Improving Document Retrieval by Automatic Query Expansion Using Collaborative Learning of Term-Based Concepts

Stefan Klink, Armin Hust, Markus Junker, and Andreas Dengel

German Research Center for Artificial Intelligence (DFKI, GmbH)
P.O. Box 2080, 67608 Kaiserslautern, Germany
{Stefan.Klink, Armin.Hust, Markus.Junker,
Andreas.Dengel}@dfki.de
http://www.dfki.de/~klink

Abstract. Query expansion methods have been studied for a long time – with debatable success in many instances. In this paper, a new approach is presented based on using term concepts learned by other queries. Two important issues with query expansion are addressed: the selection and the weighing of additional search terms. In contrast to other methods, the regarded query is expanded by adding those terms which are most similar to the *concept* of individual query terms, rather than selecting terms that are similar to the complete query or that are directly similar to the query terms. Experiments have shown that this kind of query expansion results in notable improvements of the retrieval effectiveness if measured the recall/precision in comparison to the standard vector space model and to the pseudo relevance feedback. This approach can be used to improve the retrieval of documents in Digital Libraries, in Document Management Systems, in the WWW etc.

1 Introduction

As the Internet and Digital Libraries become more and more popular, the growing number of documents has raised the problem called information overload. Typical search engines index billions of pages across a variety of categories, and return results ranked by expected topical relevance. But only a small percentage of these pages may be of a specific interest.

In Information Retrieval (IR) the number of retrieved documents is related to the number of appropriate search terms. Retrieval with short queries is typical in Web search [13], but it is much harder as compared to retrieval with long queries. This is because shorter queries often provide less information for retrieval. Modern IR systems therefore integrate thesaurus browsers to find additional search terms [24].

However, the aim of the retrieval activity is not to retrieve a large number of documents. Rather, users are interested in a high usefulness of the retrieved documents.

Another problem which is typical for the Web and for Digital Libraries is that the terminology used in defining queries is often different to the terminology used in the

representing documents. Even if some users have the same information need they rarely use the same terminology in their queries. Many intelligent retrieval approaches [5, 18, 23] have tried to bridge this terminological gap.

Research on automatic query expansion (or modification) was already under way before the 60's when initial requests were enlarged in the grounds of statistical evidence [30]. The idea was to obtain additional relevant documents through expanded queries based on the co-occurrence of the terms. However, this kind of automatic query expansion has not been very successful. The retrieval effectiveness of the expanded queries was often not greater than, or even less than the effectiveness of the original queries [21, 22, 31].

One idea involves the use of a relevance feedback environment where the system retrieves documents that may be relevant to a user's query. The user judges the relevance of one or more of the retrieved documents and these judgments are fed back to the system to improve the initial search result. This cycle of relevance feedback can be iterated until the user is satisfied with the retrieved documents. In this case, we can say that the more feedback is given to the system the better is the search effectiveness of the system. This behavior is verified by [4]. He has shown that the recall/precision effectiveness is proportional to the log of the number of relevant feedback documents.

But in a traditional relevance feedback environment the user voted documents are appropriate to the complete query. That means that the complete query is adapted to the users needs. If another user has the same intention but uses a different terminology or just one word more or less in his query then the traditional feedback environment doesn't recognize any similarities in these situations.

Another idea to solve the terminology problem is to use query concepts. The system called 'Rule Based Information Retrieval by Computer' (RUBIC) [1, 5, 18] uses production rules to capture user query concepts. In RUBIC, a set of related production rules is represented as an AND/OR tree, called a rule base tree. RUBIC allows the definition of detailed queries starting at a conceptual level. The retrieval output is determined by fuzzy evaluation of the AND/OR tree. To find proper weight values, Kim and Raghavan developed a neural network (NN) model in which the weights for the rules can be adjusted by users' relevance feedback. Their approach is different from the previous NN approaches for IR in two aspects [12, 14]. First, they handle relations between concepts and Boolean expressions in which weighted terms are involved. Second, they do not use their own network model but an already proven model in terms of its performance.

But the crucial problem of a rule-based system still exists: the automatic production of proper rules and the learning of appropriate structures of rules, not just the weights.

