3 we show statistics of the number of articles assigned to each category. The most salient difference is the total number of categories. There are 346,396 Wikipedia categories and only 5,241 WordNet categories. As a direct result of this and the statistics of Table 2, most of the WordNet categories are much bigger than the Wikipedia categories. On average, a Wikipedia category has 26 articles, while a WordNet category has 1,546 articles. The median size of both structures is much smaller, indicating a skewed distribution, but we observe the same pattern. 50% of the WordNet categories have at least 57 articles, while 50% of the Wikipedia categories has at most 5 articles. The Wikipedia category structure is thus more fine-grained than the WordNet structure.

## 3 Ad Hoc Track

For the INEX 2009 Ad Hoc Track we had two main aims. Investigating the value of element level link evidence, and the relative effectiveness of the Wikipedia and WordNet category structures available in the new INEX 2009 Wikipedia collection.

In previous years [2], we have used local link degrees as evidence of topical relevance. We took the top 100 retrieved articles, and computed the link degrees using all the links between those retrieved articles. This year, instead of looking at all local links between the top 100 retrieved articles, we consider only the links occurring in the retrieved elements. A link from article $A$ to article $B$ occurring in a section of article $A$ that is not retrieved is ignored. This link evidence is more focused on the search topic and possibly leads to less infiltration. Infiltration occurs when important pages with many incoming links are retrieved in the top 100 results. Because of their high global in-degree, they have a high probability of having links in the local set. The resulting local link degree is a consequence of their query-independent importance and pushes these documents up the ranking regardless of their topical relevance. If we use only the relevant text in document to derive link evidence, we reduce the chance of picking up topically unrelated link evidence.

The new INEX Wikipedia collection has markup in the form of YAGO elements including WordNet categories. Most Wikipedia articles are manually categorised by the Wikipedia contributors. The category structure can be used to generate category models to promote articles that belong to categories that best match the query. We aim to directly compare the effectiveness of category models based on the Wikipedia and WordNet categorisations for improving retrieval effectiveness.

We will first describe our approach and the official runs, and finally per task, we present and discuss our results.

### 3.1 Approach

We have four baseline runs based on the indexes described in the previous section:

**Article** : run on the article index with linear length prior and linear smoothing $\lambda = 0.15$.
**Section** : run on the section index with linear length prior and linear smoothing $\lambda = 0.15$.
**Article RF** : run on the article index with blind relevance feedback, using 50 terms from the top 10 results.
**Section RF** : run on the section index with blind relevance feedback, using 50 terms from the top 10 results.

These runs have up to 1,500 results per topic. All our official runs for all four tasks are based on these runs. To improve these baselines, we explore the following options.

**Category distance** : We determine two target categories for a query based on the top 20 results. We select the two most frequent categories to which the top 20 results are assigned and compute a category distance score using parsimonious language models of each category. This technique was successfully employed on the INEX 2007 Ad hoc topics by Kaptein et al. [8]. In the new collection, there are two sets of category labels. One based on the *Wikipedia* category structure and one based on the *WordNet* category labels.

**CAS filter** : For the CAS queries we extracted from the CAS title all semantic target elements, identified all returned results that contain a target element in the xpath and ranked them before all other results by adding a constant $c$ to the score per matching target element. Other than that, we keep the ranking in tact. A result that matches two target elements gets $2c$ added to its score, while a result matching one target element gets $1c$ added to its score. In this way, results matching $n$ target elements are ranked above results matching $n - 1$ target elements. This is somewhat similar to coordination level ranking of content-only queries, where documents matching $n$ query terms are ranked above documents matching $n - 1$ query terms. Syntactic target elements like `<article>`, `<sec>`, `<p>` and `<category>` are ignored.

**Link degrees** : Both incoming and outgoing link degrees are useful evidence in identifying topical relevance [5, 10]. We use the combined $indegree(d) + outdegree(d)$ as a document "prior" probability $P_{link}(d)$. Local link evidence is not query-independent, so $P_{link}(d)$ is not an actual *prior* probability. We note that for runs where we combine the article or section text score with a category distance score, we get a different score distribution. With these runs we use the link evidence more carefully by taking the log of the link degree as $P_{link}(d)$. In a standard language model, the document prior is incorporated as $P(d|q) = P_{link}(d) \cdot P_{content}(q|d)$, where $P_{content}(q|d)$ is the standard language model score.

**Focused Link degrees** : We also constructed a focused local link graph based on the retrieved elements of the top 100 articles. Instead of using all links between the top 100 articles, we only use the outgoing links from sections that are retrieved for a given topic. The main idea behind this is that link anchors appearing closer to the query terms are more closely related to the search topic. Thus, if for an article $a_i$ in the top 100 articles only section $s_j$ is retrieved, we use only the links appearing in section $s_j$ that point to other articles in the top 100. This local link graph is more focused on the search topic, and potentially suffers less from infiltration of important but off-topic articles. Once the focused local link graph is constructed, we count the number of incoming + outgoing links as the focused link prior $P_{foc_link}(d)$.

