Corpus-Based Query Expansion in Online Public Access Catalogs

Jeffry Komarjaya, Danny C. C. Poo, and Min-Yen Kan

School of Computing, National University of Singapore,
3 Science Drive 2, Singapore 117543,
jeffry.komarjaya@nus.edu.sg,{dpoo,kanmy}@comp.nus.edu.sg

Abstract. We propose a probabilistic method for query expansion in online public access catalogs that utilizes both historical query logs and the subject headings in the library catalog. Our method creates correlations between query and document terms, allowing relevant subject headings from the corpus to be retrieved and added to a query. Experiments demonstrate an average of 31.1% performance increase over currently fielded baselines.

1 Introduction

The problem of vocabulary mismatch is a deep-rooted problem in information retrieval as users often use different or too few words to describe the concepts in their queries as compared to the words that authors use to describe the concepts in their documents [1–3]. Despite this, many library online public access catalogs (OPACs), such as the INNOPAC system[1] used by *L*ibrary *IN*tegrated *C*atalogue (LINC) at the National University of Singapore, still depend on keyword matching to determine the relevant documents for queries.

Short queries damage retrieval effectiveness in two ways: 1) they lead too many results and 2) the queries themselves are ambiguous. The first phenomenon, often called information overload, makes searching difficult as users are overwhelmed with information. In our case study of LINC, from March to September 2003, queries sent to LINC have a mean length of 2.815 words. For example, for the top query "java", LINC returned 32,000 books, of which 953 books had 100% relevance, leaving the user to select between 953 alternatives. Short queries are often polysemous (having multiple senses or meanings, as in "java": the computer language or the island in Indonesia). Such queries result in ambiguity as words that could disambiguate them are missing.

Query expansion, the process of expanding a user's query with additional related words and phrases, has been suggested to address the problem. However, finding and using appropriate related words remains an open problem. Research on query expansion has focused on intranet or internet web search. However, the typical digital library OPAC contains bibliographic records which are far more structured than documents on the internet. On the other hand, traditional OPAC

[1] http://www.libdex.com/vendor/Innovative_Interfaces,_Inc.html

To appear at ECDL 2004

research has largely focused on rule-based systems that do not take advantage of corpora.

Our study melds the two approaches by analyzing library corpora for use in query expansion in the digital library OPAC. Our system combines both historical query logs and the library catalog to create a thesaurus-based query expansion that correlates query terms with document terms. Our process consists of three steps. First, historical query logs are analyzed to uncover frequent queries. These queries are sent to the OPAC to extract relevant subject headings from the top documents. In the last step, the system calculates probabilistic correlations between the retrieved subjects heading and users' queries. With these correlations, relevant subject headings can be selected from the corpus for new, unseen queries.

In Section 2, we discuss related work in query expansion. we detail the methods used to build the thesaurus (a matrix correlating query keywords and subject headings) by sending queries to the OPAC and analyzing the results, and describe how query expansion is done with the built thesaurus. Section 5 describes our case study in which we deployed our approach in our local OPAC. We conclude with a summary and directions of further work.

2 Query Expansion Techniques

In query expansion, there are two key aspects: the source of expansion terms and the method to weight and integrate expansion terms [4, 3, 5]. Existing techniques can be classified as global, local or external, based on the source of terms. Global techniques require corpus-wide statistics such as the occurrence of expansion terms in the corpus and the source of expansion terms is usually the whole corpus. Local techniques analyze a number of top-ranked documents retrieved by a query to expand it. In contrast, external techniques depend on external resources such as domain-independent thesauri for expansion terms.

2.1 Global Techniques

We define a global technique as one that analyzes the contents of a particular corpus to identify semantically similar terms. By gathering statistics of the co-occurrences of terms in the corpus, global techniques build statistical term relationships which can then be used to expand queries. Some global techniques are term clustering [6], global similarity thesauri [7], latent semantic indexing [8] and Phrasefinding [9]. Since global techniques focus only on the document and do not take into account the query, global techniques only offer a partial solution to the word mismatching problem [4]. Global techniques typically require co-occurrence information for every pair of terms. However, most global techniques compute this information offline, removing a potential computational bottleneck.

2.2 Local Techniques

As compared to global techniques, local techniques use only a subset of the documents retrieved by a query. Local techniques can be divided into two main categories: interactive (*i.e.*, relevance feedback) and automatic (*i.e.*, local feedback).

