The AID Group at TREC Genomics 2005

Leonie IJzereef[1] **Edgar Meij**[1] **Leif Azzopardi**[1] **Jaap Kamps**[1,2] **Maarten de Rijke**[1]

[1] Informatics Institute, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: This paper describes our participation in the TREC 2005 Genomics track. We took part in the ad hoc retrieval task and aimed at integrating thesauri in the retrieval model. We developed three thesauri-based methods, two of which made use of the existing MeSH theasurus and terms. One method uses blind relevance feedback on MeSH terms, the other uses an index of the MeSH thesaurus for query expansion. The third method makes use of a dynamically generated lookup list, by which gene acronyms and synonyms could be inferred. We show that, despite the relatively minor improvements in retrieval performance of individually applied methods, a combination works best and is able to deliver significant improvements over the baseline.

1 Introduction

The main focus of our participation in the TREC 2005 Genomics track was to evaluate the impact of integrating thesauri and related expansion methods in the retrieval model. We learned from interviews with biomedical researchers that the general search strategy within this domain is geared towards achieving high recall without losing early precision. We hypothesized that the structure of a controlled vocabulary could increase retrieval performance in general and recall in particular. Our working assumption was that controlled vocabulary terms can help overcome problems with synonymy and ambiguity. Thus achieving a higher recall rate by addressing the synonymy issue, but maintaining precision by removing ambiguity. To this end we investigated the results of three thesaurus-based methods.

Our first method comprises of the automatic extraction of synonyms and acronyms from the corpus and the Medical Subject Headings (MeSH) thesaurus. Gene names have a large number of possible synonyms and acronyms. We posited that using the controlled vocabulary terms from the documents and MeSH thesaurus during retrieval would minimize the negative effects of synonymy and improve recall.

Secondly, we attempted to boost precision by performing blind relevance feedback using the MeSH terms associated with the topics and MEDLINE abstracts, similar to the approach used by Kraaij et al. [7]. Finally, we attempted to exploit the textual concept descriptions within the MeSH thesaurus itself by performing query expansion using the contents of these descriptions. We found that the first two methods provided small increases in retrieval effectiveness. However, a combination of the two methods delivered significantly better precision and recall.

The remainder of this paper is organized as follows. In Section 2 we describe our data processing and models employed for this year's edition of TREC Genomics. Then we elaborate on our proposed methods in Section 3, followed by our experiments in Section 4. We present the results of our submitted runs in Section 6 and summarize our findings in a concluding section.

2 Experimental setup

2.1 Collection processing

The document collection consists of a 10-year subset of MEDLINE, which contains over 4.5 million abstracts (totaling 9 Gb in size). Before indexing, the corpus required some preprocessing. First we selected the fields that might be useful for retrieval, as shown in Table 1. We indexed each field in Lucene [8]. Standard stopwords were removed, but no form of stemming was applied.

Field	*Description*
PMID	PubMed Unique Identifier
TI	Title
AB	Abstract
MH	MeSH Terms
OAB	Other Additional Abstract (concatenated with AB)

Table 1: Citation fields

For the MeSH terms field, we only indexed the main MeSH terms. We ignored any additional qualifiers, such as the topical subheadings. Special characters, such as the asterisks used to identify a document's most important MeSH term

were also ignored. In order to preserve the complex MeSH terms we translated all terms to their unique Unified Medical Language System (UMLS) id's before indexing the document collection.

2.2 Query preprocessing

There were five generic topic templates defined for the TREC 2005 Genomics track. For each template the predefined components were identified using regular expressions. These were then removed and the remaining terms were considered the free text query submitted for that topic. For example in topic 120 (shown below), the query terms are highlighted in bold-face and the remaining terms were removed. All methods and/or runs make use of the preprocessed queries.

120. *Provide information on the role of the gene* ***nucleoside diphosphate kinase (NM23)*** *in the process of* ***tumor progression***.