**Article ranking** : based on [4], we use the article ranking of an article index run and group the elements returned by a section index run as focused results.

**Cut-off(n)** : When we group returned elements per article for the Relevant in Context task, we can choose to group all returned elements of an article, or only the top ranked elements. Of course, further down the results list we find less relevant elements, so grouping them with higher ranked elements from the same article might actually hurt precision. We set a cut-off at rank $n$ to group only the top returned elements by article.

### 3.2 Runs

Combining the methods described in the previous section with our baseline runs leads to the following official runs.

For the Thorough Task, we submitted two runs:

**UamsTAdbi100** : an article index run with relevance feedback. The top 100 results are re-ranked using the link degree prior $P_{link}(d)$.

**UamsTSdbi100** : a section index run with relevance feedback. We cut off the results list at rank 1500 and re-rank the focused results of the top 100 articles using the link prior $P_{link}(d)$. **However, this run is invalid, since it contains overlap due to an error in the xpaths.**

For the Focused Task, we submitted two runs:

**UamsFSdbi100CAS** : a section index run combined with the Wikipedia category distance scores. The results of the top 100 articles are re-ranked using the link degree prior. Finally, the CAS filter is applied to boost results with target elements in the xpath.

**UamsFSs2dbi100CAS** : a section index run combined with the Wikipedia category distance scores. The results of the top 100 articles are re-ranked using the focused link degree prior $P_{foc_link}(d)$.

For the Relevant in Context Task, we submitted two runs:

**UamsRSCMACMdbi100** : For the article ranking we used the article text score combined with the manual category distance score as a baseline and re-ranked the top 100 articles with the log of the local link prior $P_{link}(d)$. The returned elements are the top results of a combination of the section text score and the manual category distance score, grouped per article.

**UamsRSCWACWdbi100** : For the article ranking we used the article text score combined with the WordNet category distance score as a baseline and re-ranked the top 100 with the log of the local link prior $P_{link}(d)$. The returned elements are the top results of a combination of the section text score and the WordNet category distance score, grouped per article.

For the Best in Context Task, we submitted two runs:

**UamsBAfbCMdbi100** : an article index run with relevance feedback combined with the Wikipedia category distance scores, using the local link prior $P_{link}(d)$ to re-rank the top 100 articles. The Best-Entry-Point is the start of the article.

**UamsBAfbCMdbi100** : a section index run with relevance feedback combined with the Wikipedia category distance scores, using the focused local link prior $P_{foc_link}(d)$ to re-rank the top 100 articles. Finally, the CAS filter is applied to boost results with target elements in the xpath. The Best-Entry-Point is the start of the article.

Table 4: Results for the Ad Hoc Track Thorough and Focused Tasks (runs labeled "UAms" are official submissions)

| **Run id** | **MAiP** | **iP[0.00]** | **iP[0.01]** | **iP[0.05]** | **iP[0.10]** |
|---|---|---|---|---|---|
| UamsTAdbi100 | 0.2676 | 0.5350 | 0.5239 | 0.4968 | 0.4712 |
| UamsFSdocbi100CAS | 0.1726 | 0.5567 | 0.5296 | 0.4703 | 0.4235 |
| UamsFSs2dbi100CAS | 0.1928 | **0.6328** | 0.5997 | 0.5140 | 0.4647 |
| UamsRSCMACMdbi100 | 0.2096 | 0.6284 | **0.6250** | 0.5363 | 0.4733 |
| UamsRSCWACWdbi100 | 0.2132 | 0.6122 | 0.5980 | 0.5317 | 0.4782 |
| Article | 0.2814 | 0.5938 | 0.5880 | 0.5385 | 0.4981 |
| Article + Cat(Wiki) | 0.2991 | 0.6156 | 0.6150 | **0.5804** | **0.5218** |
| Article + Cat(WordNet) | 0.2841 | 0.5600 | 0.5499 | 0.5203 | 0.4950 |
| Article RF | 0.2967 | 0.6082 | 0.5948 | 0.5552 | 0.5033 |
| Article RF + Cat(Wiki) | **0.3011** | 0.6006 | 0.5932 | 0.5607 | 0.5177 |
| Article RF + Cat(WordNet) | 0.2777 | 0.5490 | 0.5421 | 0.5167 | 0.4908 |
| $(Article + CAT(Wiki)) \cdot P_{link}(d)$ | 0.2637 | 0.5568 | 0.5563 | 0.4934 | 0.4662 |
| $(Article + CAT(WordNet)) \cdot P_{link}(d)$ | 0.2573 | 0.5345 | 0.5302 | 0.4924 | 0.4567 |
| Section | 0.1403 | 0.5525 | 0.4948 | 0.4155 | 0.3594 |
| Section $\cdot P_{link}(d)$ | 0.1727 | 0.6115 | 0.5445 | 0.4824 | 0.4155 |
| Section $\cdot P_{foc\_link}(d)$ | 0.1738 | 0.5920 | 0.5379 | 0.4881 | 0.4175 |
| Section + Cat(Wiki) | 0.1760 | 0.6147 | 0.5667 | 0.5012 | 0.4334 |
| Section + Cat(WordNet) | 0.1533 | 0.5474 | 0.4982 | 0.4506 | 0.3831 |
| Section + Cat(Wiki) $\cdot P_{art\_link}(d)$ | 0.1912 | 0.6216 | 0.5808 | 0.5220 | 0.4615 |
| Section + Cat(Wiki) $\cdot P_{foc\_link}(d)$ | 0.1928 | **0.6328** | 0.5997 | 0.5140 | 0.4647 |
| Section RF | 0.1493 | 0.5761 | 0.5092 | 0.4296 | 0.3623 |
| Section RF + Cat(Wiki) | 0.1813 | 0.5819 | 0.5415 | 0.4752 | 0.4186 |
| Section RF + Cat(WordNet) | 0.1533 | 0.5356 | 0.4794 | 0.4201 | 0.3737 |
| Section RF $\cdot P_{art\_link}(d)$ | 0.1711 | 0.5678 | 0.5327 | 0.4774 | 0.4174 |