Relevance Feedback In relevance feedback systems, related terms come from user-identified relevant documents. Relevance feedback was originally designed to be used with the vector space model [10]. However, relevance feedback has also been incorporated into Boolean retrieval models [11] and probabilistic retrieval models [12, 13]. Other methods such as incremental relevance feedback [14] have also been proposed, which analyzes previous queries and relevance judgments in the same session to improve search effectiveness. [15] proposed Adaptive Relevance Feedback (ARF) on top of incremental relevance feedback to detect changes in users' information goals. However, in a real search context, users are usually reluctant to provide any type of feedback [4, 16].

Local Feedback Local feedback uses the top-ranked documents retrieved by a query as a viable source of information. The basic assumption is that the top-ranked documents retrieved are relevant, and thus the words in the top-ranked documents themselves can be used to expand the query. While the performance of local feedback can be erratic, it has shown good performance in Text REtrieval Conferences (TREC) experiments [17]. The TREC test collections are often used to evaluate query expansion techniques [1–3, 5].

Many improvements have been suggested to local feedback, such as using Boolean filters and proximity constraints to refine the set of top-ranked documents [16], exploiting potentially relevant documents from past similar queries [18] and using information theory in weighting and selecting expansion terms [1]. The idea of local context analysis was also proposed [3, 5] which combines the idea of global and local techniques to select expansion terms based on co-occurrences with the query terms within the top-ranked documents. More recently, historical user logs were also used to deduce likely user interactions [4].

2.3 External Techniques

External techniques make use of external resources, such as online thesauri which are not tailored for any particular collection, for query expansion. In past work, general reference resources such as Longman's Dictionary of Contemporary English (LDOCE) and WordNet [19] have been used. Because of the ambiguity of terms and the existence of specialized terms for certain collections, these thesauri might be difficult to use. Voorhees reported improvements of 1% for longer and less ambiguous queries but expanding shorter queries actually degraded performance. Based on these neutral results, we have decided not to pursue the use of general, external resources in our research.

3 Building a Learning Thesaurus

The Online Catalog Evaluation Projects reported that library patrons have problems matching their terms with those indexed in the online catalog and do not understand the printed LCSH (Library of Congress Subject Headings) [20]. To solve this vocabulary problem, one possible solution is to map query terms to the underlying vocabulary of the corpus by building a corpus-specific thesaurus.

To build such a thesaurus for an OPAC, frequently-occurring queries were first harvested from historical query log kept by the OPAC. Each of these queries were re-sent to the OPAC, generating a ranked list of relevant documents. We adopted local feedback to extract the top relevant documents, and extracted the subject headings for each document. We then mapped the query keywords to the frequency of the subject headings in the relevant documents.

There are two design details that are important in our system's architecture. First, we used local feedback rather than standard relevance feedback, as it requires no explicit relevance judgments or click-streams because it is fully automated and requires no user effort [15]. Second, we used frequently-occurring queries in our historical query logs to build our thesaurus for the initial queries. Subsequent new queries submitted by users need to also be analyzed and the thesaurus updated to reflect the change in query patterns.

3.1 Subject Headings

Subject headings are usually assigned by expert cataloguers. These headings are used to index the documents in OPACs. Standardized, controlled vocabulary terms or subject headings are usually employed, such as the Library of Congress Subject Headings (LCSH), the Library of Medicine's Medical Science Subject Headings (MeSH), or the Dewey Decimal Classification (DDC).

Compared to book titles, subject headings are more objective and precise. For example, subject headings "Genetics", "Evolution (Biology)" and "Behavior genetics" are clearer than the title "The Selfish Gene". Thus we feel that it would be less ambiguous to use the subject headings as expansion terms in comparison to book titles. Using subject headings also provides us with knowledge from experts, which are less prone to errors, and eliminates the need to use automatic term weighting algorithms, such as Term Frequency $\times$ Inverse Document Frequency (TF$\times$IDF), to extract terms from the corpus.

3.2 Correlating Query Terms and Document Terms

In this study, we attempt to create links between keyword query terms and the subject headings documented in OPACs by the librarians and cataloguers. Our key observation is that if queries containing a certain term often lead to the selection of documents containing another term, then we consider that there is a strong relationship between these two terms [4].