2.3 Language Modeling

To get our baseline run we used the standard version of Lucene with the ILPS extension [4, 8]. Based on the training data we concluded that a language model approach yielded better results than Lucene's vector-space based variant. All our retrieval runs, were therefore based on a multinomial language model, with tunable length prior and Jelinek-Mercer smoothing [2]. We estimated a language model for each document in the collection. For a given query we rank the documents with respect to the likelihood that the document language model generated the query. This can be viewed as estimating the probability $P(d,q)$,

$$P(d,q) = P(d) \cdot P(q|d), \tag{1}$$

where d is a document and q is the query. Thus, we need to estimate two probabilities: the prior probability of the document, $P(d)$; and the probability of generating the query, $P(q|d)$. For the probability of the query we assume the terms to be independent, and we use a linear interpolation of a document model and a collection model to estimate the probability of a query term. The probability of a query $t_1,\ldots,t_n$ is estimated as,

$$P(t_1,\ldots,t_n|d) = \prod_{i=1}^{n} \big(\lambda \cdot P(t_i|d) + (1-\lambda) \cdot P(t_i)\big), \tag{2}$$

where $P(t_i|d)$ is the probability of observing a term in a document, and $P(t_i)$ is the probability of observing the term in the collection. Both probabilities are estimated using the maximum likelihood estimate. The parameter λ is the so-called smoothing parameter. The calculation of probabilities can be reduced to the scoring formula for an indexing unit e and query $t_1,\ldots,t_n$,

$$s(d,t_1,\ldots,t_n) = \beta \cdot \log\left(\sum_t tf(t,d)\right) + \sum_{i=1}^{n} \log\left(1 + \frac{\lambda \cdot tf(t_i,d) \cdot \left(\sum_t df(t)\right)}{(1-\lambda) \cdot df(t_i) \cdot \left(\sum_t tf(t,d)\right)}\right), \tag{3}$$

where $tf(t,d)$ is the frequency of term t in document d; $df(t)$ is the count of units in which term t occurs; and λ is the weight given to the document language model when smoothing with the collection model. Note that $P(t_i)$ is proportional to $df(t_i)$. The parameter β serves as a dynamic parameter by which to tune the system when trying to bridge the length gap between an average document and an average relevant document. For $\beta = 0$ this results in a uniform distribution over length or using no length prior, and for $\beta = 1$ this results in a normal length prior.

Based on experiments with the development data, the parameters for our language model were set as follows:

- LM-lambda: smoothing parameter for Jelinek-Mercer smoothing. Set to its default value of 0.15. Higher values gave lower scores.
- LM-beta: the length prior is set to 1.0. Lower prior gave lower scores.
- LM-cmodel: the collection model is set to document frequency, because when set to collection frequency, retrieval performance degraded.

3 Methods

In this section we provide a description of our proposed methods. Our first method uses a dynamically created look-up list of gene names; the remaining two try to use the contents of the MeSH thesaurus.

3.1 Gene name expansion (`Ge`)

Since all topics were formulated based on only five different generic topic templates, we used the structure of these templates to identify possible gene names within the topics. This approach works only for topics in which there are gene names present. For example, in the following template a topic is created by filling the empty slots with respectively a gene name and a disease name:

> *Provide information about the role of the gene ... in the disease ...*

Although most gene names have several synonyms and acronyms, usually only one of these is used. To be able to identify relevant documents that contain one of the alternative names, we expanded our queries with gene name variants.

The query expansion was based on identifying synonyms and acronyms of gene names, which came from two different sources: the MeSH thesaurus and the MEDLINE corpus. Within the MeSH thsaurus, synonyms are defined between MeSH terms in a separate field.

So whilst we could use the MeSH terms directly, we had to process the MEDLINE collection in order to extract any tacit or latent acronyms within the corpora. This was performed by extracting pairs of full gene names and acronyms from the abstracts, using heuristics based on the cooccurrence of full gene names, round brackets and abbreviations. For instance:

> *...binds* ***hepatocyte nuclear factor 4 (HNF4)*** *and COUP/TF-related proteins...*

This resulted in an acronym list of 33,417 combinations (13,386 unique acronyms). The acronym list and the MeSH thesaurus were used for a simple lookup procedure; if a gene name could be found in one or both, we added all its synonyms and acronyms to the query. An additional restriction has been placed on this method; the original gene name (or one of its variants) has to be present in each retrieved document. Documents without the gene name or one of its variants were discarded. This results in the expanded query when applied to, for example, topic 111:

<111> *Provide information about the role of the gene PRNP in the disease Mad Cow Disease.*

111 *+(PRNP "protein gene" "prp gene" "prion protein gene") Mad Cow Disease*

3.2 MeSH based feedback (Fb)

For our second method we performed an initial retrieval run using the same specifications as our baseline run. Ponte [9] adds additional query words to the original query based on the log ratio of the probability of occurrence in the model for relevant documents to the probability in the whole collection. We follow his approach and identified the top *n* significant MeSH terms of the top *m* retrieved documents for every topic, using the ILPS extension for Lucene [4]. We then added these to the MeSH field of our original query and performed another retrieval run using this expanded query. So a high early precision with our baseline run implied better results from our MeSH based feedback method. Based on the 2004 TREC Genomics data, IJzereef et al. [3] have shown that blind relevance feedback on MeSH terms led to an improvement of retrieval effectiveness.