### 3.3 Thorough Task

Results of the Thorough Task can be found in Table 4. The official measure is MAiP. For the Thorough Task, the article runs are vastly superior to the section level runs. The MAiP score for the baseline Article run is more than twice as high as for the Section run. Although the Section run can be more easily improved by category and link information, even the best Section run comes nowhere near the Article baseline. The official article run UamsTAdbi100 is not as good as the baseline. This seems a score combination problem. Even with log degrees as priors, the link priors have a too large impact on the overall score. The underlying run is already a combination of the expanded query and the category scores. Link evidence might correlate with either of the two or both and lead to over use of the same information. Standard relevance feedback improves upon the baseline. The Wikipedia category distances are even more effective. The WordNet category distances are somewhat less effective, but still lead to improvement for MAiP. Combining relevance feedback with the WordNet categories hurts performance, whereas combining feedback with the Wikipedia categories improves MAiP. The link prior has a negative impact on performance of article level runs. The official run *UamsTAdbi100* is based on the *Article RF* run, but with the top 100 articles re-ranked using the local link prior. With the link evidence added, MAiP goes down considerably.

On the section runs we see again that relevance feedback and link and category information can improve performance. The Wikipedia categories are more effective than the WordNet categories and than the link degrees. The link priors also lead to improvement. On both the *Section* and *Section + Cat(Wiki)* runs, the focused link degrees are slightly more effective than the article level link degrees. For the section results, link and category evidence are complementary to each other.

For the Thorough Task, there seems to be no need to use focused retrieval techniques. Article retrieval is more effective than focused retrieval. Inter-document structures such as link and category structures are more effective.

### 3.4 Focused Task

We have no overlapping elements in our indexes, so no overlap filtering is done. Because the Thorough and Focused Tasks use the same measure, the Focused results are also shown in Table 4. However, for the Focused Task, the official measure is iP[0.01]. Even for the Focused Task, the article runs are very competitive, with the *Article + Cat(Wiki)* run outperforming all section runs. Part of the explanation is that the first 1 percent of relevant text is often found in the first relevant article. In other words, the iP[0.01 score of the article runs is based on the first relevant article in the ranking, while for the section runs, multiple relevant sections are sometimes needed to cover the first percent of relevant text. As the article run has a very good document ranking, it also has a very good precision at 1 percent recall.

The Wikipedia categories are very effective in improving performance of both the article and section index runs. They are more effective when used without relevance feedback. The link priors have a negative impact on the *Article + Cat(Wiki)* run. Again, this might be explained by the fact that the article run already has a very good document ranking and the category and link information are possibly correlated leading to a decrease in performance if we use both. However, on the *Section + Cat(Wiki)* run the link priors have a very positive effect. For comparison, we also show the official Relevant in Context run *UamsRSC-MACMdbi100*, which uses the same result elements as the *Section + Cat(Wiki)* run, but groups them per article and uses the $(Article + Cat(Wiki)) \cdot P_{link}(d)$ run for the article ranking. This improves the precision at iP[0.01]. The combination of the section run and the article run gives the best performance. This is in line with the findings in [4]. The article level index is better for ranking the first relevant document highly, while the section level index is better for locating the relevant text with the first relevant article.

In sum, for the Focused Task, our focused retrieval approach fails to improve upon standard article retrieval. Only in combination with a document ranking based on the article index does focused retrieval lead to improved performance. The whole article seems be the right level of granularity for focused retrieval

