Thesaurus-Based Feedback to Support Mixed Search and Browsing Environments

Edgar Meij and Maarten de Rijke

ISLA, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
emeij, mdr@science.uva.nl

Abstract. We propose and evaluate a query expansion mechanism that supports searching and browsing in collections of annotated documents. Based on generative language models, our feedback mechanism uses document-level annotations to bias the generation of expansion terms and to generate browsing suggestions in the form of concepts selected from a controlled vocabulary (as typically used in digital library settings). We provide a detailed formalization of our feedback mechanism and evaluate its effectiveness using the TREC 2006 Genomics track test set. As to the retrieval effectiveness, we find a 20% improvement in mean average precision over a query-likelihood baseline, whilst increasing precision at 10. When we base the parameter estimation and feedback generation of our algorithm on a large corpus, we also find an improvement over state-of-the-art relevance models. The browsing suggestions are assessed along two dimensions: relevancy and specifity. We present an account of per-topic results, which helps understand for what type of queries our feedback mechanism is particularly helpful.

1 Introduction

A query that is used to express the information need of a digital library user may fail to match the relevant words in the domain being explored. Even if the query terms do match terms and documents from the domain, they may not be used by all authors of relevant articles. Authors working in different areas may use different terms for a single concept or may even denote different concepts with the same term. Several methods exist for overcoming this vocabulary mismatch problem, many of which are based on *query expansion*. Query expansion adds terms and possibly reweighs original query terms, so as to more effectively express the original information need. Automatic approaches to query expansion have been studied extensively in information retrieval (IR). Most of these operate by using some initial set of retrieved documents to look for additional, significant terms. Much work has been dedicated to these kinds of techniques and, over time, various methods have been proposed. One class of solutions looks at the problem from a data-driven perspective, e.g., by generating expansion terms from entire documents [17], document summaries [15], or the context in which the original query terms appear [21]. Other, more knowledge-based apporaches look at "external" resources, such as ontologies, thesauri, co-occurence tables, or synonym lists [20].

In this paper, we consider query expansion in the setting of a digital library, where information access is usually a mixture of two tasks: *searching* and *browsing* [10, 13, 18]. While query expansion is typically aimed at increasing the effectiveness of the search component, we are interested in expansion techniques that can also help to improve browsing support. In particular, our goal is to achieve effective query expansion (at least as effective as state-of-the-art data-driven approaches) which, at the same time, incorporates explicit knowledge inherent in a digital library to facilitate browsing and exploring. More specifically, we aim to enhance a user's search by facilitating this kind of browsing directly and transparently from the searching process, by suggesting controlled vocabulary terms and integrating them into the retrieval model.

The research questions we address are fourfold. First, how can we use a language modeling framework to incorporate thesaurus information for generating terms to facilitate browsing? Second, can this be done in such a way that the feedback mechanism achieves state-of-the-art performance? Third, and inspired by recent work on document expansion [8, 19], what is the impact of the size of the corpus from which feedback terms are being generated? Fourth, how can we assess the quality of the thesaurus terms being proposed for browsing?

Our main contribution is the introduction of a thesaurus-biased feedback algorithm that uses generative language modeling to not only generate expansion terms to improve retrieval results, but also to propose thesaurus terms to facilitate browsing. Our algorithm, which achieves state-of-the-art performance, consists of three steps: First, we determine the controlled vocabulary terms most closely associated with a query. We then search the documents associated with these terms, in conjunction with the query, and look for additional terms to describe the query. Finally, we weigh these proposed expansion terms, again using the document-level annotations. For evaluation purposes we use the TREC 2006 Genomics track test set [9]. Specifically, we use and compare this collection and the contents of the entire PubMed database for estimation purposes.

The remainder of this paper is organized as follows. In Section 2 we describe the background of our work, as well as our proposed query expansion algorithm. In Section 3 we detail our experimental setup, and in Section 4 we present our experimental results. Related work is discussed in Section 5 and Section 6 contains our conclusions.