We assumed that subject headings from the top documents retrieved using queries containing a particular keyword were related to that keyword. For example, the terms, "macromedia" and "flash", in the query "macromedia flash" are regarded as relevant to the documents that the query retrieved, *e.g.*, "Macromedia Flash MX developer's guide", and the subject headings of these documents, "Computer animation". By acquiring and analyzing a large pool of queries and collecting their top-ranked documents, we are able to form associations between the queries and documents. These associations allow us to create a thesaurus that will aid in query augmentation by mapping keywords in the queries to the subject headings. Figure 1 shows how the subject headings are related to the query keywords.

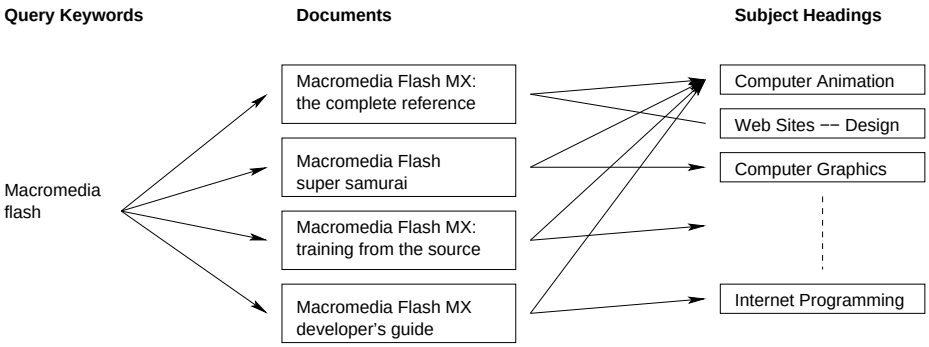

Fig. 1. Correlations between query keywords and subject headings.

4 A Corpus-based Query Expansion Model

4.1 Relation of a document term to the entire query

Xu and Croft [5] and Qiu and Frei [7] hypothesized that relevant terms tend to co-occur with all query terms in the top-ranked documents. A similar idea is applied in this study; subject headings that occur with all or most of the query keywords in the thesaurus are considered more relevant than subject headings that only occur with a few keywords. In other words, we should consider a term that is similar to the query concept rather than one that is only similar to a single term in the query.

Consider the queries Q1: "Java" and Q2: "Java Indonesia". While Q1 is ambiguous and could mean the programming language developed by Sun Microsystems, the Indonesian capital island or the coffee bean, Q2 is much less ambiguous and more likely to refer to the Indonesian capital island than to the other meanings. Subject headings that co-occur with both "Java" and "Indonesia" are likely to be relevant to Q2 and should be given a higher weight than

terms that only occur with "Java" or "Indonesia". Terms should be selected based on their similarity to the entire query instead of just a few query terms. In contrast, many query expansion techniques add a term even when the term is only strongly related to just one of the query terms [7], resulting in suboptimal performance. As a result, our approach prefers subject headings that co-occur with more query terms over those co-occurring with fewer query terms.

To determine the correlation between a query term w_q and a subject heading w_d, we calculate the degree of co-occurrence of w_d and w_q. That is, we need to calculate $co\_degree(w_d, w_q)$ [5]. We estimate $co\_degree(w_d, w_q)$ as the likelihood that w_d and w_q co-occurs non-randomly in the top-retrieved documents retrieved by queries containing w_q by using an adapted equation from the normalized TF*IDF weighting scheme [21]:

$$co\_degree(w_d, w_q) \equiv f(w_d) \times ln \frac{m}{df(w_d)} \tag{1}$$

where $f(w_d)$ is the frequency which the subject heading w_d co-occurs with the query keyword w_q in the thesaurus, m is the total number of distinct query keywords and $df(w_d)$ is the total number of distinct query keywords that co-occur with w_d in the thesaurus. A higher $f(w_d)$ will indicate that the subject heading w_d is more important over another subject heading with a lower $f(w_d)$. On the other hand, the higher $df(w_d)$ is, the more likely that w_d co-occurs with the query keyword w_q by chance or that w_d might be ambiguous because it is related to many different keywords.

