Document frequency and term specificity

Hideo Joho
Department of Computing Science
University of Glasgow
17 Lilybank Gardens
Glasgow G12 8QQ, UK.
hideo@dcs.gla.ac.uk

Mark Sanderson
Department of Information Studies
University of Sheffield
Regent Court, 211 Portobello St.
Sheffield S1 4DP, UK.
m.sanderson@sheffield.ac.uk

Abstract

Document frequency is used in various applications in Information Retrieval and other related fields. An assumption frequently made is that the document frequency represents a level of the term's specificity. However, empirical results to support this assumption are limited. Therefore, a large-scale experiment was carried out, using multiple corpora, to gain further insight into the relationship between the document frequency and terms specificity. The results show that the assumption holds only at the very specific levels that cover the majority of vocabulary. The results also show that a larger corpus is more accurate at estimating the specificity. However, the co-occurrence information is shown to be effective for improving the accuracy when only a small corpus is available.

Introduction

Topic hierarchies have long interested researchers in information retrieval (IR), computational linguistics (CL), and more recently in ontology formation for the semantic web. In IR, such structures have been employed to aid users in browsing sets of documents and in helping them formulate or later expand their queries. In CL, such hierarchies have been used as a resource for other language-based tasks.

The means of automatically creating such hierarchies remains an active subject, where the nodes of the hierarchy (composed of concepts or individual words) are arranged in some taxonomic structure with general concepts at the top of the hierarchy leading to related and more specific concepts below. Methods for locating words or phrases that would be good candidate concepts and means of determining their relationship (through some measure of co-occurrence) have been well studied (Grefenstette, 1992; Anick and Tipirneni, 1999).

Somewhat less examined, however, is the issue of term specificity: given a pair of terms/concepts that have been found to be related, how does one determine which is the more specific concept and which is the more general? One approach is to use document frequency: given a pair of terms, measure their frequency of occurrence in a collection and the term with the lower *document frequency* (*df*: the number of documents in a collection in which a term occurs) is chosen to be the more specific. The notion of determining specificity in such a manner is not new. In 1968, Salton suggested such an approach to ordering terms in a hierarchy (Salton, 1968). Spärck Jones (1972) suggested that specificity should be measured by *df*, where a less frequent term was regarded as more specific. Spärck Jones commented that this type of specificity was not necessarily the same as a semantic perspective but it was useful for retrieval systems in the form of *idf*: *inverse* document frequency. In a similar context, Barker et al (1972) estimated term specificity by determining the total number of documents containing a term, and calculating what proportion was relevant. This was designed to determine how specific a term was to a particular query.

Salton's idea of using frequency to order terms in a hierarchy was found again in work by Forsyth and Rada (1986) where a limited scale concept hierarchy was constructed and related terms were ordered by frequency, again with the assumption that the most general terms were the most frequent. It would appear, however, that throughout this early work, little actual testing of the relationship between frequency and specificity was conducted. Document frequency was used to determine term specificity in Sanderson and Croft's work building topic hierarchies (1999). As with many previous works, no test was conducted to examine the correlation between term specificity and document frequency. However, a user-based study provided some evidence of the ability for document frequency to order terms based on specificity.

Weinberg and Cunningham carried out a test to examine the relationship between subject terms in MeSH and the number of documents in MEDLINE containing those terms (1985). They selected a hierarchical tree of subject terms under the term *Endocrine Diseases*, chosen as it was a term typical of core topics in MeSH and MEDLINE. The tree was composed of around 100 terms covering four hierarchical levels. The researchers also selected another similar sized tree under the term *Environment* chosen as a more peripheral topic. They examined the *df* of terms from the different levels of the hierarchy and found a negative correlation between the depth (level) of the hierarchies and number of documents in which terms occurred: terms further down the sub-tree had a lower *df*. The negative correlation of the terms in the central tree (-0.20) was larger than in the peripheral one (-0.13), although both correlations appeared to be weak.

