Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary

Jaap Kamps

Language & Inference Technology Group
ILLC, University of Amsterdam
http://lit.science.uva.nl/

Abstract. There is a common availability of classification terms in online text collections and digital libraries, such as manually assigned keywords or key-phrases from a controlled vocabulary in scientific collections. Our goal is to explore the use of additional classification information for improving retrieval effectiveness. Earlier research explored the effect of adding classification terms to user queries, leading to little or no improvement. We explore a new feedback technique that reranks the set of initially retrieved documents based on the controlled vocabulary terms assigned to the documents. Since we do not want to rely on the availability of special dictionaries or thesauri, we compute the meaning of controlled vocabulary terms based on their occurrence in the collection. Our reranking strategy significantly improves retrieval effectiveness in domain-specific collections. Experimental evaluation is done on the German GIRT and French Amaryllis collections, using the test-suite of the Cross-Language Evaluation Forum (CLEF).

1 Introduction

Online text collections and digital libraries commonly provide additional classification information, such as controlled vocabulary terms in scientific collections. These classifications can be assigned either manually, or automatically [1]. The widespread use of additional classification terms prompts the question whether this additional information can be used to improve retrieval effectiveness. That is, when considering retrieval queries that do not use classification terms, can we make use of the fact that the retrieved documents have classification terms assigned to them? In IR parlance, this is a form of feedback.

Feedback or query expansion methods have a long history in information retrieval. This dates back, at least, to the studies of Sparck Jones [2,3] in which the collection is analyzed to provide a similarity thesaurus of word relationships. This type of approach is called *global* feedback in [4], which introduces a *local* feedback variant in which the initially retrieved documents are analyzed. There is mixed evidence on the effectiveness of global feedback. Local feedback methods are generally more effective, and the combination, by using global analysis techniques on the local document set, tends to be most effective [5].

An obvious feedback approach to exploiting classification information is to expand the original queries with (some of) the classification terms. This has received a fair amount of attention, especially in the medical domain where excellent resources exist. Srinivasan [6] investigates automatic query expansion with MeSH terms using the MEDLINE collection, based on a statistical thesaurus. Her finding is that query expansion with controlled vocabulary terms leads to improvement, but the effect is overshadowed by standard blind feedback. Hersh et al. [7] investigate various ways of expanding medical queries with UMLS Metathesaurus terms, and find a significant drop in retrieval effectiveness. Recently, French et al. [8] showed that a significant improvement of retrieval effectiveness is possible for query expansion with MeSH terms. However, they select the terms to be added by analyzing the set of human-judged, relevant documents. This gold standard experiment does not solve the problem of how to select the appropriate controlled vocabulary terms in the absence of full relevance information. Gey and Jiang [9] found a mild improvement of retrieval effectiveness when GIRT queries were expanded using thesaurus terms.

In sum, there is no equivocal evidence that fully automatically expanding queries with controlled vocabulary terms from initially retrieved documents leads to significant improvement of retrieval effectiveness. This motivated us to experiment with an alternative to expanding queries with classification terms. We explored a new feedback technique that reranks the set of initially retrieved documents based on the controlled vocabulary terms assigned to the documents. We use essentially a combination of global and local feedback techniques. On the one hand, we use a global feedback technique to analyze the usage of controlled vocabulary in the collections. The rationale for this is that we do not want to rely on the availability of special dictionaries or thesauri. Our approach is similar to latent semantic indexing [10]. We estimate the similarity of controlled vocabulary terms from their usage in the collections. Next, we apply dimensional reduction techniques, resulting in a low dimensional controlled vocabulary space. On the other hand, we use a local feedback technique for reranking the set of initially retrieved documents. Our strategy is to rerank the set of initially retrieved documents by their distance (based on the assigned controlled vocabulary terms) to the top-ranked retrieved documents.

The rest of this paper is structured as follows. Next, in section 2, we investigate the controlled vocabulary usage in scientific collections, and show how a similarity or distance measure can be used to obtain similarity vectors for the controlled vocabulary terms. Then, in section 3, we will provide some details of the experimental setup, and propose two document reranking strategies. In section 4, we investigate the results of the two reranking strategies, and their impact on retrieval effectiveness. Finally, in section 5, we discuss our results and draw some conclusions.

