Combining concepts and language models for information access

Meij, E.J.

Citation for published version (APA):
Meij, E. J. (2010). Combining concepts and language models for information access

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
There are very few things that are purely conceptual without any hard content.

Kevin Bacon

Query Modeling Using Linked Concepts

In previous chapters we have seen various ways of updating the estimate of a query model, for example through the use of feedback information (Chapter 4) or conceptual document annotations (Chapter 5). In essence, these approaches are a form of data fusion, where information from multiple sources is combined to influence a document ranking. Such fusion methods exist in a number of related tasks. For example, in web retrieval it is common to take into account anchor texts or some function of the web graph [45]. In multimedia environments, different modalities (text, video, speech, etc.) need to be combined. In cross-lingual IR, where the queries and documents are stated in different languages, evidence from multiple languages is merged to obtain a final ranking. In our query modeling case, we have combined evidence from either top-ranked or relevant documents and the initial query. In Chapter 5 we have added to this concepts in the form of document annotations. In Chapter 6 we have linked domain-restricted queries to DBpedia and the question arose “Can we apply the semantic analysis based on Linked Open Data (LOD) to the open domain?” Furthermore, can we apply these linked concepts for retrieval, using the ideas presented in Chapter 4 and Chapter 5?

Looking from a different angle, there have been several developments in web search over the last 20 years [19, 300]. Initially, web pages were ranked solely based on term frequency (TF) and inverse document frequency (IDF) of the terms a user entered in her query. Later, this was enriched with “off-page” information, such as information from the web graph, anchortext, and related hyperlinks and from user behavior such as clicks or dwell time [45, 152]. Most recently, as users are visiting the search engines for more diverse reasons [300], the major web search engines are also moving towards semantically informed responses, aiming to interpret a user’s intent and answer the information need behind the query [19]—whether the search engines “follow” changing user behavior or whether users adapt to new functionalities offered by search engines does not really matter for this discussion [143]. Aiming to answer information needs instead of queries involves rather low-level enhancements such as spelling correction, but also more fine-grained user interface enhancements such
7. Query Modeling Using Linked Concepts

as query suggestions [9]. A prime example is the Yahoo! query formulation tool called searchassist that we have mentioned as an example in the previous chapter. In [217] we have shown that blending in conceptual information in the query suggestion process can improve such suggestions, especially for rare, infrequent queries. Moving more towards determining the meaning of queries (or, indeed, the information needs behind them), current enhancements include determining the task the user aims to solve [46, 270] or determining the type of information that is being sought (through so-called verticals—which are typically defined as “domain-specific subcollections”) [11, 93]. Another way of attempting to answer the information need behind the query is through semantic analysis, for example by (semi-automatic) expansion of the query using synonyms [113]. Other approaches aim to infer the semantics behind the queries that are submitted [42].

Even other approaches try to understand the “things” that are being sought. For example, using the approach presented by Gabrilovich and Markovitch [107], we can obtain a mapping of free text to concepts (in the form of Wikipedia articles); the same ideas are applied in a more general sense by Turney and Pantel [322]. Medelyan et al. [205] present a comprehensive overview of approaches making use of Wikipedia to extract and make use of the concepts, relations, facts, and descriptions found in Wikipedia.

One of the current goals of the semantic web (in particular the LOD cloud or “web of data”) is to expose, share, and connect data [32, 37]. For this, it uses URIs to identify concepts and provides means by which to describe the concepts themselves as well as any possible relationships with other concepts. One of the current goals of major search engines is very similar: to move beyond a web of pages towards gathering and exposing web-derived knowledge and a “web of things” instead [19]. Indeed, in this chapter we explore what happens when we apply the semantic analysis method from Chapter 6, that links queries to a semantic “backbone,” (in the form of concepts in a concept language). We do so in order to “understand” open domain queries and to estimate query models based on this conceptual information.

In particular, we take the best performing machine learning method from the previous chapter and map queries from the open domain to DBpedia concepts. Then, we apply the most robust relevance feedback method, relevance model 1 (RM-1), from Chapter 4 to the Wikipedia articles associated with the found DBpedia concepts to estimate a query model. The guiding intuition is that, similar to our conceptual query models, concepts are best described by the language use associated with them. In other words, once our algorithm has determined which concepts are meant by a query, we employ the language use associated with those concepts to update the query model. We compare the performance of this approach to pseudo relevance feedback on the collection (in the same way as presented in Chapter 4) and to pseudo relevance feedback on Wikipedia (similar to the way we obtain conceptual query models in Chapter 5).
The research questions we address in this chapter are as follows.

