The University of Amsterdam at the CLEF 2008 Domain Specific Track

Parsimonious Relevance and Concept Models

Edgar Meij
emeij@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam

Abstract

We describe our participation in the CLEF 2008 Domain Specific track. The research questions we address are threefold: (i) what are the effects of estimating and applying relevance models to the domain specific collection used at CLEF 2008, (ii) what are the results of parsimonizing these relevance models, and (iii) what are the results of applying concept models for blind relevance feedback? Parsimonization is a technique by which the term probabilities in a language model may be re-estimated based on a comparison with a reference model, making the resulting model more sparse and to the point. Concept models are term distributions over vocabulary terms, based on the language associated with concepts in a thesaurus or ontology and are estimated using the documents which are annotated with concepts. Concept models may be used for blind relevance feedback, by first translating a query to concepts and then back to query terms. We find that applying relevance models helps significantly for the current test collection, in terms of both mean average precision and early precision. Moreover, parsimonizing the relevance models helps mean average precision on title-only queries and early precision on title+narrative queries. Our concept models are able to significantly outperform a baseline query-likelihood run, both in terms of mean average precision and early precision on both title-only and title+narrative queries.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Parsimonious Models, Language Models, Relevance Feedback

1 Introduction

We describe our participation in the 2008 CLEF Domain Specific track. Our main motivation for participating was to evaluate the retrieval models we have developed for another, very similar domain on the CLEF Domain Specific test collection. Our concept models have thus far been developed and evaluated on the TREC Genomics test collections, which also consists of documents which are manually annotated using concepts from a thesaurus [5, 6].

The main idea behind our approach is to model the language use associated with concepts from a thesaurus or ontology. To this end we use the document annotations as a bridge between vocabulary terms and the concepts in the knowledge source at hand. We model the language use around concepts using a generative language modeling framework, which provides theoretically sound estimation methods and builds upon a solid statistical background.

Our concept models may be used to determine semantic relatedness or to generate navigational suggestions, either in the form of concepts or vocabulary terms. These can then be used as suggestions for the user or for blind relevance feedback [13, 14, 18]. In order to apply blind relevance feedback using our models, we perform a double translation. First we estimate the most likely concepts given a query and then we use the most distinguishing terms from these concepts to formulate a new query. In a sense we are using the concepts as a pivot language [9]. To find the most distinguishing terms given a concept, we apply a technique called *parsimonization* [8]. Parsimonization is an algorithm based on expectation-maximization (EM) [3] and may be used to re-estimate probabilities of one model with respect to another. Events that are well-predicted by the latter model will lose probability mass, which in turn will be given to the remaining events. Recently, we have successfully applied parsimonization to the estimation of relevance models on a variety of tasks and collections [15]. In all of these cases, as well as with our concept models, we find that interpolating the newly found query with the original one yields the best performance—an observation which is in line with the literature [10].

The research questions we address are threefold: (i) what are the effects of estimating and applying relevance models to the collection used at the CLEF 2008 Domain Specific track [12], (ii) what are the results of parsimonizing these relevance models, and (iii) what are the results of applying our concept models for blind relevance feedback?

We find that applying relevance models helps significantly for the current test collection, in terms of both mean average precision and early precision. Moreover, we find that parsimonizing the relevance models helps mean average precision on title-only queries and early precision on title+narrative queries. Our concept models are able to significantly outperform a baseline query-likelihood run, both in terms of mean average precision and in terms of early precision on both title-only and title+narrative queries.

The remainder of this paper is organized as follows. In Section 2 we introduce the retrieval framework we have used for our submission, i.e., statistical language modeling. In Section 3 and 4 we introduce the specifics of our models and techniques. In Section 5 we describe the experimental setup, our parameter settings, and the preprocessing steps we performed on the collection. In Section 6 we discuss our experimental results and we end with a concluding section.

2 Language Modeling

Language modeling is a relatively new framework in the context of information retrieval [7, 16, 20]. It is centered around the assumption that a query as issued by a user is a sample generated from some underlying term distribution—the information need. The documents in the collection are modeled in a similar fashion and are usually considered to be a mixture of a document-specific model and a more general background model. At retrieval time, each document is ranked according to the likelihood of having generating the query (query-likelihood).