2 Controlled Vocabulary

The cross-language evaluation forum (CLEF [11]) addresses four different cross-lingual information retrieval tasks: monolingual, bilingual, multilingual, and domain-specific retrieval. For domain-specific retrieval at CLEF, the scientific collections of GIRT (German) and Amaryllis (French) are used. Table 1 gives some statistics about the test collections. Notice that the Amaryllis queries are

Table 1. Statistics about the GIRT and Amaryllis collections (stopwords are not included)

Collection	GIRT	Amaryllis
Documents	76,128	148,688
Size (Mb)	151	196
Words per Document	700	735
Queries	24	25
Words per Query	9.28	36.04
Relevant Documents per Query	40.0	80.7

long, due to the use of multi-sentence descriptions. Table 2 show one of the GIRT topics, and Table 3 shows one of the Amaryllis topics.

Table 2. GIRT topic 051 (only title and description fields)

⟨DE-title⟩ *Selbstbewusstsein von Mädchen* ⟨DE-desc⟩ *Finde Dokumente, die über den Verlust des Selbstbewusstseins junger Mädchen während der Pubertät berichten.*	⟨EN-title⟩ Self-confidence of girls ⟨EN-desc⟩ Find documents which report on the loss of self-confidence of young girls during the puberty.

Table 3. Amaryllis topic 001 (only title and description fields)

⟨FR-title⟩ *Impact sur l'environnement des moteurs diesel* ⟨FR-desc⟩ *Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution*	⟨EN-title⟩ The impact of diesel engine on environment ⟨EN-desc⟩ Air pollution by the exhaust of gas from diesel engines and methods of controlling air pollution. Pollutant emissions (NOX, SO2, CO, CO2, unburned product, ...) and air pollution control

The GIRT collection contains (abstracts of) documents from German social science literature published between 1978 and 1996 [12]. The documents are also classified by controlled vocabulary terms assigned by human indexers, using the controlled-vocabulary thesaurus maintained by GESIS [13]. The average number of controlled vocabulary terms in a document is 9.91. Table 4 gives some of the characteristics of controlled vocabulary in the GIRT and the Amaryllis collections. The Amaryllis collection contains (abstracts of) documents in

Table 4. Controlled Vocabulary Usage in the GIRT and Amaryllis collections

	GIRT	Amaryllis
Used terms	6,745	125,360
Occurrences	755,333 (704 doubles)	1,599,653 (562 doubles)
Most frequent	29,561 *Bundesrepublik Deutschland*	20,514 *Homme*
	9,246 *Frau*	17,283 *France*
	6,133 *historische Entwicklung*	7,888 *Traitement*
	4,736 *Entwicklung*	6,619 *Etude expérimentale*
	4,451 *neue Bundesländer*	5,987 *Etude cas*
	3,645 *DDR*	4,319 *Diagnostic*
	3,445 *Österreich*	4,179 *Modélisation*
	3,341 *Entwicklungsland*	4,171 *Enfant*
	3,025 *Betrieb*	4,130 *Etude comparative*
	3,012 *geschlechtsspezifische Faktoren*	3,954 *Article synthèse*

French from various scientific fields. The average number of manually assigned controlled vocabulary terms in a document is 10.75.

We want to compute the similarity of controlled vocabulary terms based on their occurrence in the collection. Our working hypothesis is that controlled vocabulary terms that are frequently assigned to the same documents, will have similar meaning. We only give an outline of the used approach here, since we apply well-known techniques. For the convenience of interested readers, a detailed description is provided in Appendix A. We determine the number of occurrences of controlled vocabulary terms and of co-occurrences of pairs of controlled vocabulary terms use in the collection, and use these to define a distance metric over the controlled vocabulary terms. Specifically, we use the Jaccard similarity coefficient on the log of (co)occurrences, and use 1 minus the Jaccard score as a distance metric [14]. For creating manageable size vectors for each of the controlled vocabulary terms, we reduce the matrix using metric multi-dimensional scaling techniques [15]. For all calculations we used the best approximation of the distance matrix on 100 dimensions. This results in a 100-dimensional vector for each of the 6,745 controlled vocabulary terms occurring in the GIRT collection. The Amaryllis collection uses a much richer set of 125,360 controlled vocabulary terms. We select only the controlled vocabulary terms occurring at least 25 times in the collection. Thus, we end up with a 100-dimensional vector for the 10,274 most frequent controlled vocabulary terms in the Amaryllis collection. Basically,