A test was conducted by Caraballo and Charniak (1999) who examined three fragments of the WordNet *hypernym*[1] hierarchy measuring the *df* of hierarchy words in a corpus, in their case, the 1987 Wall Street Journal. They showed that use of document frequency to order (by specificity) term pairs taken from the lower part of the WordNet hierarchy resulted in correct ordering 86% of the time. They also examined two fragments from the upper part of the hierarchy (i.e. the most general), here they found the accuracy of term frequency at ordering pairs was 45.5%, given that random ordering of term pairs would obtain an accuracy of 50%, they concluded that term frequency was of no use in determining specificity in this part of WordNet. The boundary between the upper and lower parts of WordNet was stated by Caraballo and Charniak to be the point in the hierarchy where they judged *basic level categories* were found[2].

More recently, Ryu and Choi (2004) conducted a similar study of the use of frequency in determining specificity focusing on multi-word medical terms, testing around 436 disease names, measuring df in 170,000 abstracts (120Mb of text) taken from the Medline collection. They found that *df* determined specificity with an accuracy of 60.6%. Quite why their numbers were so different from Caraballo and Charniak was not addressed by them.

It was decided to conduct a further study of *df* and specificity to address some of the short comings of the past work, namely,

---

[1] A hypernym is a word that is a more general concept of another word, e.g. the word amphibian is a hypernym of the word frog.

[2] A basic level category (Lakoff, 1987) is the most common level of detail in categorisation. In learning vocabulary, Lakoff explains that there is a category that one learns first. An example of an animal basic level category is bird, a word that a child might learn first; other levels, e.g. more specific such as dove, or more general such as animal, are learnt later.

- Past work has determined *df* in mid-sized collections of text, running into several tens of megabytes. It would be valuable to examine the impact of using both smaller and larger corpora when determining *df*.

- Past work has not examined in detail the accuracy of *df* at different levels of specificity, again it is likely that a more detailed examination will provide further information.

- It is often the case when considering the ordering of term pairs in a hierarchy, those terms are found to co-occur with each other in a particular set of texts. In the reviewed past work, any relationship between co-occurrence and specificity in the corpus was ignored.

The rest of the paper starts with a description of the methodology of the experiments, followed by the detailing of the three experiments measuring the relationship between *df* and specificity. The implications of the results for the development of topical hierarchies are then discussed before the paper concludes.

Methodology

The aim of our experimental work was to test on a large-scale, the ability of *df* to determine specificity. The experimental design has its roots in Caraballo and Charniak's approach of using WordNet as a source of words and phrases already manually ordered by specificity (through the hypernym relation) along with a corpus to measure *df*. The primary concerns therefore, were determining the corpus to be used to measure *df* and the part of WordNet to be examined.

Choosing and using a corpus

We started with the assumption that all 45,000 nouns and noun phrases in WordNet[3] (Miller, 1990) would be used in our experiments. Different corpora were examined for their coverage of the WordNet terms. The corpora used were a large fragment of the TREC collection (the Financial Times, 1991-94, LA Times, 1989-89 and Wall Street Journal, 1987-92) and the Web pages held by Google. With Google, each WordNet word or phrase was issued to the search engine as a query; it was assumed that if some number of documents were retrieved in response, the term was in the search engine's corpus.

WordNet	45,073	Coverage
Google	45,055	99.96%
TREC	23,705	52.59%

Table 1. Vocabulary coverage in collections.

As can be seen from an examination of the corpora (in

Table 1), Google covered almost all of WordNet while the TREC collection covered just over 50%.

Examining specificity and dealing with ambiguity

In order to examine the utility of *df* for determining specificity, the *hypernym chains* of each term in WordNet were obtained by iterating over hypernym relations from synsets at the bottom of the hierarchy to the root at the top. Figure 1 shows an example hypernym chain of the term "eye contact".

[3] v 1.5

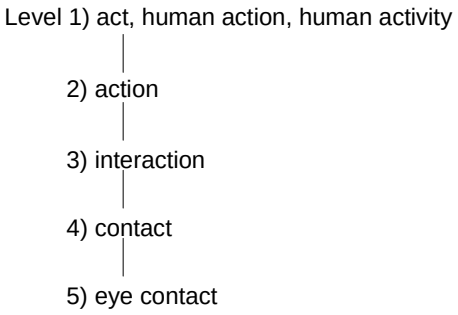

Figure 1: example extracted hypernym chain of length five.