3.3 MeSH lookup (Ml)

Currently, there are 22,997 headings in MeSH. The MeSH thesaurus itself consists of records containing individual descriptions of the MeSH concepts. These descriptions include not only synonyms, such as *Vitamin C* ***see*** *Ascorbic Acid*, but also scope notes, information about semantic types, previous indexing names, and so forth.

Each descriptive record for a MeSH term is essentially equivalent to a document about that term. Hence, we considered all the textual information about a MeSH term as a document, to which a topic can be compared. This was performed by indexing the contents of the MeSH thesaurus with Lucene. We tried to identify the MeSH terms that are most related to a topic by querying this index with the query terms extracted from the topic. When querying the index, we allowed for some fuzziness to account for spelling variances in terms. A maximal edit distance of 1 was found to be the optimal fuzziness setting, based on the training data. We then selected the top-ranked MeSH terms and these were subsequently added to the MeSH field of the original query.

4 Experiments

The retrieval performance of each of the three individual methods from the previous section was evaluated using the final topics of the TREC 2005 Genomics track. The results are shown in Table 2, with the results of the baseline run included as reference. The best scores are in bold-face.

	p10	MAP	recall
baseline	0.3755	0.2124	0.6658
Gene expansion (Ge)	**0.3939**	**0.2158**	0.6645
MeSH lookup (Ml)	0.0633	0.0286	0.1911
MeSH feedback (Fb)	0.3837	0.2023	**0.6852**

Table 2: Results of individual methods

Clearly, applying the MeSH lookup method seriously degraded retrieval performance. Based on earlier experiments using the training data, it proved, however, that applying the MeSH related query expansion methods resulted in a different set of correctly retrieved documents as compared to Gene name expansion. We therefore started evaluating the results of combinations of methods, using the CombSUM method to combine each pair of methods Fox and Shaw [1], Kamps and de Rijke [6]. The rationale was that doing so would boost the relevant documents that are found with either method. Precision would thus increase because relevant documents get higher ranks and recall would increase because more relevant documents would end up in the top 1000 retrieved documents.

We computed the best weight factor for every combination, based on the results of the training data. These experiments showed that combining Gene name expansion and MeSH based feedback (GeFb) and Gene name expansion and MeSH lookup (GeMl) delivered the best overall performance. With these findings in mind we devised our TREC submissions accordingly.

5 Runs

We submitted two runs for evaluation: `UAmscombGeFb` and `UAmscombGeMl`. For both runs the gene name expansion was applied as described in subsection 3.1. The submitted runs both use different forms of MeSH based query expansion. The weights by which the individual methods were combined differed as well, based on the results from evaluations with the training data.

`UAmscombGeFb`

- **Gene name expansion (weight 0.60).**
- **MeSH based feedback (weight 0.40).** Our feedback method has been applied to the baseline run as described in subsection 3.2. Experiments performed on the training data showed that a selection of 15 feedback MeSH terms based on the 10 top-ranking documents yielded optimal results.

`UAmscombGeMl`

- **Gene name expansion (weight 0.85).**
- **MeSH lookup (weight 0.15).** The five best matching MeSH terms were selected per topic and added to the original query as described in subsection 3.3.

6 Results

Table 3 gives an overview of the results for our submitted runs over the baseline.[1] The significance of the found results has been determined using Student's t-test.[2]

`UAmscombGeFb` gives statistically significant improvements for both recall and mean average precision as compared to the baseline. It also improves early precision, but to a lesser extent. The improvement in recall does not, contrary to common practice, adversively effect early precision. The performance of `UAmscombGeMl` is not as expected; the proposed method retrieves many non-relevant documents.