we now have a vector space for the controlled vocabulary terms, where related terms will be at a relatively short distance, and unrelated terms far apart.

Note that we only have vectors for the controlled vocabulary terms. However, there are straightforward ways to map documents and topics into the controlled vocabulary space. For each document we collect the assigned controlled vocabulary terms from the collection. We have a vector in the controlled vocabulary space for each of the controlled vocabulary terms. We define the vector for the document to be simply the mean score for each of the controlled vocabulary term vectors. We can also create vectors for topics, based on which documents are retrieved by an information retrieval system (here, we use the 10 best ranked documents). For each topic we consider the top-ranked documents, and define the vector for the topic to be the weighted mean score of the document vectors. We give each document a weight corresponding to its retrieval status value (RSV).

3 Experimental Setup

All retrieval experiments were carried out with the FlexIR system developed at the University of Amsterdam [16]. The main goal underlying FlexIR's design is to facilitate flexible experimentation with a wide variety of retrieval components and techniques. FlexIR is implemented in Perl and supports many types of preprocessing, scoring, indexing, and retrieval tools. One of the retrieval models underlying FlexIR is the standard vector space model. All our runs use the `Lnu.ltc` weighting scheme [17] to compute the similarity between a query and a document. For the experiments on which we report in this paper, we fixed the slope at 0.2; the pivot was set to the average number of unique words per document.

For the GIRT and Amaryllis collections, we index both the free-text of the documents, i.e., the title and body or abstract of articles, as well as the manually assigned controlled vocabulary terms. Our index contains the words as they occur in the collection with only limited sanitizing, i.e., we remove punctuation; apply case-folding; map marked characters to the unmarked tokens; and remove stopwords. We employ generic lists with stopwords, with 155 stopwords for French, and 231 stopwords for German. We do not apply further morphological normalization; see [18] for an overview of the effectiveness of stemming and n-gramming for monolingual retrieval in German and French.

From the indexes we obtain a baseline run per collection. For our reranking experiments, we use the controlled vocabulary space to rerank the documents initially retrieved in the baseline run. For all the retrieved documents, we extract the assigned controlled vocabulary terms from the collection. Then, we calculate document vectors for all the documents, by calculating the mean of the vectors for controlled vocabulary terms assigned to them. Finally, we calculate a vector for the topic, by calculating the weighted mean of the 10 top-ranked documents. Based on the topic and document vectors, we experiment with two reranking feedback strategies.

Naive reranking We have a vector for each of the topics, and for each of the retrieved documents. Thus, ignoring the RSV of the retrieved documents, we can simply rerank all documents by increasing euclidean distance between the document and topic vectors. Since RSVs should be decreasing by rank, we use 1 minus the distance as the new RSV.

Combined reranking We investigate a more conservative reranking by combining the two sources of evidence available: the original text-based similarity scores of the baseline run, and the controlled vocabulary-based distances which are calculated as in the naive reranking. The scores were combined in the following manner. Following Lee [19], both scores are normalized using $RSV_i' = \frac{RSV_i - min_i}{max_i - min_i}$. We assigned new weights to the documents using the summation function used by Fox & Shaw [20]: $RSV_{new} = RSV_1' + RSV_2'$. This combination results in a less radical reranking of documents.

Since we are interested in the interaction of our reranking feedback with standard blind feedback, we do three sets of experiments. In the first set we evaluate our reranking feedback using the original queries. In the second set of experiments, we apply standard blind feedback to expand the original queries with related terms from the free-text of the documents. Term weights were recomputed using the standard Rocchio method [21], where we considered the top 10 documents to be relevant and the bottom 500 documents to be non-relevant. We allowed at most 20 terms to be added to the original query. In the third set of experiments, we investigate the effectiveness of the reranking feedback using the expanded queries from the second set of experiments.