2 Query Expansion

The crucial point in query expansion is the question: Which terms (or phrases) should be included in the query formulation? If the query formulation is to be expanded by

additional terms there are two problems that are to be solved, namely how are these terms selected and how are the parameters estimated for these terms.

Many terms used in human communication are ambiguous or have several meanings [20]. But in most cases these ambiguities are resolved automatically without noticing the ambiguity. The way this is done by humans is still an open problem of psychological research, but it is almost certain, that the context in which a term occurs plays a central role.

Most attempts at automatically expanding queries failed to improve the retrieval effectiveness and it was often concluded that automatic query expansion based on statistical data was unable to improve the retrieval effectiveness substantial [22].

But this could have several reasons. Term-based query expansion approaches are mostly using hand-made thesauri or just plain co-occurrence data. They often do not use learning technologies for the query terms. On the other hand, those who use learning technologies (Neural Networks, Support Vector Machines, etc.) are query-based. That means these systems learn concepts (or additional terms) for the complete query.

In contrast to learning complete queries, the vital advantage of using term-based concepts is that other users can profit from learned concepts even if the same query is never used before. A statistical evaluation of log files has shown that the probability that a searcher uses exactly the same query than a previous searcher is much lower then the probability that parts of the query (phrases or terms) occurs in previous queries. So, even if a searcher never used the given search term, the probability that other searchers had used it is very high and then he can profit from the learned concept.

3 Traditional Document Retrieval

The task of traditional document retrieval is to retrieve documents which are relevant to a given query from a fixed set of documents, i.e. a document database. In a common way to deal with documents as well as queries, they are represented using a set of index terms (simply called terms) by ignoring their positions in documents and queries. Terms are determined based on words of documents in the database, usually during pre-processing phases where some normalization procedures are incorporated (e.g. stemming and stop-word elimination).

In the following, t_i ($1 \le i \le M$) and $\boldsymbol{d}_j$ ($1 \le j \le N$) represent a term and a document in the database, respectively, where M is the number of terms and N is the number of documents.

3.1 Vector Space Model

The most popular and the simplest retrieval model is the vector space model (VSM) [5]. In the VSM, a document $\boldsymbol{d}_j$ is represented as a M dimensional vector

$$\boldsymbol{d}_j = (w_{1j}, \ldots, w_{Mj})^T \quad (1)$$

where T indicates the transpose, w_{ij} is a weight of a term t_i in a document $\boldsymbol{d}_j$. A query is likewise represented as

$$\boldsymbol{q}_k = (w_{1k}, \ldots, w_{ik}, \ldots, w_{Mk})^T, \quad 1 \le k \le L \tag{2}$$

where w_{ik} is a weight of a term t_i in a query $\boldsymbol{q}_k$.

These weights are computed by the standard normalized tf · idf weighting scheme [27] as follows:

$$w_{ij} = tf_{ij} * idf_i \tag{3}$$

where tf_{ij} is the weight calculated using the term frequency f_{ij} and idf_i is the weight calculated using the inverse of the document frequency.

The result of the retrieval is represented as a list of documents ranked according to their similarity to the query. The similarity $sim(\boldsymbol{d}_j, \boldsymbol{q}_k)$ between a document $\boldsymbol{d}_j$ and a query $\boldsymbol{q}_k$ is measured by the standard cosine of the angle between $\boldsymbol{d}_j$ and $\boldsymbol{q}_k$:

$$sim(\boldsymbol{d}_j, \boldsymbol{q}_k) = \frac{\boldsymbol{d}_j^T \boldsymbol{q}_k}{\|\boldsymbol{d}_j\| \ \|\boldsymbol{q}_k\|} \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm of a vector.

3.2 Pseudo Relevance Feedback

A well-known method to obtain the terms for a query expansion is the pseudo relevance feedback [18]. Here, in a first step, documents are ranked with an original query, like in the VSM. Then, the highly ranked documents are assumed to be relevant and their terms are incorporated into original query. In the second step, the documents are ranked again by using the new expanded query.