Table 5: Results for the Ad Hoc Track Relevant in Context Task (runs labeled "UAms" are official submissions)

| **Run id** | **MAgP** | **gP[5]** | **gP[10]** | **gP[25]** | **gP[50]** |
|---|---|---|---|---|---|
| UamsRSCMACMdbi100 | 0.1771 | 0.3192 | 0.2794 | 0.2073 | 0.1658 |
| UamsRSCWACWdbi100 | 0.1678 | 0.3010 | 0.2537 | 0.2009 | 0.1591 |
| Article | 0.1775 | 0.3150 | 0.2773 | 0.2109 | 0.1621 |
| Article RF | 0.1880 | 0.3498 | 0.2956 | 0.2230 | 0.1666 |
| Article + Cat(Wiki) | 0.1888 | 0.3393 | 0.2869 | **0.2271** | 0.1724 |
| Article + Cat(WordNet) | 0.1799 | 0.2984 | 0.2702 | 0.2199 | 0.1680 |
| Article RF + Cat(Wiki) | **0.1950** | **0.3528** | **0.2979** | 0.2257 | **0.1730** |
| Article RF + Cat(WordNet) | 0.1792 | 0.3200 | 0.2702 | 0.2180 | 0.1638 |
| Section | 0.1288 | 0.2650 | 0.2344 | 0.1770 | 0.1413 |
| Section $\cdot P_{art\_link}(d)$ | 0.1386 | 0.2834 | 0.2504 | 0.1844 | 0.1435 |
| Section $\cdot P_{foc\_link}(d)$ | 0.1408 | 0.2970 | 0.2494 | 0.1823 | 0.1434 |
| Section + Cat(Wiki) | 0.1454 | 0.2717 | 0.2497 | 0.1849 | 0.1407 |
| Section + Cat(Wiki) $\cdot P_{art\_link}(d)$ | 0.1443 | 0.2973 | 0.2293 | 0.1668 | 0.1392 |
| Section + Cat(Wiki) $\cdot P_{foc\_link}(d)$ | 0.1451 | 0.2941 | 0.2305 | 0.1680 | 0.1409 |

with this set of Ad Hoc topics. Again, inter-document structure is more effective than the internal document structure.

### 3.5 Relevant in Context Task

For the Relevant in Context Task, we group result per article. Table 5 shows the results for the Relevant in Context Task. A simple article level run is just as effective for the Relevant in Context task as the much more complex official runs *UamsRSCMACMdbi100* and *UamsRSCWACWdbi100*, which use the *Article +* $Cat(Wiki) \cdot log(P_{link}(d))$ run for the article ranking, and the *Section + Cat(Wiki)* and *Section + Cat(WordNet)* respectively run for the top 1500 sections.

Both relevance feedback and category distance improve upon the baseline article run. The high precision of the *Article RF* run shows that expanding the query with good terms from the top documents can help reducing the amount of non-relevant text in the top ranks and works thus as a precision device. Combining relevance feedback with the Wikipedia category distance gives the best results. The WordNet categories again hurt performance of the relevance feedback run.

For the *Section* run, the focused link degrees are more effective than the article level link degrees. The Wikipedia categories are slightly more effective than the link priors for MAgP, while the link priors lead to a higher early precision. The combination of link and category evidence is less effective than either individually.

Again, the whole article is a good level of granularity for this task and the 2009 topics. Category information is very useful to locate articles focused on the search topic.

Table 6: Results for the Ad Hoc Track Best in Context Task (runs labeled "UAms" are official submissions)

| **Run id** | **MAgP** | **gP[5]** | **gP[10]** | **gP[25]** | **gP[50]** |
|---|---|---|---|---|---|
| UamsBAfbCMdbi100 | 0.1543 | 0.2604 | 0.2298 | 0.1676 | 0.1478 |
| UamsBSfbCMs2dbi100CASart1 | 0.1175 | 0.2193 | 0.1838 | 0.1492 | 0.1278 |
| UamsTAdbi100 | 0.1601 | 0.2946 | 0.2374 | 0.1817 | 0.1444 |
| Article | 0.1620 | 0.2853 | 0.2550 | 0.1913 | 0.1515 |
| Article RF | 0.1685 | **0.3203** | **0.2645** | 0.2004 | 0.1506 |
| Article + Cat(Wiki) | 0.1740 | 0.2994 | 0.2537 | **0.2069** | **0.1601** |
| Article + Cat(WordNet) | 0.1670 | 0.2713 | 0.2438 | 0.2020 | 0.1592 |
| Article RF + Cat(Wiki) | **0.1753** | 0.3091 | 0.2625 | 0.2001 | 0.1564 |
| Article RF + Cat(WordNet) | 0.1646 | 0.2857 | 0.2506 | 0.1995 | 0.1542 |

### 3.6 Best in Context Task

The aim of the Best in Context task is to return a single result per article, which gives best access to the relevant elements. Table 6 shows the results for the Best in Context Task. We We see the same patterns as for the previous Tasks. Relevance feedback helps, so do Wikipedia and WordNet categories. Wikipedia categories are more effective than relevance feedback, WordNet categories are less effective. Wikipedia categories combined with relevance feedback gives further improvements, WordNet combined with feedback gives worse performance than feedback alone. Links hurt performance. Finally, the section index is much less effective than the article index.

The official runs fail to improve upon a simple article run. In the case of *UamsBAfbCMdbi100*, the combination of category and link information hurts the *Article RF* baseline, and in the case of *UamsBSfbCMs2dbi100CASart1*, the underlying relevance ranking of the *Section RF + Cat(Wiki)* run is simply much worse than that the *Article* run.