Finally, to determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method, a non-parametric inference test [22,23]. The method has previously been applied to retrieval evaluation by, e.g., Wilbur [24] and Savoy [25]. We take 100,000 resamples, and look for significant improvements (one-tailed) at significance levels of 0.95 (*), 0.99 (**) and 0.999 (***).

4 Experimental Results

4.1 Reranking feedback

In the first set of experiments, we study the effectiveness of our new reranking feedback method using the original queries. We create baseline runs using the indexes of the free-text and controlled vocabulary terms of the documents. We use the title and description fields of the CLEF 2002 topics. The results are shown in Table 5. The resulting baseline run for GIRT has a mean average precision (MAP) of 0.2063. The resulting baseline run for Amaryllis has a MAP of 0.2778. Next, we employ the naive reranking strategy to the respective baseline runs (as described in Section 3). The result of the naive reranking is negative: we find a decrease in performance for both GIRT and Amaryllis. The drop in performance for Amaryllis is even significant. Does this mean that the calculated topic vector

Table 5. Mean average precision scores for the baseline runs, the naive rerank runs, and the combined rerank runs, using CLEF 2002 topics. Best scores are in boldface, significance $^{\star} = p < .05$, $^{\star\star} = p < .01$, $^{\star\star\star} = p < .001$

	GIRT		Amaryllis	
Run	MAP	% Change	MAP	% Change
Baseline	0.2063		0.2778	
Naive rerank	0.1973	-4.4%	0.1829	-34.2%***
Combined rerank	**0.2487**	+20.6%***	**0.3059**	+10.1%**

is not adequately representing the content of the topics? Or is it a result of our radical reranking approach?

We investigate this by employing the combined rerank strategy that takes both the text-based similarity score, as well as the distance to the topic vector into account (as described in Section 3). The results of the combined rerank runs are also shown in Table 5: for GIRT the MAP improves to 0.2487, and for Amaryllis the MAP improves to 0.3059. The respective improvements are +20.6% (GIRT) and +10.1% (Amaryllis). The improvement of both combined rerank runs is statistically significant. Thus we find evidence that the combined rerank strategy is significantly improving retrieval effectiveness.

4.2 Rocchio blind feedback

In a second set of experiments, we study the effectiveness of standard blind feedback. A possible explanation of the observed improvement due to reranking feedback is that it functions roughly like standard blind feedback. The results of applying Rocchio blind feedback (as described in Section 3) are shown in Table 6. We see that Rocchio blind feedback is promoting retrieval effective-

Table 6. Mean average precision scores for the baseline runs and the Rocchio blind feedback runs using CLEF 2002 topics. Best scores are in boldface, significance $^{\star} = p < .05$, $^{\star\star} = p < .01$, $^{\star\star\star} = p < .001$

	GIRT		Amaryllis	
Run	MAP	% Change	MAP	% Change
Baseline	0.2063		0.2778	
Blind feedback	**0.2209**	+7.1%	**0.2986**	+7.5%

ness. The resulting blind feedback runs for GIRT have a MAP of 0.2209 (an improvement of +7.1% over the unexpanded queries). The resulting blind feedback runs for Amaryllis have a MAP of 0.2986 (an improvement of +7.5%). Blind feedback is improving retrieval effectiveness, although the improvements are not significant. Note also that improvement due to blind feedback is less than the improvement due to combined reranking as discussed in our first set of

experiments. Thus, when comparing the relative effectiveness of both types of feedback, the reranking feedback meets and exceeds the effectiveness of standard Rocchio blind feedback.