In this paper, we employ (like in Kise et al. [15]) a simple variant of the pseudo relevance feedback:

Let $\mathbb{E}$ be a set of document vectors for expansion given by

$$\mathbb{E} = \left\{ \boldsymbol{d}_j^+ \;\middle|\; \frac{\mathrm{sim}(\boldsymbol{d}_j^+, \boldsymbol{q})}{\max_i \mathrm{sim}(\boldsymbol{d}_i, \boldsymbol{q})} \ge \theta \right\} \tag{5}$$

where $\boldsymbol{q}$ is an original query vector and τ is a threshold of the similarity. The sum $\boldsymbol{d}_s$ of the document vectors in

$$\boldsymbol{d}_s = \sum_{\boldsymbol{d}_j^+ \in \mathbb{E}} \boldsymbol{d}_j^+ \tag{6}$$

can be considered as enriched information about the original query. Then, the expanded query vector $\boldsymbol{q}'$ is obtained by

$$\boldsymbol{q}' = \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} + \alpha \frac{\boldsymbol{d}_s}{\|\boldsymbol{d}_s\|} \tag{7}$$

where α is a parameter for controlling the weight of the newly incorporated component. Finally, the documents are ranked again according to the similarity $sim(\boldsymbol{d}_j, \boldsymbol{q}')$ to the expanded query.

4 Learning Term-based Concepts

A problem of the standard VSM is that a query is often too short to rank documents appropriately. To cope with this problem, our approach is to enrich the original query by expanding it with terms occurring in the documents of the collection. But in contrast to traditional pseudo relevance feedback methods, where the top *i* ranked documents are assumed to be relevant and then all their terms are incorporated into the expanded query, a different technique is used to compute the relevant documents as follows:

Let $\boldsymbol{q} = t_1 \ldots t_n$ be the user query containing the terms $t_1 \ldots t_n$
and $\boldsymbol{q} = (w_1, \ldots, w_i, \ldots, w_M)^T$ be the vector representation of this query.
Let $\boldsymbol{Q} = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m\}$ be the set of all previous queries $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$
and $\boldsymbol{D}_k^+$ be the set of relevant documents of the query $\boldsymbol{q}_k$.

The goal is now to learn for each term t_i a concept $\boldsymbol{c}_i$ ($1 \le i \le n$) with the help of previous queries and their appropriate relevant documents. For this, the term t_i is searched in all previous queries and if it is found, the relevant documents of these queries are used to learn the concept.

Due to the VSM, a concept is also a weighted vector of terms and calculated with:

$$\boldsymbol{c}_i = \tau_i (0,\ldots,w_i,\ldots,0)^T + \delta_i \sum_{t_i \in \boldsymbol{q}_k} \boldsymbol{D}_k^+ \tag{8}$$

where $0 \le \tau_i, \delta_i \le 1$ are weights for the original term and the additional terms, respectively.

The expanded query vector is obtained by the sum of all term-based concepts:

$$\boldsymbol{q}' = \sum_{i=1}^{n} \boldsymbol{c}_i \tag{9}$$

Before applying the expanded query, it is normalized by

$$\boldsymbol{q}'' = \frac{\boldsymbol{q}'}{\|\boldsymbol{q}'\|} \tag{10}$$

For this approach, the complete documents (e.g. all term weights of the document vector) are summed up and added to the query. Although, in some papers it is reported that using just the top ranked terms is sufficient or sometimes better, experiments with this approach on our collections have shown that the more words are used to learn the concepts the better are the results. So, the decision was made to use always all terms of the documents and not only some (top ranked) terms.

If no ground truth of relevant documents is available, relevance feedback techniques can be used to get the ground truth. Then, concepts are learned by adding terms from retrieved relevant documents.

5 Experiments and Results

Section 5.1 describes the test collections, section 5.2 describes our evaluation methods, and section 5.3 presents the results

5.1 Test Collections

For our comparison we used four standard test collections: CACM (collection of titles and abstracts from the journal 'Communications of the ACM'), CR (congressional report), FR (federal register) and NPL (also known as the VASWANI). These collections are contained in the TREC disks [9]. All collections are provided with queries and their ground truth (a list of documents relevant to each query). For these collections, terms used for document representation were obtained by stemming and eliminating stop words.