In summary, we have seen that relevance feedback and the Wikipedia category information can both be used effectively to improve focused retrieval. The WordNet categories can lead to improvements in some cases, but are less effective than Wikipedia categories. This is probably caused by the fact that the WordNet categories are much larger and thus have less discriminative power.

Although the difference is small, focused link evidence based on element level link degrees is slightly more effective than article level degrees. Link information is very effective for improving the section index results, but hurts the article level results when used in combination with category evidence. This might be a problem of combining the score incorrectly and requires further analysis. We leave this for future work.

With this year's new Wikipedia collection, we see again that document retrieval is a competitive alternative to element retrieval techniques for focused retrieval performance. The combination of article retrieval and element retrieval can only marginally improve performance upon article retrieval in isolation. This suggests that, for the Ad Hoc topics created at INEX, the whole article is a good level of granularity and that there is little need for sub-document retrieval techniques. Structural information such as link and category evidence also remain effective in the new collection.

## 4 Book Track

In the INEX 2009 Book Track we participated in the Book Retrieval and Focused Book Search tasks. Continuing our efforts of last year, we aim to find the appropriate level of granularity for Focused Book Search. During last year's assessment phase, we noticed that it is often hard to assess the relevance of an individual page without looking at the surrounding pages. If humans find it hard to assess individual pages, than it is probably hard for IR systems as well. In the assessments of last year, it turned out that relevant passages often cover multiple pages [9]. With larger relevant passages, query terms might be spread over multiple pages, making it hard for a page level retrieval model to assess the relevance of individual pages.

Therefore, we wanted to know if we can better locate relevant passages by considering larger book parts as retrievable units. Using larger portions of text might lead to better estimates of their relevance. However, the BookML markup only has XML elements on the page level. One simple option is to divide the whole book in sequences of $n$ pages. Another approach would be to use the logical structure of a book to determine the retrievable units. The INEX Book corpus has no explicit XML elements for the various logical units of the books, so as a first approach we divide each book in sequences of pages. We created indexes using 3 three levels of granularity:

**Book index** : each whole book is indexed as a retrievable unit.
**Page index** : each individual page is indexed as a retrievable unit.
**5-Page index** : each sequence of 5 pages is indexed as a retrievable unit. That is, pages 1–5, 6–10, etc., are treated as individual text units.

We submitted six runs in total: two for the Book Retrieval (BR) task and four for the Focused Book Search (FBS) task. The 2009 topics consist of an overall topic statement and one or multiple sub-topics. In total, there are 16 topics and 37 sub-topics. The BR runs are based on the 16 overall topics. The FBS runs are based on the 37 sub-topics.

**Book** : a standard Book index run. Up to 1000 results are returned per topic.
**Book RF** : a Book index run with Relevance Feedback (RF). The initial queries are expanded with 50 terms from the top 10 results.
**Page** : a standard Page index run.
**Page RF** : a Page index run with Relevance Feedback (RF). The initial queries are expanded with 50 terms from the top 10 results.
**5-page** : a standard 5-Page index run.
**5-Page RF** : a 5-Page index run with Relevance Feedback (RF). The initial queries are expanded with 50 terms from the top 10 results.

Table 7: The impact of feedback on the number of results per topic

| Run | pages | books | pages/book |
|---|---|---|---|
| Page | 5000 | 2029 | 2.46 |
| Page RF | 5000 | 1602 | 3.12 |
| 5Page | 24929 | 2158 | 11.55 |
| 5Page RF | 24961 | 1630 | 15.31 |
| Book | – | 1000 | – |
| Book RF | – | 1000 | – |

Table 8: Results of the INEX 2009 Book Retrieval Task

| Run id | MAP | MRR | P10 | Bpref | Rel. | Rel. Ret. |
|---|---|---|---|---|---|---|
| Book | 0.3640 | 0.8120 | 0.5071 | 0.6039 | 494 | 377 |
| Book RF | 0.3731 | 0.8507 | 0.4643 | 0.6123 | 494 | 384 |

*The impact of feedback* In Table 7 we see the impact of relevance feedback on the number of retrieved pages per topic and per book. Because we set a limit of 5,000 on the number of returned results, the total number of retrieved pages does not change, but the number of books from which pages are returned goes down. Relevance feedback using the top 10 pages (or top 10 5-page blocks) leads to more results from a single book. This is unsurprising. With expansion terms drawn from the vocabulary of a few books, we find pages with similar terminology mostly in the same books. On the book level, this impact is different. Because we already retrieve whole books, feedback can only changes the set of book returned. The impact on the page level also indicates that feedback does what it is supposed to do, namely, find more results similar to the top ranked results.

At the time of writing, there are only relevance assessments at the book level, and only for the whole topics. The assessment phase is still underway, so we show results based on the relevance judgements as off 15 March 2010 in Table 8. The *Book* run has an MRR of 0.8120, which means that for most of the topics, the first ranked result is relevant. This suggests that using full text retrieval on long documents like books is an effective method for locating relevance. The impact of relevance feedback is small but positive for MRR and MAP, but negative for P@10. It also helps finding a few more relevant books.