2 Thesaurus-biased Query Models

Within the field of IR, language modeling is a relatively novel framework. It originates from speech recognition, where the modeling of speech utterances is mapped onto textual representations. The ideas behind it are intuitive and theoretically well-motivated, thus making it an attractive framework of choice. It provides us with an easily extendible setting for incorporating the information captured in document annotations. Before introducing our novel feedback mechanism we recall some general facts about language models for IR.

2.1 Generative Language Modeling

Language modeling for IR is centered around the assumption that a query, as issued by a user, is a sample generated from some underlying term distribution. The documents

in the collection are modeled in a similar fashion, and also regarded as samples from an unseen term distribution—a generative language model.

At retrieval time, the language usage in documents is compared with that of the query and the documents are ranked according to the likelihood of generating the query. Assuming independence between query terms, the probability of a document given a query can be more formally stated using Bayes' rule:

$$P(Q|\theta_d) \propto P(d) \cdot \prod_{q \in Q} P(q|\theta_d), \tag{1}$$

where θ_d is a language model of document d, and q the individual query terms in query Q. The term $P(d)$ captures the prior belief in a document being relevant, which is usually assumed to be uniform. $P(\cdot|\theta_d)$ is estimated using maximum-likelihood estimates which, in this case, means using the frequency of a query term in a document: $P(q|\theta_d) = c(q,d)/|d|$. Here, $c(q,d)$ indicates the count of term q in document d and $|d|$ the length of the particular document. This captures the notion that $P(q|\theta_d)$ is the relative frequency with which we expect to see the term q when we repeatedly and randomly sample terms from this document. The higher this frequency, the more likely it is that this document will be relevant to the query.

2.2 Smoothing

It is clear from Eq. 1, that taking the product of term frequencies has a risk of resulting in a probability of zero: "unseen" terms will produce a probability of zero for that particular document. To tackle this problem, *smoothing* is usually applied, which assigns a very small (non-zero) probability to unseen words. One way of smoothing is called Dirichlet smoothing [4, 22], which is formulated as:

$$P(Q|\theta_d) = P(d) \cdot \prod_{q \in Q} \frac{c(q,d) + \mu P(q|\theta_C)}{|d| + \mu},$$

where θ_C is the language model of a large reference corpus C (such as the collection) and μ a constant by which to tune the influence of the reference model. When comparing the language modeling framework for IR with more well-known TF.IDF schemes, the application of smoothing has an IDF like effect [11, 22].

2.3 Relevance Models

Relevance models are a special class of language models, which are used to estimate a probability distribution θ_Q over terms in a query's vocabulary [16]. The underlying intuition is that the query and the set of relevant documents are both sampled from the same (relevant) term distribution. They differ, however, in the way these distributions are modeled. While general language modeling assumes that queries are generated from documents, relevance models assume that both are generated from an unseen source—the relevance model.

So, how to create such a relevance model? A set of documents R, which has been judged to be relevant to a specific query, can be used as a model from which the terms

are sampled. In the absence of such relevance information, an initial retrieval run is performed and the top-ranked documents are assumed to be relevant. Bayes' rule is then applied to determine the probabilities of the terms given this document set. This approach normally assumes the document prior to be uniform and we obtain:

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|\theta_d). \tag{2}$$

The term $P(w|\theta_d)$ is again estimated using maximum-likelihood techniques. To obtain an estimate of $P(Q|\theta_d)$—the probability of a query, given a document model, i.e., the confidence in a particular document being relevant to the original query—Bayes' rule is applied again, together with Dirichlet smoothing. Eq. 2 thus essentially estimates the "confidence" of translating the original query Q into a particular term w, based on some set of relevant documents R.