Table 1 lists statistics about the collections after stemming and eliminated stop words. In addition to the number of documents, a difference among the collections is the document length: CACM and NPL consists of abstracts, while CR and FR contain much longer documents.

Queries in the TREC collections are mostly provided in a structured format with several fields. In this paper, the "title" (the shortest representation) is used for the CR and NPL collection whereas the "desc" (description; medium length) is used for the CACM and FR collection.

Table 1. Statistics about collections used for experiments

	CACM	CR	FR	NPL
# documents	3204	27922	19789	11429
# queries	52	34	112	93
# different terms	3029	45717	50866	4415
avg doc length [terms]	25.8	672.8	863.7	21.8
avg query length [terms]	10.9	3.1	9.8	6.6

5.2 Evaluation

The following paragraphs describe some basic evaluation methods used in this paper. For further information and a more detailed description see Kise et al [15].

5.2.1 Average Precision

A common way to evaluate the performance of retrieval methods is to compute the (interpolated) precision at some recall levels. This results in a number of recall/precision points which are displayed in recall/precision graphs [5]. However, it is sometimes convenient for us to have a single value that summarizes the performance. The average precision (non-interpolated) over all relevant documents [5, 7] is a measure resulting in a single value. The definition is as follows:

As described in section 3, the result of retrieval is represented as the ranked list of documents. Let *r(i)* be the rank of the *i*-th relevant document counted from the top of the list. The precision for this document is calculated by *i/r(i)*. The precision values for all documents relevant to a query are averaged to obtain a single value for the query. The average precision over all relevant documents is then obtained by averaging the respective values over all queries.

5.2.2 Statistical Test

The next step for the evaluation is to compare the values of the average precision obtained by different methods [15]. An important question here is whether the difference in the average precision is really meaningful or just by chance. In order to make such a distinction, it is necessary to apply a statistical test.

Several statistical tests have been applied to the task of information retrieval [11,33]. In this paper, we utilize the test called "macro t-test" [33] (called paired t-test in [11]). The following is a summary of the test described in [15]:

Let a_i and b_i be the scores (e.g., the average precision) of retrieval methods A and B for a query i and define $d_i = a_i - b_i$. The test can be applied under the assumptions that the model is additive, i.e., $d_i = \mu + \varepsilon_i$ where μ is the population mean and ε_i is an error, and that the errors are normally distributed. The null hypothesis here is $\mu = 0$ (A performs equivalently to B in terms of the average precision), and the alternative hypothesis is $\mu > 0$ (A performs better than B).

It is known that the Student's t-statistic

$$t = \frac{\bar{d}}{\sqrt{s^2 / n}} \qquad (\mathbf{11})$$

follows the t-distribution with the degree of freedom of $n - 1$, where n is the number of samples (queries), $\bar{d}$ and s^2 are the sample mean and the variance:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i , \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2 \qquad (\mathbf{12})$$

By looking up the value of t in the t-distribution, we can obtain the P-value, i.e., the probability of observing the sample results d_i ($1 \le i \le N$) under the assumption that the null hypothesis is true. The P-value is compared to a predetermined significance level σ in order to decide whether the null hypothesis should be rejected or not. As significance levels, we utilize 0.05 and 0.01.

5.3 Results and Comparison to the Standard

In this section the results of the experiments are presented. Results were evaluated using the average precision over all queries. Recall/precision graphs were generated and then significance tests were applied to the results.

5.3.1 Recall and Precision

As described above, term weights in both documents and queries are determined according to the normalized tf·idf weighting scheme and the similarity is calculated by the VSM cosine measure, see also formula (3).

The results of the pseudo relevance feedback are depending on the two parameters β (weight) and θ (similarity threshold). To get the best results, we were varying β from 0 to 5.0 with step 0.1 and θ from 0.0 to 1.0 with step 0.05. For each collection, best individual β and θ are calculated such that their average precision is highest:

Table 2. Best values for pseudo relevance feedback parameters

	CACM	CR	FR	NPL
β (weight)	1.70	1.50	0.30	2.00
θ (sim. threshold)	0.35	0.75	0.00	0.45

The results for the concept-based queries are calculated as follows: For each individual query, all concepts of the terms within this query are learned by using all other queries with the help of their relevant documents (leave-one-out test) and the expanded query is used to calculate the new recall/precision result. Of course, the relevant documents of the current query are not used to learn the concepts.