We will evaluate the page level runs once page-level and aspect-level judgements are available.

## 5 Entity Ranking

In this section, we describe our approach to the Entity Ranking Track. Our goals for participation in the entity ranking track are to refine last year's entity ranking method, which proved to be quite effective, and to explore the opportunities of the new Wikipedia collection. The most effective part of our entity ranking approach last year was combining the documents score with a category score, where the category score represents the distance between the document categories and the target categories. We do not use any link information, since last year this only lead to minor improvements [7].

### 5.1 Category information

For each target category we estimate the distances to the categories assigned to the answer entity, similar to what is done in Vercoustre et al. [16]. The distance between two categories is estimated according to the category titles. Last year we also experimented with a binary distance, and a distance between category contents, but we found the distance estimated using category titles the most efficient and at the same time effective method.

To estimate title distance, we need to calculate the probability of a term occurring in a category title. To avoid a division by zero, we smooth the probabilities of a term occurring in a category title with the background collection:

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

where $C$ is the category title and $D$ is the entire wikipedia document collection, which is used to estimate background probabilities. We estimate $P(t|C)$ with a parsimonious model [3] that uses an iterative EM algorithm as follows:

$$\begin{aligned} &\text{E-step:} && e_t = tf_{t,C} \cdot \frac{\alpha P(t|C)}{\alpha P(t|C) + (1-\alpha)P(t|D)} \\ &\text{M-step:} && P(t|C) = \frac{e_t}{\sum_t e_t}, \text{i.e. normalize the model} \end{aligned}$$

The initial probability $P(t|C)$ is estimated using maximum likelihood estimation. We use KL-divergence to calculate distances, and calculate a category score that is high when the distance is small as follows:

$$S_{cat}(C_d|C_t) = -D_{KL}(C_d|C_t) = -\sum\nolimits_{t \in D} \left( P(t|C_t) * \log\left(\frac{P(t|C_t)}{P(t|C_d)}\right)\right)$$

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 score for an answer entity in relation to a target category $S(d|C_t)$ is the highest score, or shortest distance from any of the document categories to the target category.

For each target category we take only the shortest distance from any answer entity category to a target category. So if one of the categories of the document is exactly the target category, the distance and also the category score for that target category is 0, no matter what other categories are assigned to the document. Finally, the score for an answer entity in relation to a query topic $S(d|QT)$ is the sum of the scores of all target categories:

$$S_{cat}(d|QT) = \sum\nolimits_{C_t \in QT} \operatorname*{argmax}_{C_d \in d} S(C_d|C_t)$$

Besides the category score, we also need a query score for each document. This score is calculated using a language model with Jelinek-Mercer smoothing without length prior:

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

Finally, we combine our query score and the category score through a linear combination. For our official runs both scores are calculated in the log space, and then a weighted addition is made.

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

We made some additional runs using a combination of normalised scores. In this case, scores are normalised using a min-max normalisation:

$$S_{norm} = \frac{S - Min(S_n)}{Max(S_n) - Min(S_n)}$$

A new feature in the new Wikipedia collection is the assignment of YAGO/ WordNet categories to documents as described in Section 2.2. These WordNet categories have some interesting properties for entity ranking. The WordNet categories are designed to be conceptual, and by exploiting list information, pages should be more consistently annotated. In our official runs we have made several combinations of Wikipedia and WordNet categories.

### 5.2 Pseudo-Relevant Target Categories

Last year we found a discrepancy between the target categories assigned manually to the topics, and the categories assigned to the answer entities. The target categories are often more general, and can be found higher in the Wikipedia category hierarchy. For example, topic 102 with title 'Existential films and novels' has as target categories 'films' and 'novels,' but none of the example entities belong directly to one of these categories. Instead, they belong to lower level categories such as '1938 novels,' 'Philosophical novels,' 'Novels by Jean-Paul Sartre' and 'Existentialist works' for the example entity 'Nausea (Book).' The term 'novels' does not always occur in the relevant document category titles, so for those categories the category distance will be overestimated. In addition to the manually assigned target categories, we have therefore created a set of pseudo-relevant target categories. From our baseline run we take the top $n$ results, and assign $k$ pseudo-relevant target categories if they occur at least 2 times as a document category in the top $n$ results. Since we had no training data available we did a manual inspection of the results to determine the parameter settings, which are $n = 20$ and $k = 2$ in our official runs. For the entity ranking task we submitted different combinations of the baseline document score, the category score based on the assigned target categories, and the category score based on the

Table 9: Target Categories

| **Topic** | **olympic classes dinghie sailing** | **Neil Gaiman novels** | **chess world champions** |
|---|---|---|---|
| **Assigned** | dinghies | novels | chess grandmasters<br>world chess champions |
| **PR** | dinghies<br>sailing | comics by Neil Gaiman<br>fantasy novels | chess grandmasters<br>world chess champions |
| **Wikipedia** | dinghies<br>sailing at the olympics<br>boat types | fantasy novels<br>novels by Neil Gaiman | chess grandmasters<br>chess writers<br>living people<br>world chess champion<br>russian writers<br>russian chess players<br>russian chess writers<br>1975 births<br>soviet chess players<br>people from Saint Petersburg |
| **Wordnet** | specification<br>types | writing<br>literary composition<br>novel<br>written communication<br>fiction | entity<br>player<br>champion<br>grandmaster<br>writer<br>chess player<br>person<br>soviet writers |

pseudo-relevant target categories. For the list completion task, we follow a similar procedure to assign target categories, but instead of using pseudo-relevant results, we use the categories of the example entities. All categories that occur at least twice in the example entities are assigned as target categories.