A total of 59,920 hypernym chains were found in WordNet ranging in length from two to sixteen (here, length is measured by the number of nodes in the chain). As with almost any study involving the determination from corpora of a term's attributes, the problem of word sense was considered. The atomic units of WordNet are *synsets*: sets of word senses that are synonyms of each other. Each word in a synset refers to a particular sense of that word (e.g. "contact" in Figure 1: a type of interaction as opposed to, for example, the short form of "contact lens"). Ideally in order to compute the *df* of the word sense(s) of a synset, a sense tagged corpus should be used. However, there are no large corpora of this type. In order to estimate frequency of occurrence of senses from a corpus of words, it was decided to focus our study on a sub-set of WordNet that one could be more confident in measuring.

Sanderson and Van Rijsbergen (1999) showed that across 15,000 tested words, the commonest sense of a word accounted for the outright majority of the word's occurrences in a corpus (regardless how many senses that word has). By focusing the study in this paper on hypernym chains composed only of synsets composed of a terms used in their commonest sense (as defined by WordNet[4]), one could assume that the frequency of occurrence of the term in a corpus was reasonably well correlated to the occurrence of its prevailing sense. A similar approach to dealing with term ambiguity was taken by Caraballo and Charniak in their work. Therefore, from WordNet's 59,920 hypernym chains, a subset of 25,242 chains that consisted of only commonest sense synsets was used. Table 2 shows the number of chains found in Wordnet organized by chain length.

Chain length	Commonest Sense	Chain length	Commonest Sense
2	0		
3	157	10	1,480
4	733	11	863
5	2,696	12	368
6	5,297	13	262
7	5,681	14	115
8	4,793	15	24
9	2,773	16	0
		Total	25,242

Table 2. Length and distribution of chain length.

[4] The means that the creators of WordNet used to determine the frequency of occurrence of a word's sense was two fold: first if the word occurred in the Sense Eval corpus (a subset of the Brown corpus, where the senses of its constituent words were manually disambiguated), the commonest sense was measured from there; if the word did not occur in that corpus, then the commonest sense was determined by the WordNet creators based on their lexical/world knowledge (Miller, 1995).

The best corpus for determining specificity

In order to determine which corpus to measure *df* in, a test was conducted to examine the relationship between corpora size and accuracy of *df* to determine specificity. Two corpora were used: the Web as held by Google and the TREC newspaper subset as defined above. Document frequency was determined in the Web pages indexed by Google by reading the number of results estimated to be retrieved in response to the WordNet noun or phrase being issued as a query, see Figure 2. For those synsets that were composed of more than one word, the *df* of a synset was best calculated by averaging the document frequency of each member word.

Figure 2. Reading the estimated *df* for a WordNet noun or phrase from Google.

Table 3 shows the result of the test showing that document frequency obtained from Google was slightly more accurate at determining specificity than from the TREC collection. A significance test applied to the averages using the t-test showed the null hypothesis (of the Google and TREC corpora producing the same *df* specificity accuracy) did not hold with $p<0.05$. This result was unexpected as it was expected that *df* measured in more homogeneous collections (such as a newspaper corpus) was likely to be more accurate as the corpus was topically coherent compared to a heterogeneous collection such as the Web. However the result shows that *df* measured from Google was best to identify specificity and it alone was used in subsequent analyses of the WordNet hypernym chains.

Collection	Accuracy
TREC	70.93%
Google	72.36%

Table 3. Accuracy of *df* in two corpora.

Analysis of hypernyms

Two analyses were undertaken to investigate the relationship between document frequency and term specificity. The first examined hypernym chains at different levels of specificity; the second examined the relationship of document frequency, specificity and co-occurrence.

Levels of specificity and document frequency

The first analysis measured the accuracy of *df* at determining specificity by investigating the number of cases where a WordNet synset in a hypernym chain had a higher document frequency than the *df* of the synset in the level immediately below. When examining individual chains extracted from WordNet, inevitably, most of the parts of each chain overlap. As can be seen, chains extracted from a hierarchy (as seen in Figure 3) would share almost every node (i.e. synsets) with at least one other chain. In this analysis, overlapping nodes from different chains were examined only once.