4.3 Rocchio blind feedback plus reranking feedback

In the third set of experiments, we investigate whether the improvement of retrieval effectiveness we found in the first set of experiments is supplementary to the effects of standard blind feedback we found in the second set of experiments. That is, the difference with the first set of experiments is that we now use queries that have been expanded by Rocchio blind feedback. The results are shown in Table 7, note that we now compare the improvement relative to the expanded queries, and not relative to the earlier baseline run. The results of the

Table 7. Mean average precision scores for the Rocchio blind feedback runs, the naive rerank runs, and the combined rerank runs, using CLEF 2002 topics. Best scores are in boldface, significance $^{\star} = p < .05$, $^{\star\star} = p < .01$, $^{\star\star\star} = p < .001$

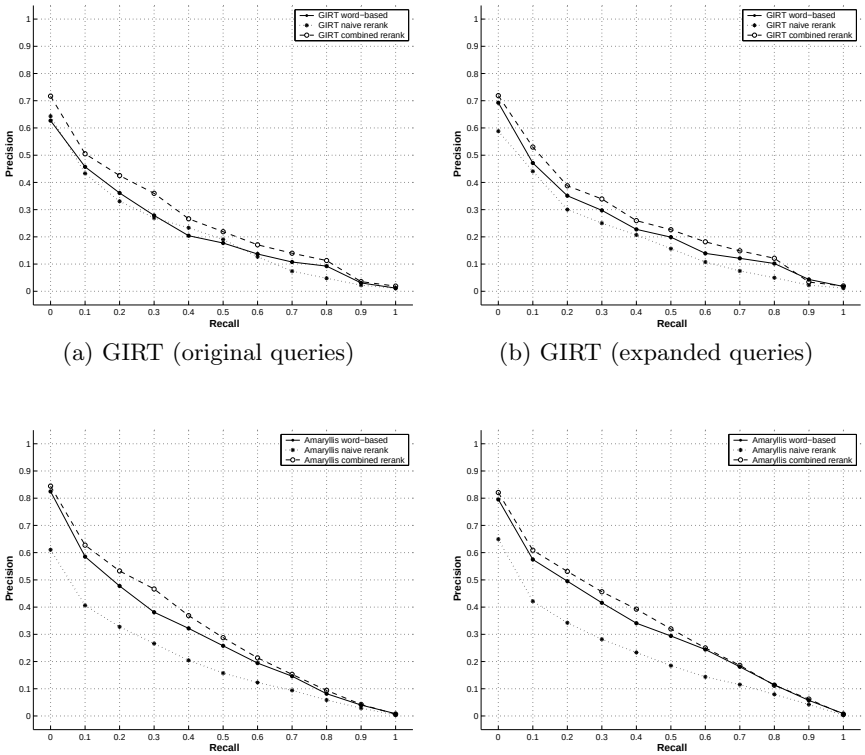
	GIRT		Amaryllis	
Run	MAP	% Change	MAP	% Change
Blind feedback	0.2209		0.2986	
Naive rerank	0.1831	-17.1%	0.2025	-32.2%***
Combined rerank	**0.2481**	+12.3%*	**0.3197**	+7.1%*

naive reranking strategy are no better than in the first set of experiments: both runs show a drop in performance, and the decrease in performance is significant for Amaryllis. The combined rerank strategy turns out to be effective again. For GIRT the MAP is 0.2481 (+12.3% over the expanded queries) and for Amaryllis the MAP is 0.3197 (+7.1%). Both improvements are statistically significant. So, we find evidence that the combined rerank strategy is significantly improving retrieval effectiveness, on top of the effect due to blind feedback.

When comparing the combined reranking scores with those obtained in the first set of experiments, we notice the following. The combined reranking score for Amaryllis expanded queries is 4.5% higher than the score for the unexpanded queries. However, the combined reranking score for the expanded GIRT queries is 0.2% lower than the score for the unexpanded queries. Thus in this case, the use Rocchio blind feedback is hurting the score, possibly due to topic drift influencing the retrieved top 10 documents for some of the topics.[1]

Figure 1 plots the recall-precision curves for the reranking feedback experiments we conducted. We have shown that the combined reranking strategy leads to a significant improvement of retrieval effectiveness. This also shows that

[1] Recent evidence suggest that Rocchio feedback is promoting overall performance, but hurts performance on the poorly performing topics [26].