2.4 Biasing Relevance Models

We now introduce a new latent variable into Eq. 2, which is derived from documents categorized with thesaurus terms m. Through this model we bias the generation of a relevance model towards terms associated with the thesaurus terms. For any given query we take the l thesaurus terms that are most likely to generate the query, based on some corpus of annotated documents, and then condition the generation of a relevance model on these terms:

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(d|m_1, \ldots, m_l) \cdot P(Q|\theta_d). \tag{3}$$

We assume the thesaurus terms to be independent, so we can express their joint probability $P(d|m_1, \ldots, m_l)$ as the product of the marginals: $\prod_{i=1,\ldots,l} P(d|m_i)$. Each term $P(d|m_i)$ can be estimated using Bayes' rule, by determining the following posterior distribution, based on documents annotated with that particular term:

$$P(d|m) = \frac{P(m|d) \cdot P(d)}{P(m)},$$

where $P(d)$ is again assumed to be uniform. We estimate the prior probability of seeing a thesaurus term as $P(m) = c(m) \cdot |M|^{-1}$ for any given thesaurus term m, where $c(m)$ is the total number of times this thesaurus term is used to categorize a document and $|M| = \sum_{m \in M} c(m)$. Doing so ensures that frequently occuring, more general (and thus less discriminative thesaurus terms) receive a relatively lower probability mass. $P(m|d)$ is estimated in a similar fashion: it is 0 if m is not associated with d, and the reciprocal of the number of thesaurus terms associated with document d otherwise.

2.5 Clipped Relevance Model

Relevance models generally perform better when they are linearly interpolated with the original query estimate—the so-called "clipped relevance model" [14]—using a mixing weight λ:

$$P(w|\theta_Q) = \lambda \cdot \frac{c(w,Q)}{|Q|} + (1-\lambda) \cdot P(w|\hat{\theta}_Q). \tag{4}$$

The final query is thus composed of an initial and an expanded query part, with terms and weights in the latter chosen according to either Eq. 2 or 3. When λ is set to 1, the ranking function reduces to the regular query-likelihood ranking algorithm.

3 Experimental setup

Now that we have put forward our proposed thesaurus-biased expansion algorithm, we turn to answering our research questions. In this section we detail the test collection and experimental setup and in the next we present our findings.

3.1 Test Collection

As our test collection we take the TREC 2006 Genomics test set [9]. The 2006 edition of the TREC Genomics track provides a set of queries (topics), a document collection of full-text biomedical articles, and relevance assessments. The task put forward by the organizers of this particular year's track was to identify relevant documents given a topic and to extract the relevant passages from these documents. The topics themselves were based on four distinct topic templates, and instantiated with specific genes, diseases or biological processes. Relevance was measured at three levels: the document, passage and aspect level. For our experiments, we use the judgments at the document level and those at the aspect level.

All of the documents in this collection are also accessible through PubMed, a bibliographic database maintained by the National Library of Medicine (NLM). It contains bibliographical records of almost all publications from the major biomedical research areas, conferences, and journals and uses controlled vocabulary terms to index the documents. This vocabulary, called MeSH (Medical Subject Headings), is a thesaurus containing 22,997 hierarchically structured concepts, and is used by trained annotators from the NLM to assign one or more MeSH terms to every document indexed in PubMed. These terms can then later be used to restrict, refine, or focus a query, much in the same way a regular library categorization system does.

We base our estimations of the relevance models and the thesaurus-biased models either on the TREC Genomics 2006 document collection or on all the abstracts as found in PubMed. The former collection contains 162,259 full-text documents, whereas the entire PubMed database contains 15,806,221 abstracts. Both have document-level MeSH terms categorizing their content, with an average of around 10 MeSH terms per publication.

3.2 Runs

We created five runs. As a baseline, we perform a regular query-likelihood run (QL) based on Eq. 4 with λ set to 1. We will refer to the run implementing our thesaurus-biased relevance models as MeSH-biased models (MM); it uses Eq. 4 in conjunction with Eq. 3. We compare the results with standard relevance models (RM) which are also estimated using Eq. 4, but with the expanded query portion based on Eq.2. As stated before, we estimate the expanded part of the query either on the TREC Genomics

	λ	$\|R\|$	k	l	MAP	Change	P10	Change
QL	1	-	-	-	0.359		0.45	
RM (collection)	0.10	10	50	-	0.426	+19%	0.48	+7%
RM (PubMed)	0.35	1	50	-	0.425	+18%	0.48	+7%
MM (collection)	0.05	10	10	20	0.424	+18%	0.48	+7%
MM (PubMed)	0.45	1	30	20	**0.429**	+20%	**0.49**	+9%

Table 1. Comparison between different query models and a query-likelihood baseline (best scores in boldface.)