The above calculates the relevance of the subject headings with individual query terms. We also need to measure the relationship of the subject heading with regards to the entire query. Many researchers [4, 7, 5] have proposed methods to measure the degree of co-occurrence with all query terms. We measure the relationship of a term to the entire query using the following cohesion weight calculation [5]:

$$g(w_d, w_q) \equiv \prod_{all\ query\ terms} (\delta + co\_degree(w_d, w_q)) \tag{2}$$

where, δ is a smoothing factor to assign a small, non-zero probability to subjects that only co-occur with only one query term that would otherwise receive a weight of zero. With a small δ, subject headings that co-occur with all query terms are ranked higher and with a large δ, subject headings having significant co-occurrences with individual query terms are ranked higher [5]. As we prefer subject headings that co-occur with more query terms over those co-occurring with fewer, we set the smoothing factor δ to a small value of 0.001.

After the cohesion weights of the subject headings related to the query have been calculated and the weights normalized, subject headings for query expansion have to be selected. We select subject headings which have weights above the threshold β. The default value of is 0.03. The new query Q' will be reformulated by adding these subject terms into the original query. Q' will then be used to retrieve documents.

5 Experimental Evaluations

In this section, we describe the methodology and data collection of the experiment before illustrating our experimental findings.

5.1 Evaluation Methodology

The objective of the experiments reported in this section is to test whether query expansion using the thesaurus we have constructed can be used to improve the retrieval effectiveness compared to the original (unexpanded) queries. Our local OPAC, an INNOPAC-based system, allows the user to select the method for sorting the results. The two most common methods are to sort by relevance or date; *i.e.* either the most relevant documents or the most recent documents are ranked higher. Thus, we compare our query expansion results using both date and relevance sorting methods.

IR performance are usually assessed using standard metrics of precision and recall. However, boolean retrieval (used in many OPACs) with or without query expansion retrieves the same set of documents and thus query expansion changes only the ranking of the documents within the ranked list. As such, absolute precision and recall are not as suitable metrics. Instead, we use precision of the top k documents or precision-at-k [1, 2, 22] as our performance metric.

We measured the precision over first k documents for both our system and the baseline method, sorted by date and relevance, where k = 12, 24, 36, 48 or 60, as our INNOPAC shows twelve documents per screen. The objective of this experiment is to determine which solution will retrieve more relevant documents in the first k retrieved documents. The user model for this experiment is that the user typically reads only the first k documents and not all the documents [22]. In addition, users are usually more interested in the precision of the results displayed in the first page of the list of retrieved documents [23]. The default threshold β-value is set to 0.03 and the δ-value is set to 0.001 in this experiment.

To determine the optimal threshold β-value, we also tested the effect of using different thresholds, β-values, for query expansion. We experimented with β-values of 0.01, 0.03, 0.05 and 0.1, and we measured the precision over first k documents, where k = 12, 24, 36, 48 or 60. The δ-value is set to 0.001 for this experiment. In addition, we also tested the effect of using different δ-values in the cohesion weight Equation 2 to find out the optimal δ-value in our cohesion weight equation. For the δ-values of 0.001, 0.01, 0.05 and 0.1, we measured the precision over first k documents for query expansion, where k = 12, 24, 36, 48 or 60. We used the default β-value of 0.03 for this experiment.

5.2 Data Collection

For our experiments, we collected queries from real OPAC users in the School of Computing at National University of Singapore (NUS) and at various discipline specific libraries. We conducted short interviews with the users to document their information needs. A total of 39 queries and their descriptions were collected.

The users were requested to provide the query keywords they used, describe what they were searching for in detail, and identify topics that are likely to be relevant as well as topics that are likely to be irrelevant. Based on the descriptions given, we were able to judge the relevance of the documents.

These queries had an average length of 2.05 words and cover various topics from computer science to medicine. The experiments were conducted on the heterogeneous NUS LINC corpus, which consisted of 1,209,509 unique titles as at June 2003[2].

5.3 Experimental Results

We now present the experimental results of query expansion on LINC, using the metric discussed earlier, precision-at-k. The original (unexpanded) queries, sorted by date and relevance, were used as the baseline in the experiments. The results are presented in Table 1.

k	Baseline (Date)	Baseline (Relevance)	Query Expansion
12	0.4209	0.4209	0.5747 (+36.55%, +36.55%)
24	0.3536	0.3856	0.5128 (+45.02%, +32.96%)
36	0.3169	0.3589	0.4686 (+47.87%, +30.56%)
48	0.2932	0.3440	0.4375 (+49.18%, +27.17%)
60	0.2743	0.3192	0.4038 (+47.20%, +26.51%)
Average	0.3318	0.3657	0.4795 (+44.51%, +31.10%)

Table 1. Comparison of baseline and query expansion results.