(a) GIRT (original queries)

(b) GIRT (expanded queries)

(c) Amaryllis (original queries)

(d) Amaryllis (expanded queries)

Fig. 1. Interpolated recall-precision averages for the naive rerank runs and the combined rerank runs on the original and expanded queries using CLEF 2002 topics

the topic vector can be used to capture the content of the topic. In turn, this demonstrates the viability of our approach to derive the meaning of controlled vocabulary terms from their occurrence in the collection.

5 Discussion and Conclusions

This paper introduced a new feedback technique that reranks the set of initially retrieved documents based on the controlled vocabulary terms assigned to the documents. Our reranking strategy significantly improved retrieval effectiveness in domain-specific collections, above and beyond the use of standard Rocchio blind feedback.

Our method is specifically tailored for collections that provide additional classification of documents, such as manually assigned controlled vocabulary terms in scientific collections. We derived a controlled vocabulary thesaurus based on their (co)occurrences in the collections. Similar approaches have been proposed since the advent of information retrieval. For example, Sparck Jones [3] discusses the clustering of words based on their co-occurrence. The dimensional reduction techniques we used are similar to those used in latent semantic indexing [10]. Our focus was on the classication terms in the collection, although the same techniques can be applied to all, or a selection of, words in the collection. Gauch et al. [27,28] use a corpus analysis approach for query expansion. Schütze and Pedersen [29] use a cooccurrence-based thesaurus to derive context vectors for query words. Our approach differs from earlier work by its focus on the reranking of the initially retrieved document, based on the controlled vocabulary terms assigned to the documents. The queries only play a role in the retrieval of the initial set of documents. Perhaps closest in spirit is the work of Jin et al. [30], proposing a language model that takes classification labels into account.

Experimental evaluation was done on the German GIRT and French Amaryllis collections, using the test-suite of the Cross-Language Evaluation Forum [11]. We experimented with two reranking strategies. The first strategy, a naive ranking based solely on the distances, generally showed a drop in performance. The second strategy, a combined reranking using evidence from both the text-based relevance score and the controlled vocabulary-based distances, showed a significant improvement of retrieval effectiveness. To investigate how the improvement due to reranking relates to standard blind feedback, we conducted further experiments and showed that reranking feedback is more effective than Rocchio blind feedback. Moreover, we can apply reranking to expanded queries leading, again, to a significant improvement. For one of the collections, however, the score for combined reranking feedback is lower for the expanded queries than for the original queries. Thus, were in earlier research the gain due to query expansion with controlled vocabulary was overshadowed by the gain due to standard blind feedback, we here see that reranking feedback is overschadowing the gain due to Rocchio feedback.

There are obvious differences between standard blind feedback and reranking feedback, for example, an important effect of query expansion is the retrieval

of additional relevant documents, i.e., an improvement of recall, whereas a re-ranking strategy can only improve the ranking of relevant documents, i.e., an improvement of precision. Further experimentation is needed to fully assess the relative impact of both feedback methods, and to uncover the underlying mechanisms responsible for their effectiveness. This should take into account similar results in interactive retrieval, where relevance feedback tends to produce more accurate results than query reformulation [31].

6 Acknowledgments

This research is supported by the Netherlands Organization for Scientific Research (NWO, grants 400-20-036 and 612.066.302). Many thanks to Maarten de Rijke and Börkur Sigurbjörnsson for comments and suggestions.

A Appendix

We analyzed the keyword space using multi-dimensional scaling techniques [15]. The first step is to compute dissimilarities for the controlled vocabulary terms.

A natural candidate for measuring the similarity of the controlled vocabulary terms is the *Jaccard* coefficient. Let $|i|$ denote the number of document having controlled vocabulary term i. For each pair of controlled vocabulary terms i and j, we determine

$$\mathsf{J}(i,j) \;=\; \frac{|i \cap j|}{|i \cup j|} \;=\; \frac{|i \cap j|}{|i| + |j| - |i \cap j|}.$$

Note that for i we have that $\mathsf{J}(i,i) = 1$ and for disjoint i and j we have $\mathsf{J}(i,j) = 0$. From the Jaccard similarity coefficient, we can make a dissimilarity coefficient by considering $\mathsf{d}_1(i,j) = (1 - \mathsf{J}(i,j))$ or $\mathsf{d}_2(i,j) = \sqrt{(1 - \mathsf{J}(i,j))}$. These dissimilarity coefficients have the following desirable properties, d_1 is metric and d_2 is both metric and euclidean [14].