The results of our concept-based expansion are also depending on weights. For the experiments described in this paper we just used the default value: $\tau_i = \delta_i = 1$.

Figure 1 shows the recall/precision results of the original query with the standard vector space model (red line, VSM), the pseudo relevance feedback (blue line, PRF) and the expanded query using learned concepts (green line, concepts).

The recall/precision graphs in figure 1 indicate that the automatic query expansion method based on learned concepts yields a considerable improvement in the retrieval effectiveness in 3 collections over all recall points compared to the standard vector space model and to the pseudo relevance feedback method (except with the NPL collection). There is no indication that the improvement is depending on the size of the collection, the number of documents nor on the number or size of the queries. The method performs good on CACM but only somewhat better than the VSM on the NPL. On FR it performs better than on the CR collection. Looking at the figures the impression could arise that our method performs better with longer queries. But experiments with the CR collection have shown that 'title' queries result a better precision than 'description' or 'narrative' queries. This behavior is in contrast to the first impression of the figures.

Additionally, as described above, the more words and the more documents are used to learn the concept the better are the results. Experiments have shown that the precision continues to increase as more documents are used.

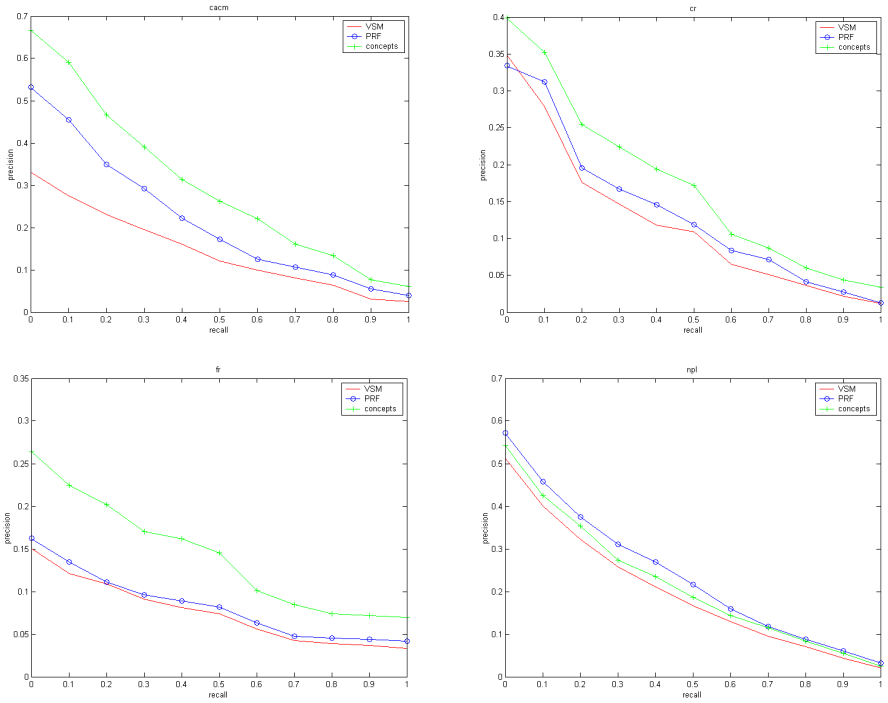

Fig. 1. recall / precision of CACM, CR, FR, and NPL

5.3.2 Statistical tests

The results above are very exciting. But in order to be sure that these results are really meaningful and not just by chance, it is necessary to apply a statistical test. As described above, we used the "macro t-test". The results of the macro t-test for all pairs of methods are shown in table 3. The meaning of the symbols such as "", ">" and "~" is summarized at the bottom of the table. For example, the symbol "<" was obtained in the case of the concept method compared to the VSM for the NPL collection. This indicates that, at the significance level $\sigma = 0.05$, the null hypothesis "concept method performs equivalently to the VSM" is rejected and the alternative hypothesis "concept method performs better than the VSM" is accepted. (At $\sigma = 0.01$, however, the null hypothesis cannot be rejected.)