**RQ 4.** What are the effects on retrieval performance of applying pseudo relevance feedback methods to texts associated with concepts that are automatically mapped from ad hoc queries?

a. What are the differences with respect to pseudo relevance estimations on the collection? And when the query models are estimated using pseudo relevance estimations on the concepts' texts?

b. Is the approach mainly a recall- or precision-enhancing device? Or does it help other aspects, such as promoting diversity?

The main contribution presented in this chapter is to provide an indication to what extent the LOD-based semantic analysis presented in the previous chapter can be applied for query modeling in the open domain. In this chapter, we therefore make use of the TREC Terabyte 2004–2006 (TREC-TB) and TREC Web 2009, Category A (TREC-WEB-09) test collections as introduced in Section 3.3. Recall that TREC Terabyte uses the .GOV2 document collection, a large crawl of the .gov domain. TREC Web 2009 uses the ClueWeb09 document collection, a realistically sized web collection. In the experiments in this chapter we use the largest subset, Category A. The topics associated with the TREC Web 2009 test collection are taken from a search engine's log and representative of queries submitted to a web search engine.

We continue this chapter in Section 7.1 by introducing our method for obtaining DBpedia concepts from ad hoc queries. In Section 7.2 we detail how we estimate the query models as well as the experimental setup used. We discuss results in Section 7.3 and end with a concluding section.

### 7.1 Linking queries to Wikipedia

To be able to derive query models based on the concepts meant by the query, we first need to link queries to concepts (in the form of Wikipedia articles or, equivalently, DBpedia concepts). To this end, we follow the approach from Chapter 6, which maps queries to DBpedia concepts. In this case, however, we subsequently apply query modeling. We take the best performing settings from that chapter, i.e., SVM with a polynomial kernel using full queries. Instead of using the Sound and Vision dataset, however, we employ two ad hoc TREC test collections in tandem with a dump of the English version of Wikipedia (dump date 20090920).

In order to classify concepts as being relevant to a query, the approach uses manual query-to-concept annotations to train the SVM model. During testing, a retrieval run is performed on Wikipedia for new, unseen queries. The results of which are then classified using the learned model. The output of this step is a
label for each concept, indicating whether it is relevant or not. This dichotomy represents our binary classification problem.

Wikipedia and supervised machine learning have previously been used to select optimal terms to include in the query model [347]. We, however, are interested in selecting those concepts that best describe the query and use those to sample terms from. This is similar to the unsupervised manner used, e.g., in the context of retrieving blogs [337]. Such approaches are completely unsupervised in that they only consider a fixed number of pseudo relevant Wikipedia articles. As we will see below, focusing this set using machine learning improves overall retrieval performance.

The features that we use include those pertaining to the query, the Wikipedia article, and their combination. See Section 6.2.3 for an extensive description of each. Since we are using ad hoc test collections in this case, we do not have session information and omit the history-based features. In order to obtain training data, we have asked 4 annotators to manually identify all relevant Wikipedia articles for queries in the same fashion as presented in the previous chapter. The average number of Wikipedia articles the annotators identified per query is around 2 for both collections. The average number of articles identified as relevant per query by SVM is slightly different between the test collections, with 1.6 for TREC Terabyte and 2.7 for TREC Web 2009. This seems to be due to the differences in queries; the TREC Web queries are shorter and, thus, more prone to ambiguity.

Let’s look at some examples. Table 7.1 shows examples of concepts that are identified by the SVM model on the TREC Web 2009 test collection. We first observe that, as pointed out above, the queries themselves are short and ambiguous. For query (#48) “wilson antenna,” it predicts ROBERT WOODROW WILSON as the only relevant concept, classifying concepts such as MOUNT WILSON (CALIFORNIA) as not relevant. For the query “the music man” (#42) it identifies the company, song, film, and musical which indicates the inherent ambiguity that is typical for many web queries. The same effect can be observed for the query “disneyland hotel” (#39) with concepts TOKYO DISNEYLAND HOTEL, DISNEYLAND
## 7.1. Linking queries to Wikipedia