Lafferty and Zhai [11] propose to generalize the query likelihood model to the KL-divergence scoring method, in which the query is modeled separately. Scoring documents then comes down to measuring the divergence between a query model $P(t|\theta_Q)$ and each document model $P(t|\theta_D)$, in which the divergence is negated for ranking purposes. When the query model is generated using the empirical, maximum-likelihood estimate (MLE) on the original query as follows:

$$P(t|\tilde{\theta}_Q) = \frac{n(t;Q)}{|Q|}, \tag{1}$$

where $n(t;Q)$ is the number of occurrences of term t in query Q and $|Q|$ the length of the query, it can be shown that documents are ranked in the same order as using the query likelihood model [20]. More formally, the score for each document given a query using the KL-divergence retrieval model is:

$$\text{Score}(Q,D) = -\text{KL}(\theta_Q||\theta_D) = -\sum_{t\in\mathcal{V}} P(t|\theta_Q)\log P(t|\theta_D) + \sum_{t\in\mathcal{V}} P(t|\theta_Q)\log P(t|\theta_Q), \tag{2}$$

where $\mathcal{V}$ denotes the vocabulary. The entropy of the query—$\sum_{t\in\mathcal{V}} P(t|\theta_Q)\log P(t|\theta_Q)$—remains constant per query and can be ignored for ranking purposes.

2.1 Smoothing

Each document model is estimated as the MLE of each term in the document $P(t|\theta_D)$, linearly interpolated with a background language model $P(t)$, which in turn is calculated as the likelihood of observing t in a sufficiently large collection, such as the document collection:

$$P(t|\theta_D) = \beta P(t|\theta_D) + (1-\beta)P(t). \tag{3}$$

We smooth using Bayesian smoothing with a Dirichlet prior and set $\beta = \frac{\mu}{|D|+\mu}$ and $(1-\beta) = \frac{|D|}{|D|+\mu}$, where μ is the Dirichlet prior that controls the influence of smoothing [2, 22].

2.2 Query Modeling

Relevance feedback can be applied to better capture a user's information need [1, 12, 19]. In the context of statistical language modeling, this is usually performed by estimating a new query model, viz. $P(t|\theta_Q)$, in Eq. 2 [16, 21]. Automatically reformulating queries (or *blind* relevance feedback) entails looking at the terms in some set of (pseudo-)relevant documents and selecting the most informative ones with respect to the set or the collection. These terms may then be reweighed based on information pertinent to the query or to the documents and—in a language modeling setting—be used to estimate a query model.

Relevance modeling is one specific technique by which to estimate a query model given a set of (pseudo-)relevant documents $\mathcal{D}_Q$. The query and documents are both taken to samples of an underlying generative model—the relevance model. There are several ways by which to estimate the parameters of this model given the observed data (query and documents), each following a different independence assumption [12]. For our current experiments we use method 2, which is formulated as:

$$P(t|\hat{\theta}_Q) \propto P(t) \prod_{q_i\in Q} \sum_{D_i\in\mathcal{D}_Q} P(q_i|\theta_{D_i})P(\theta_{D_i}|t), \tag{4}$$

where $q_1,\ldots,q_k$ are the query terms, D a document, and t a term. Bayes' rule is used to estimate the term $P(\theta_D|t)$:

$$P(\theta_D|t) = \frac{P(t|\theta_D)P(\theta_D)}{P(t)}, \tag{5}$$

where we assume the document prior $P(\theta_D)$ to be uniform. Similar to Eq. 3, the term $P(t|\theta_D)$ may be interpreted as a way of accounting for the fact that the (pseudo-)relevant documents contain terms related to the information need as well as terms from a more general model. We set it to the following mixture:

$$P(t|\theta_D) \quad = \quad \alpha\frac{n(t;D)}{|D|} + (1-\alpha)P(t), \tag{6}$$

where $P(t)$ is the probability of observing t in a sufficiently large collection such as the entire document collection. Query models obtained using relevance models perform better when they are subsequently interpolated with the initial query using a mixing weight λ [10]:

$$P(t|\theta_Q) \quad = \quad \lambda P(t|\tilde{\theta}_Q) + (1-\lambda)P(t|\hat{\theta}_Q) \tag{7}$$

3 Concept Models

In order to leverage the explicit knowledge encapsulated in the GIRT/CSASA thesauri, we perform blind relevance feedback using the concepts defined therein. We leverage the dual document representations—concepts and terms—to create a generative language model for each concept, which bridges the gap between terms and concepts. Related work has also used textual representations to represent concepts, see e.g., [4, 17], however, we use statistical language modeling techniques to parametrize the concept models, by leveraging the dual representation of the documents.