Roughly speaking, "A $\gg$($\ll$) B", "A > (<) B" and "A~B" indicate that "A is almost guaranteed to be better (worse) than B", "A is likely to be better (worse) than B" and "A is equivalent to B", respectively.

The macro t-tests confirm our results. Our new method for expanding queries which is based on term-based concepts outperforms the standard VSM and the pseudo-relevance feedback (except on the NPL collection).

Due to the low P-values in the '$\gg$' cases it is proved that the results are not obtained by chance and that the improvements are significant.

Table 3. Results of the macro t-test.

methods (A vs. B)	CACM	CR	FR	NPL
PRF vs. VSM	≫	~	~	≫
concepts vs. PRF	≫	>	≫	<
concepts vs. VSM	≫	≫	≫	>

≫, ≪ : P-value ≤ 0.01
>, < : 0.01 ≤ P-value ≤ 0.05
~ : 0.05 ≤ P-value

6 Conclusions and Future Work

We have described an approach for bridging the terminology gap between user queries and potential answer documents using term-based concepts. In this approach for all query terms concepts are learned from previous queries and relevant answer documents. A new query is transformed by replacing the original query terms by the learned concepts. This is in contrast to traditional query expansion methods which do not take into account previous queries but mostly rely on term statistics of the underlying document collection or hand-made thesauri.

Our experiments made on four standard test collections with different sizes and different document types have shown considerable improvements vs. the original queries in the standard vector space model and vs. the pseudo relevance feedback (here, except on the NPL collection). This is true even through the parameters for pseudo relevance feedback were optimized, while for our approach we did not touch the parameters. The improvements seem not to depend on the type nor on the size of the collection and they are not obtained by chance.

The vital advantage of the approach is that in a multi-user scenario users can benefit from concepts learned by other users. The approach is expected to perform better as more users and queries are involved. It can be used to improve the retrieval of documents in Digital Libraries, Document Management Systems, WWW etc.

In the near future it is planned to make some experiments on the influence of the weights τ_i and δ_i (cmp. equation (**8**)), and to develop functions for calculating these parameters for each individual concept. Further experiments are planned using user-voted relevance feedback instead of collection-given ground-truth to test the performance on 'real-life' data. For this we are currently collecting queries and click data from a search engine [10].

The approach on passage-based retrieval by Kise [15] has shown good improvements vs. LSI and Density Distribution. Instead of using the complete relevant documents for expanding the query or using the n top ranked terms, an interesting idea for the future is to use just terms of relevant passages within the documents. This should increase the quality of the expanded queries.

7 Acknowledgements

This work was supported by the German Ministry for Education and Research, bmb+f (Grant: 01 IN 902 B8).

References

1. Aalbersberg I.J.: Incremental relevance feedback. In *Proceedings of the Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11 - 22, 1992
2. Allan J.: Incremental relevance feedback for information filtering. In *Proceedings of the 19th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 270 - 278, 1996
3. Alsaffar A.H., Deogun J.S., Raghavan V.V., Sever H: Concept-based retrieval with minimal term sets. In Z.W. Ras and A. Skowon, editors, *Foundation of Intelligent Systems: 11th Int. Symposium*, ISMIS'99, pp. 114-122, Springer, Warsaw, Poland, June 1999
4. Buckley C., Salton G., Allen J.: The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 292 - 300, 1994
5. Baeza-Yates R., Ribeiro-Neto B.: *Modern Information Retrieval*. Addison-Wesley Pub. Co., 1999. ISBN 020139829X
6. Croft W.B.: Approaches to intelligent information retrieval. *Information Processing and Management*, 1987, Vol.23, No.4, pp. 249-254
7. ftp://ftp.cs.cornell.edu/pub/smart/
8. Harman D.: Towards Interactive Query Expansion. In: Chiaramella Y. (editor): *11th International Conference on Research and Development in Information Retrieval*, pp. 321 – 331, Grenoble, France, 1988
9. http://trec.nist.gov/
10. http://phibot.org/
11. Hull D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329 - 338, 1993
12. Iwayama M.: Relevance Feedback with a Small Number of Relevance Judgments: Incremental Relevance Feedback vs. Document Clustering. In *Proceedings of the 23rd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 10 - 16, Athens, Greece, July 2000
13. Jansen B.J., Spink A., Bateman J. and Saracevic T.: Real Life Information Retrieval: A Study of User Queries on the Web, In *SIGIR Forum*, Vol. 31, pp. 5-17, 1988
14. Kim M., Raghavan V.: Adaptive concept-based Retrieval Using a Neural Network, In *Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, Athens, Greece, July 2000
15. Kise K., Junker M., Dengel A., Matsumoto K.: Passage-Based Document Retrieval as a Tool for Text Mining with User's Information Needs, In *Proceedings of the 4th International Conference of Discovery Science*, pp. 155-169, Washington, DC, USA, November 2001