6.1 Topic analysis

The results from Table 3 can be broken down into the scores of the individual topics. A graphical representation can be found in Figures 1 and 2. As can be seen in these figures, `UAmscombGeFb` improves recall for more topics than

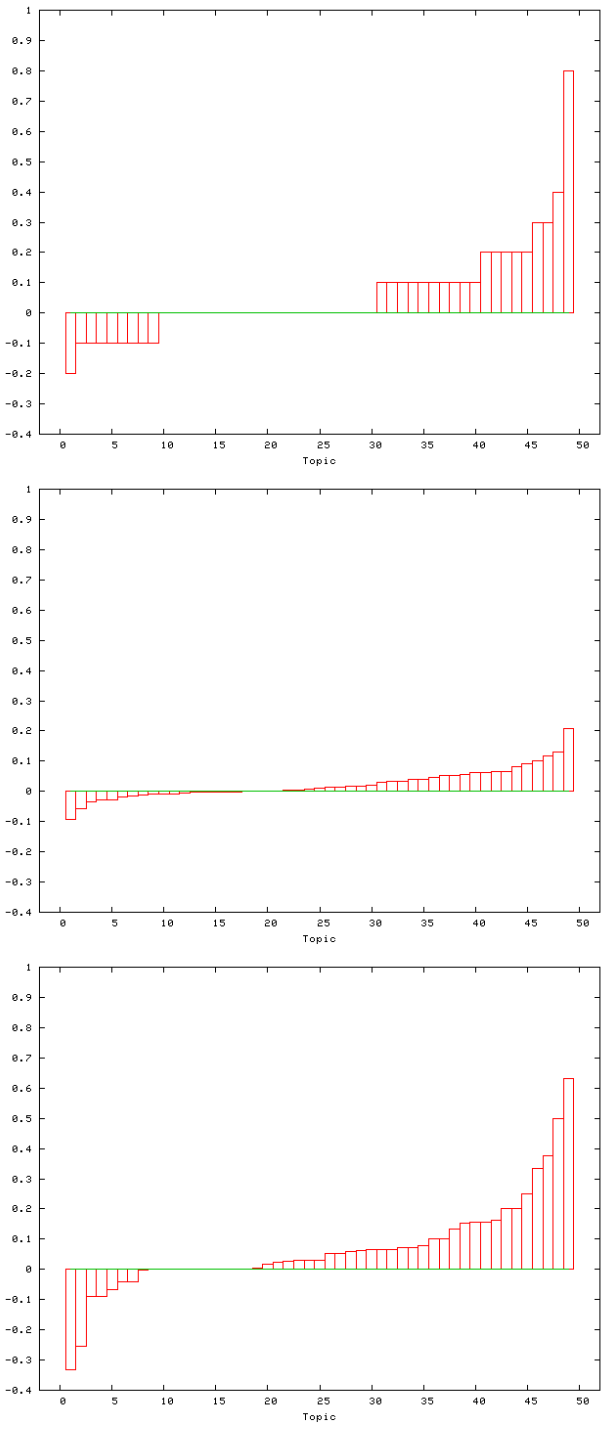

Figure 1: Per-topic breakdown of the effect of applying `GeFb` strategy, as compared to baseline: p10 (top), mean average precision (middle) and recall (bottom).

[1] Shortly after submitting these runs we discovered a flaw in the used term extractor. Due to this fact the results were slightly worse than could be expected when the proper tokenizer would have been used. For the remainder of this paper we will therefore be using the results of runs using the corrected term extractor instead of the actually submitted runs.

[2] There have been extensive discussions as to whether this particular test can be applied in this context, because of the assumption of normality of the distribution. However, recent work has shown that it in fact it is just as reliable as non-parametric tests [10].