The Jaccard scores for the collections give values close to 0 for almost all pairs of controlled vocabulary terms. To allow for greater variation, we use the logarithm of the values, thus we determine the distance between two controlled vocabulary terms i and j as

$$\mathsf{Dist}(i,j) \;=\; 1 - \frac{\log_{10}(|i \cap j|)}{\log_{10}(|i \cup j|)} \;=\; 1 - \frac{\log_{10}(|i \cap j|)}{\log_{10}(|i| + |j| - |i \cap j|)}.$$

This, again, gives a value in the range $[0, 1]$, a value 1 for terms not appearing in the same document, a value 0 for terms only occurring in the same documents.

The distance Dist is a metric, i.e, it gives a non-negative number such that

1. $\mathsf{Dist}(i,j) = 0$ if and only if $i = j$,
2. $\mathsf{Dist}(i,j) = \mathsf{Dist}(j,i)$, and
3. $\mathsf{Dist}(i,j) + \mathsf{Dist}(j,k) \geq \mathsf{Dist}(i,k)$.

The third (triangle) inequality will hold due to the fact that all values for distinct i and j are above 0.5.

Based on the above, we can now construct a squared matrix of dissimilarities $\{\mathsf{Dist}(i,j)\}$, of size 6,745 by 6,745 in case of GIRT and of size 10,274 by 10,274 in

case of Amaryllis. Our aim is to find a set of points in a lower dimensional space such that each of these points represents one of the controlled vocabulary terms, and that the euclidean distances between points approximate the original dissimilarities as well as possible.

For this, we follow the standard procedure of metric multi-dimensional scaling [15, pp.22–39]. From the dissimilarities, we obtain a matrix $\mathbf{A}$ of elements $-\frac{1}{2}(\mathsf{Dist}(i,j))^2$. Next, we obtain the double-centered matrix $\mathbf{B}$, build from $\mathbf{A}$ by subtracting row and column mean, and adding matrix mean.

Then spectral decomposition gives

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ the diagonal matrix of eigenvalues and $\mathbf{V} = (v_1, \ldots, v_n)$ the matrix of corresponding eigenvectors. We assume that the eigenvalues are ordered such that $\lambda_i \geq \lambda_{i+1}$, and that the eigenvectors have unit length.

Following [10], we choose to look at the first 100 eigenvalues $\mathbf{\Lambda}_{100} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{100})$ and associated eigenvectors $\mathbf{V}_{100} = (v_1, \ldots, v_{100})$. The best approximation of $\mathbf{B}$ on 100 dimensions is matrix $\mathbf{X}_{100}$ such that

$$\mathbf{X}_{100} = \mathbf{V}_{100}\mathbf{\Lambda}_{100}^{1/2}$$

The resulting matrix has dimensions 6,745 by 100 in case of GIRT, and 10,274 by 100 in case of Amaryllis. For each controlled vocabulary term, we now have a vector of length 100.

References

1. Lewis, D.D.: An evaluation of phrasal and clustered representations on a text categorization task. In Belkin, N., Ingwersen, P., Pejtersen, A.M., Fox, E., eds.: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York NY, USA (1992) 37–50
2. Sparck Jones, K., Needham, R.: Automatic term classification and retrieval. Information Processing & Management **4** (1968) 91–100
3. Sparck Jones, K.: Automatic Keyword Classification for Information Retrieval. Butterworth, London (1971)
4. Attar, R., Fraenkel, A.S.: Local feedback in full-text retrieval systems. Journal of the Association of Computing Machinery **24** (1977) 397–417
5. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In Frei, H.P., Harman, D., Schabie, P., Wilkinson, R., eds.: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York NY, USA (1996) 4–11
6. Srinivasan, P.: Query expansion and MEDLINE. Information Processing & Management **34** (1996) 431–443
7. Hersh, W., Price, S., Donohoe, L.: Assessing thesaurus-based query expansion using the UMLS metathesaurus. In: Proceedings of the 2000 AMIA Annual Fall Symposium. (2000) 344–348
8. French, J.C., Powell, A.L., Gey, F., Perelman, N.: Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. In: Proceedings of the tenth International Conference on Information and Knowledge Management, ACM Press (2001) 199–206