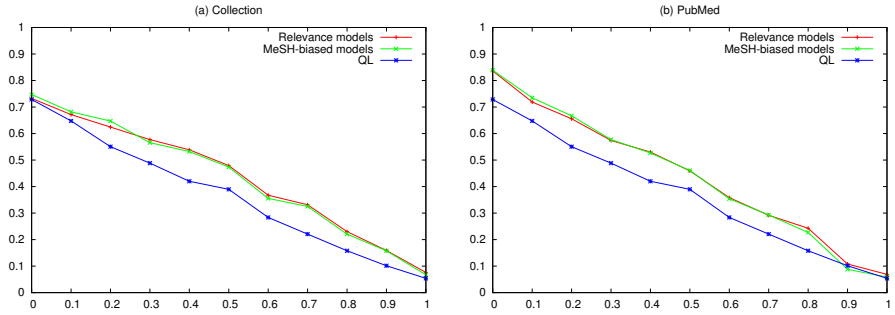

Fig. 1. Precision-recall graphs comparing relevance models and MeSH-biased models, estimated on (a) the collection or (b) PubMed. The results of the baseline are included as reference.

2006 document collection (MM/RM (collection)), or on the contents of the much larger PubMed collection (MM/RM (PubMed)). All runs are morphologically normalized as described by Huang et al. [12] and stemmed using a Porter stemmer .

3.3 Parameters and Optimization

Based on earlier experiments, we fix $\mu = 100$ and focus on the following dimensions; the number of documents used to construct the relevance model ($|R|$), the number of expansion terms (k), the number of MeSH terms used to describe the query (l), and the value of λ. We have compared an exhaustive range of values and, for the sake of conciseness, we only report the optimal ones found.

3.4 Evaluation measures

We compare the five runs (QL, RM (collection/PubMed), MM (collection/Pubmed)) in terms of retrieval effectivess, using precision at 10 (P@10) and mean average precision (MAP). In addition, we look at the thesaurus terms returned by the MM runs, and determine their relevancy as follows. We do not have the resources to recruit domain experts capable of assessing the broad range of topics included in the TREC 2006 Genomics track test collection. Instead, we created "pseudo-relevance judgments." from

Relevance models	
Collection terms	PubMed terms
receptor	ethanol
nicotin	nicotinic
subunit	nicotine
of	chronic
acetylcholin	cells
the	treatment
alpha7	receptor
abstract	mrna
alpha	nachr
medlin	m10
2003	**levels**

MeSH-biased models	
Collection terms	PubMed terms
receptor	ethanol
nicotin	nicotinic
of	nicotine
the	chronic
subunit	cells
humans	treatment
acetylcholin	receptor
animals	mrna
icotinic	nachr
study	m10
alpha7	subunit

Table 2. Comparison of top expansion terms for topic 173: "How do alpha7 nicotinic receptor subunits affect ethanol metabolism?", using estimations from the collection and PubMed. The terms associated with MeSH-biased models, were based on the MeSH terms as described in Table 4. Terms specific to a method are marked in boldface.

the additional assessments provided by TREC Genomics. Besides judging document-level relevance, the assessors for the 2006 Genomics track also used MeSH terms to categorize each relevant passage (the so-called "aspects" [9]). So, for each topic we have a list of MeSH terms which the assessors judged as being descriptive of the relevant passages. We compare this list (per topic) with the top-10 MeSH terms found by the MM runs.

4 Results and Discussion

We present our experimental results in two sets. First, we focus on the retrieval effectiveness of our thesaurus-biased query expansion method. After that we zoom in on the browsing suggestions being generated.