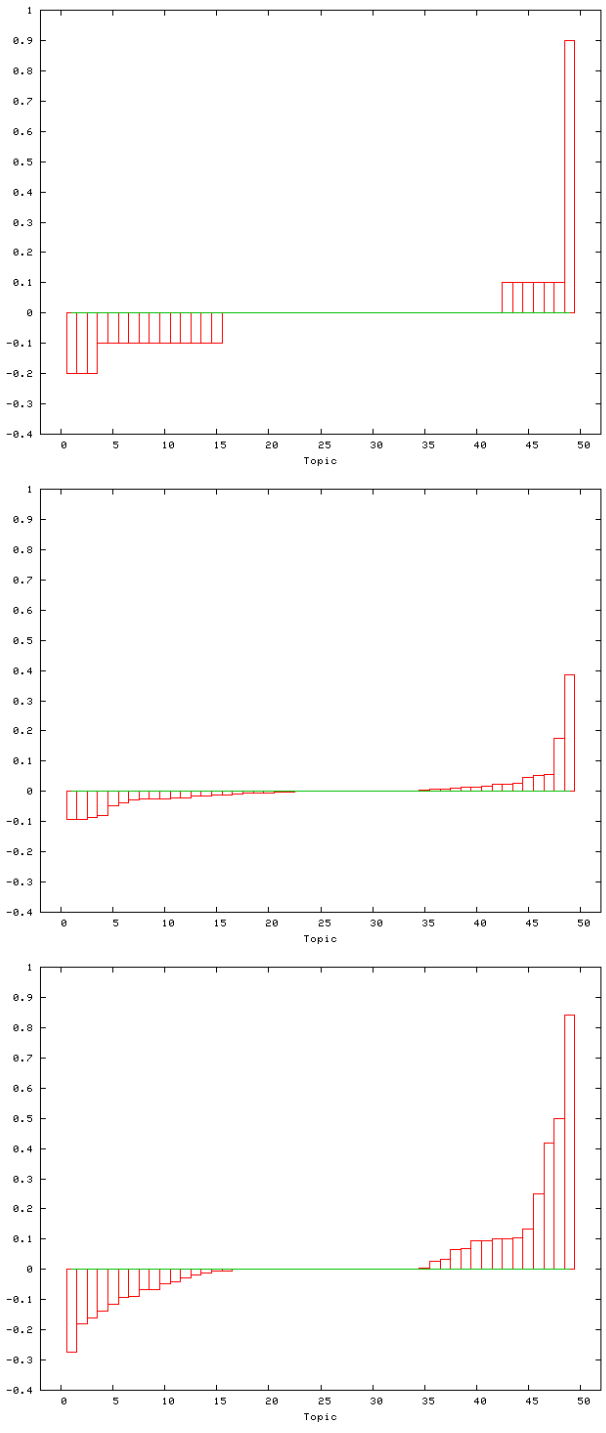

Figure 2: Per-topic breakdown of the effect of applying `GeMl` strategy, as compared to baseline: p10 (top), mean average precision (middle) and recall (bottom).

	p10	%Change
baseline	0.3755	
`UAmscombGeMl`	0.3694	-1.62%
`UAmscombGeFb`	**0.4327**	+15.22%*

	MAP	%Change
baseline	0.2124	
`UAmscombGeMl`	0.2164	+1.88%
`UAmscombGeFb`	**0.2430**	+14.30%**

	recall	%Change
baseline	0.6658	
`UAmscombGeMl`	0.6658	+0.00%
`UAmscombGeFb`	**0.7000**	+5.14%**

Table 3: Results of combinations of methods. Best scores are in bold-face. Significance *: $p < 0.05$, **: $p < 0.01$.

`UAmscombGeMl` when compared to the baseline. Our combined thesaurus-based approach does, in general, improve recall as well as mean average precision.

There are still some topics on which recall decreases, when compared to the baseline. These topics typically contain gene names for which incorrect acronyms or synonyms are stored. Another cause for a drop in recall are the occurrences of multi-term gene names, such as *Insulin receptor gene*. In our baseline run each of these terms is considered individually as query terms. During the application of our gene name expansion method, these separate terms are taken together and as such considered as a single, combined query term. This in turn leads to reduced recall.

On a few topics our baseline run outperforms `UAmscombGeFb` and `UAmscombGeMl` on both recall and mean average precision. The results for these topics are mainly influenced by an inaccurate gene name lookup. For example, the application of gene name expansion resulted in a drop in recall of 0.1818 on topic 134. This topic is shown below (in original as well as expanded form). As can be seen from this example, the term *conductance regulator protein* was included, but is not indicative of the information need of the topic. In biomedical research it is uncommon to speak about proteins when referring to the gene that encodes for them, thus the drop in performance. This is also reflected in the fact that this particular term does not appear in any of the relevant documents for this topic.

134 *Provide information about the genes CFTR and Sec61 in degradation of CFTR which leads to cystic fibrosis.*

134 *+((CFTR "cl" "cystic fibrosis gene"*
"conductance regulator"
"cystic fibrosis transmembrane conductance regulator"
"conductance regulator protein"
"cystic fibrosis transmembrane regulator"

"conductance regulator gene") (Sec61)) degradation of CFTR which leads to cystic fibrosis

There are some topics on which `UAmscombGeFb` performs better, whereas `UAmscombGeMl` performs worse than the baseline. These are topics which achieve high early precision with our baseline run, on which the blind feedback method is based. If there are relatively many relevant documents returned within the top ranked documents, the chance of a correct query expansion using the associated MeSH terms increases.