To incorporate concepts in the retrieval process, we propose a conceptual query model which is an interpolation of the initial query with another query model. This model is obtained from a double concept translation. In this translation, concepts are used as a pivot language [9]; the initial query is translated to concepts and back to expanded query terms:

$$P(t|\theta_Q) = \lambda P(t|\tilde{\theta}_Q) + (1-\lambda) \sum_{c \in \mathcal{C}} P(t|c)P(c|Q). \tag{8}$$

Note that we assume that the probability of selecting a term is no longer dependent on the query once we have selected a concept given that query. Then, two components need to be estimated: the probability of a concept given a query $P(c|Q)$ and of a term given a concept $P(t|c)$. To acquire $P(t|c)$, we will use the assignments of GIRT/CSASA thesaural concepts to the documents in the collection and aggregate over the documents $\mathcal{D}_c$ which are labeled with a particular concept c:

$$P(t|c) = \sum_{D \in \mathcal{D}_c} P(t|D,c)P(D|c).$$

We drop the conditional dependence of t on c given D, again assume the document prior to be uniform, and apply Bayes' rule to obtain:

$$P(t|c) = \frac{1}{P(c)} \sum_{D \in \mathcal{D}_\mathcal{C}} P(t|\theta_D)P(c|\theta_D), \tag{9}$$

where $P(c)$ is a maximum likelihood (ML) estimate on the collection:

$$P(c) = \frac{\sum_D n(c;D)}{\sum_{c'} \sum_{D'} n(c';D')}$$

and $P(c|\theta_D)$ is determined using the ML of the concepts associated with that document

$$P(c|\theta_D) = \frac{n(c;D)}{\sum_{c'} n(c';D)}. \tag{10}$$

Next, we also need need a way of estimating concepts for each query, which means that we are looking for a set of concepts $\mathcal{C}_Q$ such that $c \in \mathcal{C}_Q$ have the highest posterior probability $P(c|Q)$. We approach this by looking at the assignment of concepts to documents ane again consider documents which are related to the original query, by using the top ranked documents $\mathcal{D}_Q$ from the initial retrieval run:

$$P(c|Q) = \sum_{D \in \mathcal{D}_Q} P(c|\theta_D)P(D|Q), \tag{11}$$

where $P(D|Q)$ is determined using the retrieval scores. Note that we again assume that the probability of observing a concept is independent of the query, once we have selected a document given the query, i.e., $P(c|D,Q) = P(c|\theta_D)$. This enables us to directly use Eq. 10.

4 Parsimonization

If $P(t|\theta_D)$ and $P(c|\theta_D)$ in Eq. 6 and Eq. 10 are estimated based on MLE, general terms and concepts may acquire too much probability mass, simply because they occur more frequently. Parsimonization may be

Table 1: Empirical results of our submitted runs, in terms of mean average precision (MAP) and precision@10 (P10). Best scores are marked in boldface. A †/‡ indicates a statistically significant difference as compared to the baseline at the 0.05/0.01 level respectively (tested using a Wilcoxon signed rank test).

	title		title+narrative	
	MAP	P10	MAP	P10
UAmsBaseline	0.2433	0.4560	0.2077	0.4600
UAmsRelModels	0.2737‡	0.5040	0.2396	0.4400
UAmsParsRelModels	0.2779†	0.5000	0.2271	0.4800†
UAmsConceptModels	**0.2922**†	**0.5160**†	**0.2581**†	**0.5240**†