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Query</th>
<th>Concepts</th>
</tr>
</thead>
</table>
| 2       | french lick resort and casino                   | FRENCH LICK RESORT CASINO  
FRENCH LICK, INDIANA                                                      |
| 13      | map                                             | MAP  
TOPOGRAPHIC MAP  
WORLD MAP  
The National Map                                                          |
| 14      | dinosaurs                                       | DINOSAURS  
HARRY AND HIS BUCKET FULL OF DINOSAURS  
Walking with Dinosaurs                                                     |
| 15      | espn sports                                     | ESPN Star Sports  
ESPN  
ESPN on ABC                                                                  |
| 16      | arizona game and fish                           | ARIZONA GAME AND FISH DEPARTMENT  
LIST OF LAKES IN ARIZONA                                                   |
| 17      | poker tournaments                               | POKER TOURNAMENT  
ULTIMATE POKER CHALLENGE                                                   |
| 23      | yahoo                                           | YAHOO!  
YAHOO! MUSIC  
YAHOO! News                                                                |
| 24      | diversity                                       | SPECIES DIVERSITY  
GENETIC DIVERSITY  
CULTURAL DIVERSITY                                                          |
| 26      | lower heart rate                                | HEART RATE  
HEART RATE VARIABILITY  
DOPPLER FETAL MONITOR                                                       |
| 28      | inuyasha                                        | INUYASHA  
LIST OF INUYASHA EPISODES  
LIST OF INUYASHA CHARACTERS                                                |
| 39      | disneyland hotel                                | DISNEYLAND HOTEL (CALIFORNIA)  
DISNEYLAND HOTEL (PARIS)  
Tokyo DISNEYLAND HOTEL                                                     |
| 41      | orange county convention center                 | ORANGE COUNTY CONVENTION CENTER  
ORANGE COUNTY, CALIFORNIA  
LIST OF CONVENTION & EXHIBITION CENTERS                                    |
| 42      | the music man                                   | THE MUSIC MAN  
THE MUSIC MAN (1962 FILM)  
MUSIC MAN  
The Music Man (song)                                                        |
| 45      | solar panels                                    | PHOTOVOLTAIC MODULE                                                        |
| 48      | wilson antenna                                  | ROBERT WOODROW WILSON                                                      |
| 49      | flame designs                                   | FLAME OF RECCA  
GEORDIE LAMP                                                               |

**Table 7.1:** Examples of topics automatically linked to concepts on the TREC Web 2009 test collection.
Query Modeling Using Linked Concepts

HOTEL (CALIFORNIA), and DISNEYLAND HOTEL (PARIS). There are also mistakes, however, such as predicting the concepts FLAME OF RECCA and GEORDIE LAMP for the query (#49) “flame designs.” The first concept is a Japanese manga series, whereas ‘Geordie’ was the nickname of the designer of the mine lamp that served as a solution to explosions due to firedamp in coal mines.

In the next stage, we take the predicted concepts for each query and estimate query models from the Wikipedia article associated with each concept. For this, we adopt the language modeling approach detailed in Section 2.2.2 and as query model we use the linear interpolation from Eq. 2.10. Recall that there, \( P(t|\hat{\theta}_Q) \) indicates the empirical estimate on the initial query and \( P(t|\tilde{\theta}_Q) \) the expanded part. In Chapter 4, relevance model 1 (RM-1, cf. Eq. 2.24) had the most robust performance. We therefore use this model to obtain \( P(t|\hat{\theta}_Q) \) and estimate it on the contents of the Wikipedia articles associated with the concepts. In essence, this method is similar to the one we presented in Chapter 5. There, we used conceptual document annotations to (i) obtain a conceptual representation of each query and to (ii) “translate” the found concepts to vocabulary terms. In this chapter, we use the learned SVM model to obtain the first step. Since each concept is now associated with a single document (the Wikipedia article), we use those to update the estimate of the query model.

Figure 7.1 shows two example query models for topics #42 and #39 from the TREC Web 2009 test collection. We note that the initial query terms receive the largest probability mass and that the terms that are introduced seem mostly related to the topic.

7.2 Experimental Setup

To determine whether the automatically identified concepts are a useful resource to improve retrieval performance by updating the query model, we compare our approach (WP-SVM) against a query likelihood (QL) baseline and against RM-1 estimated on pseudo relevant documents. In particular, we obtain the set of pseudo relevant documents in three ways:

1. on the collection (“normal” pseudo relevance feedback—similar to the approach presented in Chapter 4),

2. on Wikipedia (similar to the approach presented in Chapter 5 as well as so-called “external expansion” [92, 337]), and

3. on automatically linked Wikipedia articles (linked using the approach from Chapter 6), as introduced in the previous section.