Finally, there are many topics which benefit from both the proposed strategies. The next example returned no relevant documents during our baseline run, as opposed to a recall of 0.6316 using `UAmscombGeFb`. This improvement can be attributed mostly to the accurate gene name expansion:

129 *Provide information on the role of the gene Interferon-beta in the process of viral entry into host cell.*

129 *+(Interferon-beta "beta-interferon" "fibroblast interferon" "interferon beta" "beta 1 interferon" "interferon beta1" "beta interferon" "beta-1 interferon" "interferon beta 1" "interferon-beta1" "ifn-beta" "fiblaferon" "interferon beta-1" "interferon fibroblast" "ifnbeta") viral entry into host cell*

7 Conclusions and future work

Our main focus while participating in this year's TREC Genomics has been to evaluate the integration of thesauri in the retrieval model. We posited that the use of a controlled vocabulary would help the system overcome synonymy and ambiguity issues and come closer towards the information need of an end-user. To this end we have developed three thesauri-based methods. One method uses automatically extracted synonym/acronym pairs from the corpus and MeSH thesaurus (`Ge`). The other two use the contents (`Fb`) and structure (`Ml`) of the assigned MeSH terms respectively.

Based on the provided training data we arrived at the conclusion that in fact combinations of methods work best. When applied individually, the proposed methods do not achieve significant improvements over our baseline run in terms of retrieval effectiveness. Each method was able to identify different relevant documents, given a single topic. We therefore submitted two runs based on a weighed combination of either `Ge` with `Fb` and `Ge` with `Ml`. A statistically significant improvement was measured when comparing the retrieval results of one of our submitted runs (`Ge+Fb`) with our baseline run. When examining the results of the individual topics, we found however that some topics benefited more from our proposed strategies than others.

We performed some additional experiments based on our language model. This model assumes that users select query terms that are very likely to be present in documents which would fulfill their information need. Using this model we attempted to form ideal queries; the best possible queries that users could pose to the system. The results are much lower than would be expected and it seems therefore that a language model might not be the best approach to retrieve MEDLINE abstracts.

Based on our interviews with biomedical researchers, we gained further insight in their search behavior and strategies. Besides using a strictly keyword-based search, they also use additional metadata. After an initial keyword-based retrieval run, they continue their search based on citations or authors of one or more top-ranked documents. These are all issues we intend to address during our participation in the TREC Genomics track next year.

Additionally, we would like to investigate the use of thesauri and/or ontologies within the retrieval model further. Reranking documents using a controlled vocabulary improves retrieval effectiveness in domain-specific collections, such as the Cross-Language Evaluation Form (CLEF) [5]. We believe this might be the same when applied within the biomedical domain.

8 Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project (`http://www.vl-e.nl`). This project 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).

Leig Azzopardi was supported by grants from the Netherlands Organization for Scientific Research (NWO) under project number 612.000.106. Jaap Kamps was supported by grants from NWO under project numbers 612.066.302 and 640.001.501. Maarten de Rijke was supported by grants from NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, and 640.001.501.

9 References

[1] E. A. Fox and J. A. Shaw. Combination of multiple searches. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215, 1994.

[2] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

[3] L. IJzereef, J. Kamps, and M. de Rijke. Biomedical retrieval: How can a thesaurus help? In R. Meersman and Z. Tari, editors, *CoopIS/DOA/ODBASE*, pages 1432–1448. LNCS 3761, 2005.

[4] ILPS. The ILPS extension of the Lucene search engine, 2005. `http://ilps.science.uva.nl/Resources/`.

[5] J. Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In S. McDonald and J. Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 283–295. Springer, 2004.

[6] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for european languages. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1073–1077, New York, NY, USA, 2004. ACM Press.

[7] W. Kraaij, M. Weeber, S. Raaijmakers, and R. Jelier. Mesh based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings of TREC 2004*. NIST, 2005.

[8] Lucene. The Lucene search engine, 2005. `http://jakarta.apache.org/lucene/`.

[9] J. M. Ponte. Language models for relevance feedback. In W. B. Croft, editor, *Advances in Information Retrieval*, The Kluwer International Series in Information Retrieval, chapter 3, pages 73–95. Kluwer Academic Publishers, Boston, 2000.

[10] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA, 2005. ACM Press.