4.1 Thesaurus-biased Relevance Models

Table 1 displays the results of the evaluated runs (best scores in boldface). We note that the MAP score of our baseline is well above the median score achieved by participants of the TREC 2006 Genomics track [9] (which was 0.279). Our first two research questions asked for an effective query expansion method that combines feedback term generation with browsing term generation. We observe that the retrieval effectiveness of our thesaurus-biased models is in the same range as that of relevance models, both when using the collection and when using PubMed as the source for feedback terms, in terms of MAP and P@10 scores—RM and MM statistically significantly outperform the baseline.

We see a mixed picture when the size of the feedback corpus is changed (our third research question). Let us look at the precision-recall graphs. Figure 1 clearly shows

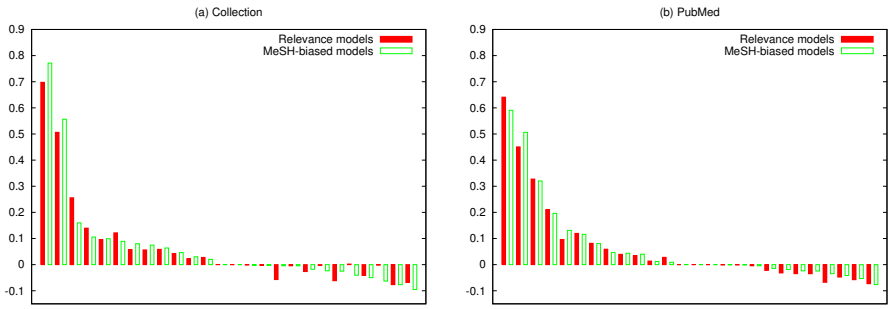

Fig. 2. Sorted difference in per-topic MAP values when comparing MeSH-biased models and relevance models with the query-likelihood baseline, estimated on either (a) the collection or (b) PubMed.

that both models succeed in exceeding the baseline on almost all levels of recall and that estimating on a larger collection mostly helps to improve early precision, i.e., precision at lower recall levels. The improvement of MeSH-biased models over relevance models is marginal, however, and visible only at the lower recall levels.

A Closer Look: Feedback Terms Next, we look at the specific expansion terms which each model finds from the vocabulary. Table 2 provides a detailed example of the top-10 vocabulary terms which are found for the same topic. While the terms themselves change little (viz. the second and last column of Table 2), the assigned term weights do, which is the main cause of the increase in performance. The effect of basing the estimations on PubMed are visible in the specificity of the expansion terms. This is witnessed, for example, in the addition of low content-bearing terms such as "the" and "of" using the collection.

Topic Details Figure 2 displays the per-topic change in MAP scores for the baseline run and the MeSH-biased model over the baseline. When zooming in on these individual topics, we find that applying MeSH-biased relevance models only helps in half of the cases (13 out of 26). Our model performs slightly better than relevance models, but this result is not significant (when tested with a Wilcoxon signrank test at $p = 0.05$)—an effect which is probably due to the small size of the topic set. Put more positively, the performance of the models is at a comparable level, while our approach readily facilitates browsing activities through the found thesaurus terms.

We observe that some queries benefit from applying thesaurus-biased relevance models, whilst others are helped by the estimation of a traditional relevance model. Because these models perform differently on different topics, we investigate possible ways of predicting which model to use on which topic. There are several methods of predicting and classifying a priori classes of query difficulty. One of these is through determining the *query clarity*, which is a way of quantifying the possible ambiguity in a query [1, 3, 7]. According to Cronen-Townsend et al. [7], it correlates at a signifi-

Bin	Topic type	Topics
1.	Find articles describing the role of a gene involved in a given disease.	6
2.	Find articles describing the role of a gene in a specific biological process.	8
3.	Find articles describing interactions between two or more genes in the function of an organ or disease.	7
4.	Find articles describing one or more mutations of a given gene and its biological impact.	7

Table 3. Generic topic types on which each topic is based together with the number of topics which were created with it.

cant level with the resulting retrieval performance of that query. However in our case, we find no significant correlation between the query clarity scores and the resulting performance for the current topic set.