### 5.3 Results

Before we look at at the results, we take a look at the categories assigned by the different methods. In Table 9 we show a few example topics together with the categories as assigned ("Assigned") by each method. As expected the pseudo-relevant target categories ("PR") are more specific than the manually assigned target categories. The number of common Wikipedia categories in the example entities ("Wikipedia") can in fact be quite large. More categories is in itself not a problem, but also non relevant categories such as '1975 births' and 'russian writers' and very general categories such as 'living people' are added as target categories. Finally, the WordNet categories ("WordNet") contain less detail than the Wikipedia categories. Some general concepts such as 'entity' are included. With these kind of categories, a higher recall but smaller precision is expected.

The official results of the entity ranking runs can be found in Table 10. The run that uses the official categories assigned during topic creation performs best, and significantly better than the baseline when we consider Average Precision (xinfAP). The pseudo-relevant categories perform a bit worse, but still significantly better than the baseline. Combining the officially assigned categories and the pseudo-relevant categories does not lead to any additional improvements. Looking at the NDCG measure the results are unpredictable, and do not correlate well to the AP measure. In addition to the official runs, we created some additional runs using min-max normalisation before combining scores. For each combinations, only the best run is given here with the corresponding $\lambda$.

In our official list completion runs we forgot to remove the example entities from our result list. The results reported in Table 11 are therefore slightly better than the official results. For all runs we use $\lambda = 0.9$. We see that the run based on the WordNet categories outperforms the runs using the Wikipedia categories, although the differences are small. Again the AP results, do not correspond well to the NDCG measure.

Table 10: Results Entity Ranking

| Run | AP | NDCG |
|---|---:|---:|
| Base | 0.171 | 0.441 |
| Off. cats ($\lambda = 0.9$) | $0.201^{\bullet}$ | $0.456^{\circ}$ |
| Off. cats norm. ($\lambda = 0.8$) | $\mathbf{0.234}^{\bullet}$ | $\mathbf{0.501}^{\bullet}$ |
| Prf cats ($\lambda = 0.9$) | $0.190^{\circ}$ | $0.421^{\circ}$ |
| Off. cats ($\lambda = 0.45$) + Prf cats ($\lambda = 0.45$) | $0.199^{\bullet}$ | $0.447^{-}$ |

Table 11: Results List Completion

| Run | AP | NDCG |
|---|---:|---:|
| Base | 0.152 | 0.409 |
| Wiki ex. cats | $0.163^{\bullet}$ | $0.402^{-}$ |
| Wiki ex. + prf cats | $0.168^{\textcolor{gray}{\bullet}}$ | $0.397^{\circ}$ |
| WordNet ex. cats | $\mathbf{0.181}^{\textcolor{gray}{\bullet}}$ | $\mathbf{0.418}^{-}$ |
| Wiki + Wordnet ex. cats | $0.173^{\bullet}$ | $0.411^{-}$ |

Compared to previous years the improvements from using category information are much smaller. In order to gain some information on category distributions within the retrieval results, we analyse the relevance assessment sets of the current and previous years. We show some statistics in Table 12.

When we look at the Wikipedia categories, the most striking difference with the previous years is the percentage of pages belonging to the target category. In the new assessments less pages belong to the target category. This might be caused by the extension of the category structure. In the new collection there are more categories, and the categories assigned to the pages are more refined than before. Also less pages belong to the majority category of the relevant pages, another sign that the categories assigned to pages have become more diverse. When we compare the WordNet to the Wikipedia categories, we notice that the WordNet categories are more focused, i.e. more pages belong to the same

Table 12: Relevance assessment sets statistics

| Year | 07 | 08 | 09 | 09 |
|---|---|---|---|---|
| Cats | Wiki | Wiki | Wiki | WordNet |
| Avg. # of pages | 301 | 394 | 314 | 314 |
| Avg. % relevant pages | 0.21 | 0.07 | 0.20 | 0.20 |
| Pages with majority category of all pages: | | | | |
| all pages | 0.232 | 0.252 | 0.254 | 0.442 |
| relevant pages | 0.364 | 0.363 | 0.344 | 0.515 |
| non-relevant pages | 0.160 | 0.241 | 0.225 | 0.421 |
| Pages with majority category of relevant pages: | | | | |
| all pages | 0.174 | 0.189 | 0.191 | 0.376 |
| relevant pages | 0.608 | 0.668 | 0.489 | 0.624 |
| non-relevant pages | 0.068 | 0.155 | 0.122 | 0.317 |
| Pages with target category: | | | | |
| all pages | 0.138 | 0.208 | 0.077 | |
| relevant pages | 0.327 | 0.484 | 0.139 | |
| non-relevant pages | 0.082 | 0.187 | 0.064 | |

categories. This is in concordance with the previously calculated numbers of the distribution of articles over Wikipedia and WordNet categories, and vice versa in Section 2.2.