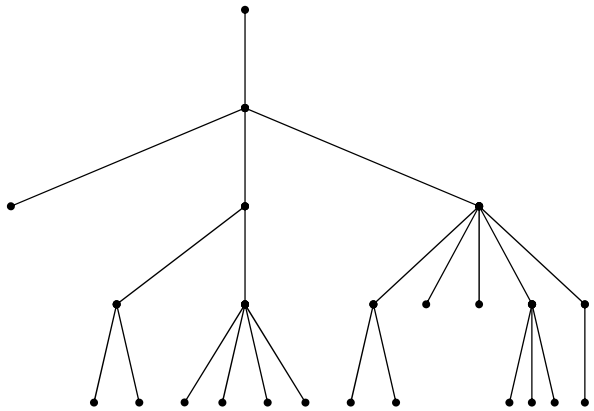

Figure 3: Example hypernym hierarchy, showing chains ranging in length three to five.

The results of the analysis are shown in Table 4, where for each chain length, the accuracy of using *df* to order pairs of nodes (by specificity) at each level of the chain is shown. When examining the node pairs, the node at the higher level is referred to as the *parent* node and the one below is its *child*.

Chain length	No. of chains	Level 1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11
3	157	60.71 (28)	**80.89** (157)								
4	733	55.56 (27)	58.92 (185)	**81.82** (673)							
5	2696	56.52 (23)	60.98 (123)	65.02 (486)	**82.44** (2170)						
6	5297	61.11 (18)	56.96 (79)	66.78 (298)	65.05 (910)	**78.51** (4142)					
7	5681	60.00 (15)	58.00 (50)	67.48 (163)	60.67 (417)	68.95 (1269)	**81.75** (4532)				
8	4793	61.54 (13)	60.00 (30)	67.01 (97)	57.51 (201)	62.42 (471)	68.35 (1185)	**78.25** (4033)			
9	2773	60.00 (10)	72.73 (22)	62.26 (53)	59.34 (91)	55.26 (190)	65.44 (353)	60.40 (808)	**79.75** (2454)		
10	1480	57.14 (7)	68.75 (16)	57.14 (35)	56.00 (50)	46.75 (36)	69.23 (90)	50.23 (215)	56.17 (486)	**76.41** (1361)	
11	863	60.00 (5)	66.67 (12)	72.22 (18)	50.00 (26)	43.24 (37)	71.70 (53)	48.89 (90)	50.66 (152)	55.67 (300)	**74.66** (817)

Table 4. Shows the percentage of parents that have a higher *df* than their child. The figures in brackets are the number of unique parent/child pairs.

From the figures in Table 4 a number of points can be drawn. The number of distinct parent child pairs (the numbers in brackets) increased as one dropped to lower levels. Note that the number of pairs at the lowest level (the far right of each row in the table) was almost always not as high as the actual number of distinct chains being analyzed. This difference in values reveals that the WordNet hypernym structure is not a strict hierarchy where every child node

has only one parent, but in fact a *directed acyclic graph*. Occasionally a child node in the hypernym structure has more than one parent.

The accuracy of *df* at ordering parent child pairs correctly by specificity is relatively consistent across the different levels of the chain length, however, the highest accuracy (highlighted in **bold**) was always found at the lowest level. The figures shown here contradict somewhat the results presented by Caraballo and Charniak in their 1999 work: there they stated that frequency could determine specificity 83% of the time for parent-child pairs below the basic level and 45.5% above the level. Here it would appear that the high levels of accuracy are really only apparent for the parent-child pairs found at the very bottom of each hypernym chain and to a lesser extent for the pair above the bottom pair. For all other pairs, accuracy is lower and relatively constant. Aggregating all the chain accuracy figures into a single graph confirms this analysis. The overall accuracy of *df* at determining specificity across all WordNet parent child pairs was 74%. The average is influenced by the great many pairs at the bottom two levels of WordNet.

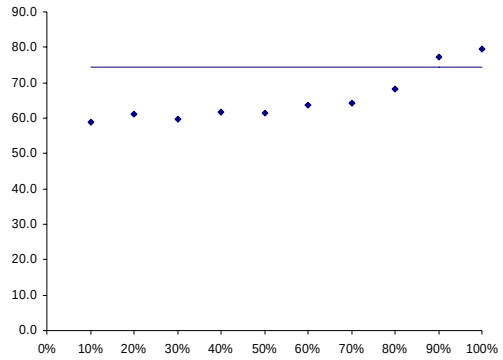

Figure 4: Graph of *df* accuracy against chain depth; values taken from Table 4. The x-axis is the percentage position of the chain depth; the y-axis is a weighted average of *df* accuracy; values are calculated at standard values using an interpolation method for calculating a recall precision graph from Baeza-Yates & Ribeiro-Neto (1999). The average *df* accuracy is drawn in as a line.