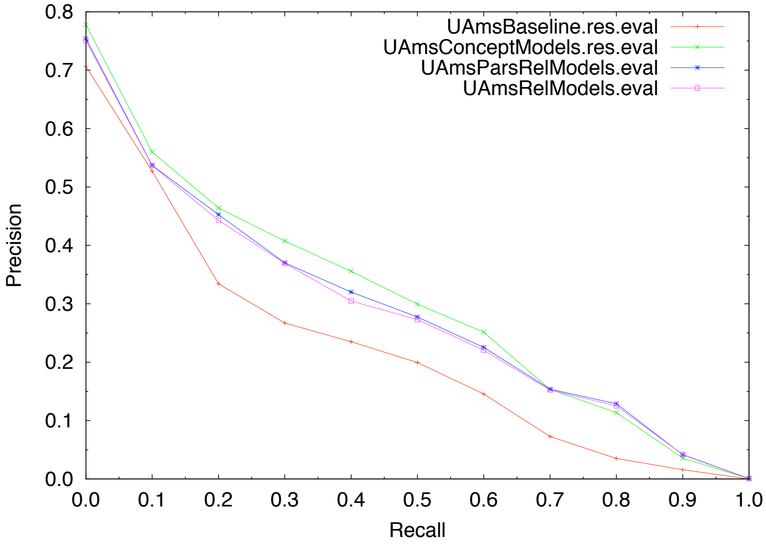

Figure 1: Precision-recall graph of the various runs.

used to reduce the amount and probability mass of non-specific terms in a language model by iteratively adjusting the individual term probabilities based on a comparison with a large reference corpus, such as the collection [8]. While one of the introduced models may already contain a way of incorporating a reference corpus, viz. Eq. 6, we propose to make the estimates more sparse. Doing so enables more specific terms to receive more probability mass, thus making the resulting model more to the point. In order to achieve this, we consider both models to be a mixture of a document model $P(x|\theta_D)$ and a background model $P(x)$, where $x \in \{t, c\}$, and we "parsimonize" the estimates through applying the following EM algorithm until the estimates do not change significantly anymore:

E-step:
$$e_x = n(x;D)\frac{\gamma P(x|\theta_D)}{(1-\gamma)P(x)+\gamma P(x|\theta_D)}$$

M-step:
$$P(x|\theta_D) = \frac{e_x}{\sum_{x'} e_{x'}}$$

5 Experimental Setup

We did not perform any preprocessing on the document collection, besides replacing German characters as well as HTML entities. To estimate our concept models, we have used the CONTROLLED-TERM-EN field in the documents. We submitted the following runs:

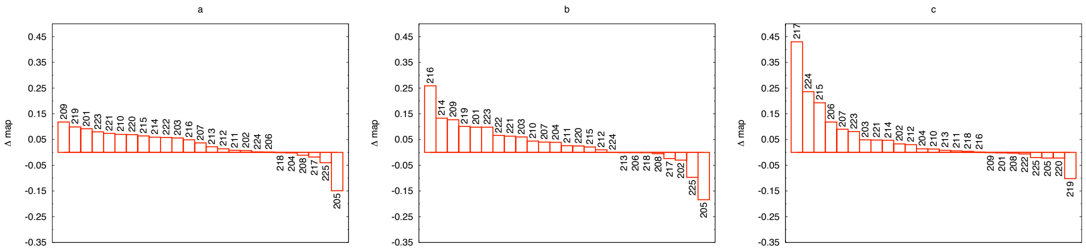

Figure 2: Per-topic breakdown of the difference in terms of average precision between the baseline run and `UAmsRelModels` (a), `UAmsParsRelModels` (b), and `UAmsConceptModels` (c). Topics are sorted by increasing difference, the labels indicate the respective topic identifiers.

`UAmsBaseline` – a baseline run based with $P(t|\theta_Q)$ in Eq. 2 set to the empirical estimate on the query (Eq. 1),

`UAmsRelModels` – a run based on relevance models, viz. Eq. 4,

`UAmsParsRelModels` – a run based on parsimonious relevance models, viz. Eq. 4 in conjunction with the E- and M-steps described in the previous section,

`UAmsConceptModels` – a run based on concept models, viz. Section 3 and Section 4.

Something went wrong with the submitted `UAmsParsRelModels` run based on parsimonious relevance models, making it identical to the `UAmsRelModels` run. In this paper we report on the corrected version.