A specific feature of the current topic set is that the topics are generated based on four templates, or so-called "generic topic types" [9], which are represented in Table 3. The table shows each template, as well as the actual number of topics based on it. The topic templates emphasize different search tasks, which may in turn influence the effectiveness of the various approaches. If this is the case, then that would indicate that this particular "class" of topics is sensitive to the chosen model. To understand the issue, we bin the topics per topic type and determine the means of all the results per model (and bin). We test if the differences between the means of MAP of all runs, grouped by the four combinations of models and collections, are significantly different for each topic type. We use a Newman-Keuls test [6] to do a pair-wise comparison and test the null hypothesis that the means of a group is equal to that of another group.

When tested, we find that only topic type 1 and 2 have a significantly different performance for each model ($p < 0.01$) and only when estimation is done using PubMed. In these case, the mean of the MeSH-biased model is higher than the mean of the relevance model. We conclude that our proposed algorithm does a better job at finding relevant documents for these topic types. We argue that they have a more "open" nature than the other two, suggesting that our method favors more open queries. Whether this is the case and whether it holds for a larger topic set remains as future work.

4.2 Thesaurus Terms Generated

Finally, we turn to another aspect of our algorithm's output: the MeSH terms being generated for browsing purposes. Table 4 shows the MeSH terms found for topic 173 (the first topic in the topic set has number 160), using our proposed approach. Estimating $P(m)$ on a smaller corpus (first column) has the effect of introducing slightly more general terms, e.g., "Research support" and "Humans," which might account for the slightly lower scores for this particular method of estimation. The MeSH terms estimated from PubMed are more specific, e.g., "Bungarotoxins" and "Nicotinic (ant)agonists." We can quantify this observation (thesaurus terms generated by MM(collection) tend to be somewhat more general than thesaurus terms generated by MM(PubMed)), by computing for every topic the average distance to the root of the MeSH thesaurus of the sug-

MeSH-biased models	
Collection MeSH terms	PubMed MeSH terms
Animals	Receptors, Nicotinic
Humans	Ethanol
Research Support, Non-U.S. Gov't	Nicotinic Agonists
Receptors, Nicotinic	Animals
Research Support, U.S. Gov't, P.H.S.	Central Nervous System Depressants
Brain	Nicotine
Mice	Mice
Comparative Study	Bungarotoxins
Ion Channels	Nicotinic Antagonists
Membrane Proteins	Rats
Immunohistochemistry	Receptors, Serotonin

Table 4. Comparison of top MeSH terms for topic 173: "How do alpha7 nicotinic receptor subunits affect ethanol metabolism?", using estimations from the collection and PubMed.

gested thesaurus; so, the lower distance, the more abstract the terms. For MM(collection) the average distance to the root was 4.46 while for MM(medline) it was 4.78.

Finally, what is the quality of the generated thesaurus terms, using the evaluation criteria put forward in Section 3? When we estimate the MeSH-biased model on the collection, on average 2.3 MeSH terms per topic match. When we look at the estimation from PubMed, 3 out of the 10 MeSH terms match. The difference between these two is significant at the $p < 0.05$ level, using a Wilcoxon signrank test.

5 Related Work

Besides earlier mentioned query expansion work, most other related work can be found among language modeling approaches to information retrieval. Tao et al. [19] describe a method to create augmented language models, based on the documents in a collection. The authors assume, in a similar fashion as with relevence models, that a document is itself generated from an unseen language model. So, instead of expanding queries, they expand the documents to better describe this underling generative model. The authors argue that such an *enlarged* document is a better representation, which is reflected in the reported increases in retrieval performance.

The method most closely in line with the current work, however, is described by Collins-Thompson and Callan [5]. The authors describe a way of combining multiple sources of evidence to predict relationships between query and vocabulary terms, which uses a Markov chain framework to integrate semantic and lexical features into a relevance model. The semantic features they investigate are general word associations and synonymy relations as defined in WordNet. Cao et al. [2] describe a more principled way of integrating WordNet term relationships into statistical language models, but they do not use relevance models. Both methods are evaluated on "general" corpora—viz. news collections—and result in consistent improvements. We, however, place our

work in a digital library setting, where document-level annotations play an important role. Our work differs from these approaches in the fact that we particularly focus on, and utilize. the knowledge that has gone into the construction and assignment of controlled vocabulary terms to documents. Doing so enables our approach to assist the user in browsing a collection, while keeping end-to-end retrieval performance comparable with other state-of-the-art approaches.