So, as reference, we use either the collection (RM (C)) or top-ranked Wikipedia articles (RM (WP)) for query modeling. RM (WP) is obtained using a full-text index of Wikipedia, containing all the fields introduced in the previous chapter and
including within-Wikipedia anchortexts and titles. For both RM (WP) and RM (C) we use the top 10 retrieved documents and include the 10 terms with the highest probability in \( P(t|\hat{θ}_Q) \), similar to the experimental setup used in Chapter 4 (there, on the TREC-PRF-08 collection, RM-1 obtained its highest retrieval performance when 10 terms were used).

To train the SVM model, we split the topic set of each test collection in a training and test set. For TREC Terabyte 2004–2006, we have 149 topics of which 74 are used for training and 75 for testing. For TREC Web 2009 we have 50 topics and use 5-fold cross validation [344]. Similar to the experiments presented in Chapter 4 and described in Section 3.4 (cf. page 50), we perform a line search of the parameter space to determine the optimal value for \( λ_Q \).

### 7.3 Results and Discussion

Before we report on the experimental results, we first note the performance of results reported in the literature on the test collections employed in this chapter. For the TREC Terabyte test collection, this number is not available since we (i) use an aggregation of the topic sets from all TREC Terabyte 2004–2006 tracks and (ii) split this new topic set in a training and test set. We do note, however, that the average MAP score of all systems participating in the TREC Terabyte 2004–2006 tracks is roughly 0.30. For TREC Web 2009, we cannot compare our absolute scores with those presented in the literature, since we use the mtc-eval evaluation methodology [61]. Hence, we determine the probability of relevance for each unjudged document retrieved by the runs presented in this chapter using the expert tool.\(^1\)

Table 7.2 lists the results on the TREC Terabyte test collection, optimized for MAP. Here, applying RM-1 to pseudo relevant documents from the collection yields highest MAP, although the difference with respect to the MAP values for RM (WP) and WP-SVM is very small. All models obtain significant improvements over the baseline in terms of MAP. When the query models are estimated on Wikipedia, the highest mean reciprocal rank (MRR) is obtained, with WP-SVM following closely; only WP-SVM and RM (WP) obtain significant improvements in terms of MRR and recall. WP-SVM retrieves the most relevant documents of all the models on this collection. Interestingly, it also obtains the highest early precision.

Figure 7.2 shows a per-topic plot of the performance of WP-SVM relative to the baseline (a positive value indicates an improvement over the baseline). The first thing to note is that there are a number of topics that are neither helped nor hurt. One of the properties of the conceptual mapping approach is that the SVM may decide that none of the candidate concepts are relevant. The query model is

\(^1\)See http://ir.cis.udel.edu/~carteret/downloads.html.
Table 7.2: Results on the TREC Terabyte test collection, optimized for MAP.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\lambda_Q$</th>
<th>P10</th>
<th>RelRet</th>
<th>MRR</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>0.0</td>
<td>0.439</td>
<td>0%</td>
<td>6965</td>
<td>0%</td>
</tr>
<tr>
<td>RM (C)</td>
<td>0.3</td>
<td>0.515*</td>
<td>+17%</td>
<td>7872*</td>
<td>+13%</td>
</tr>
<tr>
<td>RM (WP)</td>
<td>0.2</td>
<td>0.527*</td>
<td>+20%</td>
<td>7836*</td>
<td>+13%</td>
</tr>
<tr>
<td>WP-SVM</td>
<td>0.2</td>
<td>0.532*</td>
<td>+21%</td>
<td>7902*</td>
<td>+13%</td>
</tr>
</tbody>
</table>

Figure 7.2: Per-topic breakdown of the improvement of WP-SVM over the QL baseline on the TREC Terabyte test collection.

left as is in that case, yielding the same performance as the baseline. This is the case for 30 out of the 75 TREC Terabyte topics. We further observe that, although about as many topics are helped as hurt in terms of MAP, there are more topics that are helped more using WP-SVM on early precision. So, in those cases where concepts are identified, early precision is helped most.