In all runs which use blind relevance feedback, we use the 5 terms with the highest probability from the 10 highest ranked documents to estimate our query models. We have then used the 2007 CLEF Domain Specific topics to find the optimal parameter settings for α (Eq. 6) and λ (Eq. 7 and Eq. 8). For our current experiments we set $\mu = 50$ and fix $\gamma = 0.15$ [8].

6 Results and Discussion

Table 1 lists the results of our runs. On the 2007 data, we found that adding the narrative field of the topics helps retrieval effectiveness. For comparative purposes we have included results for both title-only and title+narrative runs. On the 2008 topics, we do not find the same improvement when adding the narrative field, besides slightly improving precision@10. When looking at the longer topics (title+narrative), applying parsimonization to the relevance models hurts mean average precision, but helps early precision. This precision-enhancing effect is in line with earlier results [15].

The proposed concept models improve significantly over the query-likelihood baseline, both in terms of mean average precision and precision@10 and for both title-only and title+narrative topics. From the precision-recall plot in Figure 1 (title-only) it is clear that our concept model improves slightly in early precision and that the biggest gain is obtained between recall levels 0.2 and 0.7. It also shows that the relevance modeling approaches mainly help to improve recall and not so much precision.

Figure 2 displays a per-topic comparison between the query-likelihood run and each of the other runs. From these contrastive plots it emerges that topics 205 and 225 are hurt most when using relevance models. Further analysis should indicate which characteristics of these topics are responsible for this result. Interestingly, these two topics are hurt less when we apply our concepts models, whereas topic 219 is hurt most in this run. On the other side of the graph, there are quite a few topics which are helped using either relevance models or concept models. Especially topic 216 is improved when applying parsimonious relevance models ($>$ 0.25 increase in mean average precision). The positive difference when applying concept models is even more distinctive; topic 217 is nearly improved by a 0.5 increase in mean average precision.

7 Conclusion

We have described our participation in this year's CLEF Domain Specific track. Our aim was to evaluate blind relevance feedback models as well as concept models on the CLEF Domain Specific test collection. The results of our experiments show that applying relevance modeling techniques has a significant positive effect on the current topics, in terms of both mean average precision and precision@10. Moreover, we find that parsimonizing the relevance models helps mean average precision on title-only queries and early precision on title+narrative queries. When we apply concept models for blind relevance feedback, we observe an even bigger as well as significant improvement over the query-likelihood baseline, also in terms of mean average precision and early precision. Moreover, unlike (parsimonious) relevance models, our concept model improves title-only as well as title+narrative queries on both measures.

8 Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project, 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 E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815.

9 References

[1] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR '03*, 2003.

[2] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *ACL '96*, 1996.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07*, 2007.

[5] W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 Genomics track overview. In *TREC '05*, 2005.

[6] W. Hersh, A. Cohen, and P. Roberts. TREC 2007 Genomics track overview. In *TREC '07*, 2007.

[7] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL '98*, 1998.

[8] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR '04*, 2004.

[9] W. Kraaij and F. de Jong. Transitive probabilistic CLIR models. In *RIAO '04*, 2004.

[10] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing?: iterative pseudo-query processing using cluster-based language models. In *SIGIR '05*, 2005.

[11] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*, 2001.

[12] V. Lavrenko and B. W. Croft. Relevance based language models. In *SIGIR '01*, 2001.

[13] E. Meij and M. de Rijke. Thesaurus-based feedback to support mixed search and browsing environments. In *ECDL '07*, 2007.

[14] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Parsimonious concept modeling. In *SIGIR '08*, 2008.

[15] E. Meij, W. Weerkamp, K. Balog, and M. de Rijke. Parsimonious relevance models. In *SIGIR '08*, 2008.

[16] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, 1998.

[17] D. R. Recupero. A new unsupervised method for document clustering by using wordnet lexical and conceptual relations. *Inf. Retr.*, 10(6):563–579, 2007.

[18] D. Trieschnigg, E. Meij, M. de Rijke, and W. Kraaij. Measuring concept relatedness using language models. In *SIGIR '08*, 2008.

[19] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96*, 1996.

[20] C. Zhai. *Risk Minimization and Language Modeling in Text Retrieval*. PhD thesis, Carnegie Mellon University, 2002.

[21] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, 2001.

[22] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.