16. Kwok K.: Query Modification and Expansion in a Network with Adaptive Architecture. In *Proceedings of the 14th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192 - 201, 1991
17. Lu F., Johnsten Th., Raghavan V.V., Traylor D.: Enhancing Internet Search Engines to Achieve Concept-based Retrieval, In *Proceedings of Inforum'99*, Oakridge, USA
18. Manning C.D. and Schütze H.: *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
19. Maglano V., Beaulieu M., Robertson S., : Evaluation of interfaces for IRS: modeling end-user search behaviour. *20th Colloquium on Information Retrieval*, Grenoble, 1988
20. McCune B.P., Tong R.M., Dean J.S., Shapiro D.G.: RUBRIC: A System for Rule-Based Information Retrieval, In *IEEE Transaction on Software Engineering*, Vol. SE-11, No.9, September 1985
21. Minker J., Wilson, G.A. Zimmerman, B.H.: An evaluation of query expansion by the addition of clustered terms for a document retrieval system, *Information Storage and Retrieval*, vol. 8(6), pp. 329-348, 1972
22. Peat H.J., Willet, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems, *Journal of the ASIS*, vol. 42(5), pp. 378-383, 1991
23. Pirkola A.: Studies on Linguistic Problems and Methods in Text Retrieval: The Effects of Anaphor and Ellipsis Resolution in Proximity Searching, and Translation and query Structuring Methods in Cross-Language Retrieval, *PhD dissertation*, Department of Information Studies, University of Tampere. Acta Universitatis Tamperensis 672. ISBN 951-44-4582-1; ISSN 1455-1616. June 1999
24. Qiu Y.: ISIR: an integrated system for information retrieval, In *Proceedings of 14th IR Colloqium*, British Computer Society, Lancaster, 1992
25. van Rijsbergen C.J., Harper D.H., etal.: The Selection of Good Search Terms. *Information Processing and Management* 17, pp. 77-91, 1981
26. Resnik P.: Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th Int. Joint Conference on Artificial Intelligence*, pp. 448-453, 1995
27. Salton G., Buckley C.: Term weighting approaches in automatic text retrieval. Information *Processing & Management* 24(5), pp. 513 - 523, 1988
28. Salton G., Buckley C.: Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science* 41(4), pp. 288 – 297, 1990
29. Sanderson M., Croft B.: Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 206 - 213, Berkeley, CA, August 1999
30. Sparck-Jones K.: Notes and references on early classification work. In *SIGIR Forum*, vol. 25(1), pp. 10-17, 1991
31. Smeaton A.F., van Rijsbergen C.J.: The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, vol. 26(3), pp. 239 – 246, 1983
32. Stucky D.,: Unterstützung der Anfrageformulierung bei Internet-Suchmaschinen durch User Relevance Feedback, *diploma thesis*, German Research Center of Artificial Intelligence (DFKI), Kaiserslautern, November 2000
33. Yang Y. and Liu X.: A Re-Examination of Text Categorization Methods. In *Proceedings of the 22nd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42 - 49, Berkeley, CA, August 1999