Topic #847 (“Portugal World War II”) is a topic that is hurt when applying WP-SVM. Here, the two concepts that are returned (LIST OF MILITARY VEHICLES and LIST OF SHIPWRECKS IN 1943) are vaguely related but not relevant to the query. Topics that are helped using WP-SVM include “train station security measures” (#711), caused by the suggested concept SECURITY ON THE MASS RAPID TRANSIT. Another topic that is helped on this test collection is topic #733 “Airline overbooking”. Here, the concept AIRLINE is the only suggestion. For topic #849 (“Scalable Vector Graphics”), the concepts SCALABLE VECTOR GRAPHICS and VECTOR GRAPHICS are returned, causing 42 more relevant documents to be returned. These findings provide evidence that Wikipedia is a useful resource for query modeling; the approach functions as both a recall- and a precision-enhancing device.

As to TREC Web 2009, Table 7.3 shows the results on this test collection, using mtc-eval measures [61], which were introduced in Chapter 3. On this collection
7.3. Results and Discussion

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_Q$</th>
<th>eP10</th>
<th>eR-prec</th>
<th>eMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>1.0</td>
<td>0.077</td>
<td>0%</td>
<td>0.272</td>
</tr>
<tr>
<td>RM (C)</td>
<td>0.5</td>
<td>0.070</td>
<td>-9%</td>
<td>0.278</td>
</tr>
<tr>
<td>RM (WP)</td>
<td>0.5</td>
<td>0.082*</td>
<td>+6%</td>
<td>0.268*</td>
</tr>
<tr>
<td>WP-SVM</td>
<td>0.0</td>
<td>0.241*</td>
<td>+213%</td>
<td>0.348*</td>
</tr>
</tbody>
</table>

Table 7.3: Results on the TREC Web test collection (Category A), optimized for eMAP.

Figure 7.3: Per-topic breakdown of the improvement of WP-SVM over the QL baseline on the TREC Web test collection using ad hoc measures.

of web pages, we observe that merely relying on the baseline approach yields very low retrieval performance. Applying pseudo relevance feedback on the collection does not help; in fact, retrieval performance in terms of early precision is degraded in that case.

When estimating a relevance model from Wikipedia (RM (WP)), we find a slight decrease in terms of eMAP. It does yield a significant improvement in terms of eP10, however. Moreover, in this case, eR-prec is also significantly improved. WP-SVM improves the performance on all metrics. Interestingly, the best results here are obtained when $\lambda_Q = 1.0$, i.e., when all probability mass is given to the expanded query part.

Figure 7.3 again shows per-topic plots, this time for the TREC Web test collection. From these plots it is clear that WP-SVM helps to substantially improve early precision on this test collection; eMAP is also improved over almost all topics. Topics that are helped most include #46 (“alexian brothers hospital”—caused by the concepts ALEXIANS and LIST OF HOSPITALS IN ILLINOIS) and #25 (“euclid”—caused by the single matching concept EUCLID). Topic #12 (“djs”) is hurt most in terms of eP10 and is the only topic that is hurt on eMAP. Here, three DJs are identified as concepts (QUAD CITY DJ’S, PLUMP DJS, and SOULWAX) but they do not help to improve on early precision. In sum, the results presented so far in-
Table 7.4: Results on the TREC Web test collection in terms of diversity, optimized for α-NDCG@10.

<table>
<thead>
<tr>
<th>Method</th>
<th>λ_Q</th>
<th>IA-P@10</th>
<th>α-NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>1.0</td>
<td>0.017</td>
<td>0.041</td>
</tr>
<tr>
<td>RM (C)</td>
<td>0.5</td>
<td>0.013</td>
<td>-24%</td>
</tr>
<tr>
<td>RM (WP)</td>
<td>0.5</td>
<td>0.016</td>
<td>-6%</td>
</tr>
<tr>
<td>WP-SVM</td>
<td>0.6</td>
<td>0.035</td>
<td>+106%</td>
</tr>
</tbody>
</table>

dicate that supervised query modeling using Wikipedia is helpful for large, noisy collections.

The TREC Web 2009 track featured a sub-track in which the aim was to improve upon diversity in the document ranking, as introduced in Chapter 3. Recall that diversity aims to reward those document rankings in which documents that are related to subtopics of the query appear at the top. Moreover, rankings that retrieve documents relating to many subtopics are preferred to those that cover fewer subtopics. The subtopics for the TREC Web 2009 track are based on information extracted from the logs of a commercial search engine and roughly balanced in terms of popularity. When we evaluate WP-SVM on the TREC Web 2009 collection using the diversity track's measures, cf. Table 7.4, we arrive at the same picture as for ad hoc retrieval.\(^1\) Pseudo relevance feedback on the collection hurts diversity using both measures. We observe the same results, although to a lesser extent, when applying pseudo relevance feedback on Wikipedia. When we use WP-SVM, however, the diversity of the document rankings is improved, as measured by both IA-P@10 and α-NDCG@10, although not significantly so.