9. Gey, F.C., Jiang, H.: English-German cross-language retrieval for the GIRT collection–exploiting a multilingual thesaurus. In: Proceedings of the Eighth Text REtrieval Conference (TREC-8), National Institute for Standards and Technology, Washington, DC (1999)
10. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41** (1990) 391–407
11. CLEF: Cross language evaluation forum (2003) `http://www.clef-campaign.org/`.
12. Kluck, M., Gey, F.C.: The domain-specific task of CLEF - specific evaluation strategies in cross-language information retrieval. In Peters, C., ed.: Cross-Language Information Retrieval and Evaluation, CLEF 2000. Volume 2069 of Lecture Notes in Computer Science., Springer (2001) 48–56
13. Schott, H., ed.: Thesaurus Sozialwissenschaften. Informationszentrum Sozialwissenschaften, Bonn (2002) 2 Bände: Alphabetischer und systematischer Teil.
14. Gower, J.C., Legendre, P.: Metric and euclidean properties of dissimilarity coefficients. Journal of Classification **3** (1986) 5–48
15. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling. Chapman & Hall, London UK (1994)
16. Monz, C., de Rijke, M.: Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001. Volume 2406 of Lecture Notes in Computer Science., Springer (2002) 262–277
17. Buckley, C., Singhal, A., Mitra, M.: New retrieval approaches using SMART: TREC 4. In Harman, D.K., ed.: The Fourth Text REtrieval Conference (TREC-4), National Institute for Standards and Technology. NIST Special Publication 500-236 (1996) 25–48
18. Hollink, V., Kamps, J., Monz, C., de Rijke, M.: Monolingual document retrieval for European languages. Information Retrieval **7** (2004) 33–52
19. Lee, J.H.: Combining multiple evidence from different properties of weighting schemes. In Fox, E.A., Ingwersen, P., Fidel, R., eds.: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York NY, USA (1995) 180–188
20. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In Harman, D.K., ed.: The Second Text REtrieval Conference (TREC-2), National Institute for Standards and Technology. NIST Special Publication 500-215 (1994) 243–252
21. Rocchio, Jr., J.J.: Relevance feedback in information retrieval. In Salton, G., ed.: The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs NJ (1971) 313–323
22. Efron, B.: Bootstrap methods: Another look at the jackknife. Annals of Statistics **7** (1979) 1–26
23. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)
24. Wilbur, J.: Non-parametric significance tests of retrieval performance comparisons. Journal of Information Science **20** (1994) 270–284
25. Savoy, J.: Statistical inference in retrieval effectiveness evaluation. Information Processing and Management **33** (1997) 495–512
26. Jijkoun, V., Kamps, J., Mishne, G., Monz, C., de Rijke, M., Schlobach, S., Tsur, O.: The University of Amsterdam at TREC 2003. In: TREC 2003 Working Notes, National Institute for Standards and Technology (2003)

27. Gauch, S., Wang, J.: A corpus analysis approach for automatic query expansion. In: Proceedings of the Sixth International Conference on Information and Knowledge Management, ACM Press (1997) 278–284
28. Gauch, S., Wang, J., Rachakonda, S.M.: A corpus analysis approach for automatic query expansion and its extension to multiple databases. ACM Transactions on Information Systems (TOIS) **17** (1999) 250–269
29. Schütze, H., Pedersen, J.O.: A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing & Management **3** (1997) 307–318
30. Jin, R., Si, L., Hauptman, A.G., Callan, J.: Language model for IR using collection information. In Järvelin, K., Beaulieu, M., Baeza-Yates, R., Myaeng, S.H., eds.: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York NY, USA (2002) 419–420
31. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. Information Processing & Management **36** (2000) 95–108