We are still investigating if there are other reasons that explain why the performance does not compare well to the performance in previous years. Also we expect some additional improvements from optimising the normalisation and combination of scores.

## 6 Conclusion

In this paper we discussed our participation in the INEX 2009 Ad Hoc, Book, and the Entity Ranking Tracks.

For the Ad Hoc Track we conclude focused link evidence outperforms local link evidence on the article level for the Focused Task. Focused link evidence leads to high early precision. Using category information in the form of Wikipedia categories turns out to be very effective, and more valuable than WordNet category information. These inter-document structures are more effective than document internal structure. Our focused retrieval approach can only marginally improve an article retrieval baseline and only when we keep the document ranking of the article run. For the INEX 2009 Ad Hoc topics, the whole article level seems a good level of granularity.

For the Book Track, using the full text of books gives high early precision and even good overall precision, although the small number of judgements might lead to an over-estimated average precision. Relevance feedback seems to be very effective for further improving early precision, although it can also help finding more relevant books. The Focused Book Search Task still awaits evaluation because there are no page-level relevance judgements yet.

Considering the entity ranking task we can conclude that in the new collection using category information still leads to significant improvements, but that the improvements are smaller because the category structure is larger and categories assigned to pages are more diverse. WordNet categories seem to be a good alternative to the Wikipedia categories. The WordNet categories are more general and consistent categories.

This brings up an interesting difference between ad hoc retrieval and entity ranking. We use the same category distance scoring function for both tasks, but for the former, the highly specific and noisy Wikipedia categories are more effective, while for the latter the more general and consistent WordNet categories are more effective. Why does ad hoc search benefit more from the more specific Wikipedia categories? And why does entity ranking benefit more from the more general WordNet categories? Does the category distance in the larger Wikipedia category structure hold more focus on the topic and less on the entity type? And vice versa, are the more general categories of the WordNet category structure better for finding similar entities but worse for keeping focus on the topical aspect of the search query? These questions open up an interesting avenue for future research.

*Acknowledgments* Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.-501). Rianne Kaptein was supported by NWO under grant # 612.066.513 Marijn Koolen was supported by NWO under grant # 640.001.501.

## Bibliography

[1] K. N. Fachry, J. Kamps, M. Koolen, and J. Zhang. Using and detecting links in Wikipedia. In *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *LNCS*, pages 388–403. Springer Verlag, Heidelberg, 2008.

[2] K. N. Fachry, J. Kamps, M. Koolen, and J. Zhang. Using and detecting links in Wikipedia. In N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors, *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *Lecture Notes in Computer Science*, pages 388–403. Springer Verlag, Heidelberg, 2008.

[3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM Press, New York NY, 2004.

[4] J. Kamps and M. Koolen. The impact of document level ranking on focused retrieval. In *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5631 of *LNCS*. Springer Verlag, Berlin, Heidelberg, 2009.

[5] J. Kamps and M. Koolen. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*. ACM Press, New York NY, USA, 2009.

[6] J. Kamps, M. Koolen, and B. Sigurbjörnsson. Filtering and clustering XML retrieval results. In *Comparative Evaluation of XML Information Retrieval Systems: Fifth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2006)*, volume 4518 of *LNCS*, pages 121–136. Springer Verlag, Heidelberg, 2007.

[7] R. Kaptein and J. Kamps. Finding entities in Wikipedia using links and categories. In *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5631 of *LNCS*. Springer Verlag, Berlin, Heidelberg, 2009.

[8] 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.

[9] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 452–459, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1572019.

[10] M. Koolen and J. Kamps. What's in a link? from document importance to topical relevance. In *Proceedings of the 2nd International Conferences on the Theory of Information Retrieval (ICTIR 2009)*, volume 5766 of *LNCS*, pages 313–321. Springer Verlag, Berlin, Heidelberg, 2009.

[11] R. Schenkel, F. Suchanek, and G. Kasneci. YAWN: A semantically annotated wikipedia xml corpus. In *12th GI Conference on Databases in Business, Technology and Web (BTW 2007)*, March 2007.

[12] B. Sigurbjörnsson and J. Kamps. The effect of structured queries and selective indexing on XML retrieval. In *Advances in XML Information Retrieval and Evaluation: INEX 2005*, volume 3977 of *LNCS*, pages 104–118, 2006.

[13] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An Element-Based Approach to XML Retrieval. In *INEX 2003 Workshop Proceedings*, pages 19–26, 2004.

[14] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Mixture models, overlap, and structural hints in XML element retreival. In *Advances in XML Information Retrieval: INEX 2004*, volume 3493 of *LNCS 3493*, pages 196–210, 2005.

[15] 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.

[16] 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.