Figure 7.4 shows per-topic plots of the diversity measures, comparing the baseline to WP-SVM. From these plots it is clear why we do not obtain significant improvements; diversity is only helped on a small number of topics. Topic #49 (“flame designs”) is the only topic that is hurt. For this topic, the concepts **FLAME OF RECCA** and **GEORDIE LAMP** are retrieved. Both do not seem relevant to the topic, causing the decline in terms of diversity performance. In contrast, topics #1 (“obama family tree”) and #46 (“alexian brothers hospital”) are examples of topics that are helped. For the first, the concepts **FAMILY TREE**, **MICHELLE OBAMA**, and **RULES OF RUSSIA FAMILY TREE** are identified. For the second, the concepts **ALEXIANS** and **LIST OF HOSPITALS IN ILLINOIS** are identified. In both cases, each concept refers to a different aspect of the query. Hence, the estimated query models are also diverse in these aspects which in turn helps to improve di-

\(^1\)The absolute values shown in Table 7.4 are low as compared to those obtained by the participants of that particular track (the median IA-P@10 score lies around 0.054). We note, however, that the runs presented in this chapter do not incorporate any information pertaining to the graph structure associated with the web pages, nor do they explicitly incorporate diversity information [126, 167]. The method presented here may be applied in conjunction with any diversity-improving algorithm.
versity in the resulting document ranking. These findings, in conjunction with the examples provided earlier, indicate that our query modeling approach caters for multiple interpretations of the query since prominent terms from the Wikipedia article associated with each identified concept are included in the query model.

7.4 Summary and Conclusions

In this chapter we have presented a query modeling method that brings together intuitions from the preceding chapters. It proceeds by using the conceptual mapping approach from Chapter 6 to map open domain queries to DBpedia. Next, we use the natural language associated with each concept (in the form of the text of the accompanying Wikipedia article) to estimate a query model. This approach serves as a means to (i) understanding a query, by identifying concepts meant by it and (ii) leveraging the natural language associated with those concepts to improve end-to-end retrieval performance.

The research questions we have addressed in this chapter are as follows.

RQ 4. What are the effects on retrieval performance of applying pseudo relevance feedback methods to texts associated with concepts that are automatically mapped from ad hoc queries?

On a relatively small web collection, we have found small but significant improvements over a query likelihood baseline. On a much larger web corpus, we have achieved improvements on all metrics, whether precision or recall oriented, especially when relying exclusively on externally derived contributions to the query model. In some cases, the concept selection stage does not classify any concepts as being relevant to the query, which results in obtaining the same performance as the baseline. Averaged over all topics, however, the estimated query models
using the found concepts result in significantly improved retrieval performance in terms of precision.

RQ 4a. What are the differences with respect to pseudo relevance estimations on the collection? And when the query models are estimated using pseudo relevance estimations on the concepts’ texts?

On the TREC Terabyte collection, we have found improvements of our model over RM-1 estimated on pseudo relevant documents from the collection in terms of both recall and early precision. When estimated on the concepts’ texts, we have observed that RM-1 yields the highest MRR (although only slightly better than WP-SVM).

On the TREC Web 2009 test collection, we have found that our approach improves over pseudo relevance feedback on all measures. Applying pseudo relevance feedback for query modeling does not seem to help on this test collection, neither when estimated on documents from the collection, nor when estimated on Wikipedia. In the latter case, early precision is slightly (and significantly) improved over the baseline, whereas eMAP is significantly worse.

RQ 4b. Is the approach mainly a recall- or precision-enhancing device? Or does it help other aspects, such as promoting diversity?

On the TREC Terabyte test collection, we have found significant increases in terms of both recall and early precision; a finding corroborated on the TREC Web test collection. There, we have observed substantial gains in terms of both traditional metrics and diversity measures. When considering diversity, we have observed major improvements using our approach.

In sum, we have shown that employing the texts associated with automatically identified concepts for query modeling can improve end-to-end retrieval performance. This effect is most notable on a recent, realistically sized document collection of crawled web pages. Using diversity measures put forward on that test collection, we have also noted that WP-SVM is able to substantially improve the diversity of the result list.