Our thesaurus-based query expansion performed very well as compared to using LINC without query expansion, with an improvement of 44.51% and 31.10% performance improvement over the average precision-at-k, for date and relevance sorting, respectively. This suggests that our version of query expansion is indeed useful in improving the retrieval effectiveness of the search. The reason for the improved performance is that some relevant documents which are ranked low by the original queries are propelled to the top of the ranked output because they contain many subject headings. In addition, query expansion was able to improve the retrieval performance of ambiguous queries. An example is the query "erp", in which the user's intention was to find books related to Enterprise Resource Planning (ERP) but some of the documents retrieved by the unexpanded query included terms such as expressway robbery permit and Event-Related Potentials, which were irrelevant to the user's information needs.

Determining the threshold, β-value In Table 2, we list the retrieval performance of query expansion using different β-values of 0.01, 0.03, 0.05 and 0.1.

[2] http://www.lib.nus.edu.sg/about/stats02-03.html

k	$\beta = 0.01$	$\beta = 0.03$	$\beta = 0.05$	$\beta = 0.1$
12	0.5512	0.5747	0.5534	0.5427
24	0.4882	0.5138	0.4967	0.4679
36	0.4487	0.4707	0.4537	0.4301
48	0.4209	0.4380	0.4262	0.4059
60	0.3863	0.4038	0.3910	0.3756
Average	0.4591	0.4802	0.4642	0.4445

Table 2. Effect of threshold β-value on performance of query expansion.

Table 2 shows the effect of β-value on the performance of query expansion. We can see that the average precision-at-k tends to be slightly higher for β-values of 0.03 and 0.05. This is because when β-value gets too large, potentially relevant subject headings lower than the threshold are sometimes not selected, for example, for the query "statistics", the relevant subject heading "Mathematical Statistics" was omitted. Thus, setting β-value too high will degrade retrieval performance by omitting potentially relevant subject headings. On the other hand, when β-value is too small, irrelevant subject headings are selected, for example for the query "culture shock", irrelevant subject heading "Cell Culture" was added. A small β-value will allow irrelevant subject headings to be added, which also decreases retrieval performance.

δ-value To find out the optimal value for δ in the cohesion weight equation, we measured the precision over the first k documents retrieved by our system. We use different values for δ in the cohesion weight Equation 2 and compare the results below.

k	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.1$
12	0.5747	0.5384	0.5128	0.4914
24	0.5128	0.4850	0.4529	0.4423
36	0.4686	0.4423	0.4166	0.4002
48	0.4375	0.4145	0.3931	0.3755
60	0.4038	0.3833	0.3662	0.3504
Average	0.4795	0.4527	0.4283	0.4120

Table 3. Effect of δ-value on performance of query expansion.

Table 3 shows the effect of δ-value on the performance of query expansion. We can see that the average precision-at-k tends to decrease as δ-value increases. Using a δ-value of 0.001 as the baseline, average precision-at-k fell by 5.59%, 10.67% and 14.08% when the δ-value increases to 0.01, 0.05 and 0.1 respectively. This is because when δ-value gets too large, it dominates the cohesion weight equation that we discussed earlier, making the more crucial factor co_weight less important. The cohesion weights of the subject headings then become inaccurate, which often causes relevant subject headings to be omitted. To illustrate,

the relevant subject heading "Evolution (Biology)" was omitted for the query "evolution" and for the query "C", the relevant subject heading "C (Computer Program Language)" was omitted. If a small δ-value is used, subject headings co-occurring with more terms are given heavier weights. Xu and Croft [5] mentioned "concepts co-occurring with all query terms are good for precision". Our experimental results also imply that δ-value should not be too large, as it is only a smoothing factor and should not dominate the cohesion weight equation.