6 Conclusion

We have described a transparent method to integrate document-level annotations in a retrieval model based on statistical language models. Our goal was to incorporate the information and semantics stored in a document categorization system to achieve effective query expansion, while at the same time facilitating browsing.

We evaluated our algorithm in a biomedical setting, using the TREC 2006 Genomics track test set and the MeSH thesaurus. We used the terms from this thesaurus and the documents annotated with them to bias the estimation of a relevance model—a special class of statistical language models. We determined the impact of increasing the size of the document set on which we base our estimations on the quality of the found thesaurus terms, and found a significant difference in favour of the larger PubMed database.

We have found a 20% improvement in mean average precision when comparing the end-to-end retrieval results of our model with a query-likelihood baseline. When we look at estimating our model from the much smaller evaluation collection, we find a 19% increase in mean average precision over the same query-likelihood baseline. These results would have put these runs in the top segment of the runs submitted to the TREC 2006 Genomics track. We have looked at ways to determine, which type of topic benefits from which approach. When we group the topics together based on the topic template, we find a statistically significant difference in favour of our method for two out of four particular topic "classes."

Future work includes an analysis on a larger set of topics, as well as incorporating the tree-like structure inherent in a thesaurus—we have assumed thesaurus terms to be independent, while in fact they may not be. Finally, we will look for additional ways to predict if and when biased query modeling is beneficial.

7 Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project (http://www.vl-e.nl), which is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

8 References

[1] J. Allan and H. Raghavan. Using part-of-speech patterns to reduce query ambiguity. In *SIGIR '02*, pages 307–314, 2002.

[2] G. Cao, J.-Y. Nie, and J. Bai. Integrating word relationships into language models. In *SIGIR '05*, pages 298–305, 2005.
[3] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *SIGIR '06*, pages 390–397, 2006.
[4] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *ACL*, pages 310–318, 1996.
[5] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *CIKM '05*, pages 704–711, 2005.
[6] W. Cooley and R. Lohnes. *Multivariate data analysis*. Wiley, 1971.
[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02*, pages 299–306, 2002.
[8] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, pages 154–161, 2006.
[9] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli. TREC 2006 Genomics track overview. In *TREC Notebook*. NIST, 2006.
[10] J. R. Herskovic, L. Y. Tanaka, W. Hersh, and E. V. Bernstam. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *J Am Med Inform Assoc*, 14 (2):212–220, 2007.
[11] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL '98*, pages 569–584, 1998.
[12] X. Huang, Z. Ming, and L. Si. York University at TREC 2005 Genomics track. In *Proceedings of the 14th Text Retrieval Conference*, 2005.
[13] T. Koch, A. Ardö, and K. Golub. Browsing and searching behavior in the renardus web service a study based on log analysis. In *JCDL '04*, pages 378–378, 2004.
[14] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing?: Iterative pseudo-query processing using cluster-based language models. In *SIGIR '05*, pages 19–26, 2005.
[15] A. M. Lam-Adesina and G. J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *SIGIR '01*, pages 1–9, 2001.
[16] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, 2001.
[17] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *SIGIR '98*, pages 206–214, 1998.
[18] K. F. Tan, M. Wing, N. Revell, G. Marsden, C. Baldwin, R. MacIntyre, A. Apps, K. D. Eason, and S. Promfett. Facts and myths of browsing and searching in a digital library. In *ECDL '98*, pages 669–670, 1998.
[19] T. Tao, X. Wang, Q. Mei, and C. Zhai. Accurate language model estimation with document expansion. In *CIKM '05*, pages 273–274, 2005.
[20] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *SIGIR '93*, pages 171–180, 1993.
[21] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th ACM SIGIR conference*, pages 4–11, 1996.
[22] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.