6 Conclusion

We proposed a method for automatic query expansion in OPACs based on the domain knowledge contained in an automatically constructed thesaurus which maps query keywords to document subject headings. To build this thesaurus, historical query logs were analyzed to find out the most frequently-occurring queries and keywords library patrons use. Document terms were then extracted from the top-ranked documents retrieved by these queries and statistical correlations between query keywords and document subject headings were created to support query expansion. The experimental results show that our solution is practical and offers significantly better performance than the unexpanded baseline. Precision on the first screen of ranked documents improved by over 30% in our experiments.

Although we have successfully incorporated our query expansion technique into a widely-used OPAC and demonstrated its effectiveness, there is still much room for improvment. In our future work, we plan to investigate how phrase structure can refine the terms collected in our OPAC-specific thesaurus. In addition, we are exploring how document metadata such as MARC metadata, can be harnessed for further query expansion.

7 Acknowledgments

We wish to thank our colleagues over at NUS Libraries for their generous contribution of the LINC query logs for our research use and their continued support of our on-going work to improve OPAC usability. We also would like to thank the anonymous reviewers for their helpful suggestions.

References

1. Carpineto, C., De Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems **19** (2001) 1–27
2. Carpineto, C., Romano, G., Giannini, V.: Improving retrieval feedback with multiple term-ranking function combination. ACM Transactions on Information Systems **20** (2001) 259–290

3. Xu, J., Croft, W.: Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '96), Zurich, Switzerland, ACM Press (1996) 4–11
4. Cui, H., J.-R., W., Nie, J.Y., Ma, W.Y.: Query expansion by mining user logs. IEEE Transactions on Knowledge and Data Engineering **15** (2003) 829–839
5. Xu, J., Croft, W.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems **18** (2000) 79–112
6. Sparck Jones, K.: Automatic Keyword Classification for Information Retrieval. Butterworths, London, UK (1971)
7. Qiu, Y., Frei, H.: Concept based query expansion. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '93), Pittsburgh, USA, ACM Press (1993) 160–169
8. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47
9. Jing, Y., Croft, W.: An association thesaurus for information retrieval. In: Proceedings of the Intelligent Multimedia Information Retrieval Systems. (RIAO '94), New York, USA (1994) 146–160
10. Salton, G., C., B.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science and Technology **41** (1990) 288–296
11. Radecki, T.: Incorporation of relevance feedback into boolean retrieval systems. In: Proceedings of the 5th Annual ACM Conference on Research and Development in Information Retrieval, West Berlin, Germany, ACM Press (1982) 133–150
12. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science **27** (1976) 129–146
13. Sparck Jones, K.: Search term relevance weighting given little relevance information. Journal of Documentation **35** (1979) 30–48
14. Aalbersberg, I.: Incremental relevance feedback. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark (1992) 11–22
15. Eguchi, K., Ito, H., A., K., Y., K.: Adaptive and incremental query expansion for cluster-based browsing. In: Proceedings of the 6th International Conference on Database Systems for Advanced Applications, (DASFAA '99,, Hsinchu, Taiwan, IEEE Computer Society (1999) 25–34
16. Mitra, M., Singhal, A., C., B.: Improving automatic query expansion. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '98), Melbourne, Australia, ACM Press (1998) 275–281
17. Voorhees, E.M., Harman, D.: Overview of the 6th text retrieval conference (trec-6). In: Proceedings of the 6th Text Retrieval Conference (TREC-6). Number 500-240 in NIST Special Publication (1998)
18. Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: Social searching? In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 1997), Philadelphia, USA, ACM Press (1997) 306–313
19. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '94), Dublin, Ireland, ACM Press (1994) 61–69

20. Markey, K.: Subject searching in library catalogs: Before and after the introduction of online catalogs. Number 4 in OCLC Library, Information and Computer Science Series. OCLC Online Computer Library Center, Dublin, Ohio (1984)
21. Wu, H., Salton, G.: A comparison of search term weighting: term relevance vs. inverse document frequency. In: Proceedings of the 4th Annual International ACM SIGIR Conference on Information storage and retrieval: theoretical issues in information retrieval, Oakland, California, ACM Press (1981) 30–39
22. Davis, E.: Web search engines: Retrieval. http://www.cs.nyu.edu/courses/fall02/G22.3033-008/lec5.html (2002)
23. Kobayashi, M., Takeda, K.: Information retrieval on the web. ACM Computing Surveys **32** (2000) 144–173