Impact of co-occurrence information

Up to this point, term specificity was determined through frequency of occurrence alone, however, it is often the case when determining which of a pair of terms is the more specific those terms are found to co-occur with each other in a particular set of texts. In information retrieval, often those texts are documents retrieved in response to a query. Therefore, examining collocated word pairs in corpora and retrieved documents constituted the last part of our analysis: examining the impact of co-occurrence information when considering term specificity.

To examine such effects, the analysis described in the Section titled “The best corpus…” was repeated, but with additional corpora composed of sets of retrieved documents. To form these corpora, one hundred TREC topics (301-400) were run against a BM25 based search engines retrieving from the each of the three component newspaper collections of the TREC corpus: the FT, LA Times and WSJ. The three corpora were collectively referred to as the *Top500 corpora*. Those WordNet noun pairs where both a parent and child occurred in at least one of the top 500 retrieved documents were noted. In addition, those pairs found to collocate within at least one of the 500 documents were also recorded. Counts of the number of cases where parent terms held a higher *df* than the child were made. The pairs used in the analysis were limited to those pairs found to exist in the smallest of the corpora, which was the Top500. The

accuracy of *df* at ordering parent child pairs by specificity for the co-occurring term pairs (found in the Top500) was also measured in the TREC and Google collections. Note, the Google collection was not checked to see if the term pairs co-occurred in the Google collection.

	Without co-occurrence			With co-occurrence		
Collections	FT	LA	WSJ	FT	LA	WSJ
No. of queries	100	100	100	100	100	100
No. of pairs	122,913	166,712	155,421	9,951	15,824	14,246
Top500	65.1%	67.1%	66.1%	82.0% (+25.8)	83.2% (+23.8)	81.4% (+23.1)
TREC	70.6%	71.3%	70.8%	83.2% (+17.8)	84.0% (+17.9)	83.4% (+18.7)
Google	72.0%	72.5%	72.5%	83.4% (+15.9)	84.3% (+16.2)	83.4% (+14.9)

Table 5. Effect of co-occurrence information on *df* at determining term specificity

Table 5 shows the results of the analysis examining pairs found to occur in the corpora and the subset of pairs that co-occur in at least one document. There are certain points to note.

- The experimental results showed that determining specificity using *df* from co-occurring term pairs is more accurate. The impact of co-occurrence was found to be more significant when *df* was obtained in a smaller size of documents. The improvement of accuracy was around 24% for the *df* in the Top 500 documents but only ~16% in *df* determined from Google.

- The *df* obtained from terms co-occurring in a larger size of corpus was found to be more accurate at determining specificity, but was much less pronounced than the increases seen when determining specificity from terms that do not co-occur.

- The constancy of the result across the collections should be emphasized. We were aware of the potential problem of WordNet's definition of commonest sense when considering the domain and heterogeneity of the tested corpora. If the senses of terms used in the Web and TREC Collections were significantly different, one would expect the results to have varied across the collections. However, our experiment indicated this was not the case. It would appear that on average, the usage of sense across the Web corpus was similar to sense usage in the newspaper corpora, which is perhaps surprising given the differences in age and domain of the corpora.

Summary

This paper addressed several aspects of the relationship between document frequency (*df*) and term specificity by using a significantly larger size of vocabulary and corpus than previous works and by examining the relationship in new ways compared to past work. The series of analyses involved measuring average document frequency at various levels in hypernym chains, comparing parent-child term pairs from WordNet, and evaluating the impact of co-occurrence information on determination of specificity.

The first analysis examined the impact of corpus size on the accuracy of *df* at determining term specificity; it was shown that use of larger corpora facilitated higher accuracy.

The second experiment focused on parent-child pairs in hypernym chains to reveal the accuracy of *df* to identify specificity from a given term pair. Across chains of different length, it was found that the highest probability of parent synsets holding a higher document frequency than their child synsets was found between the pairs of the last two levels in most cases. From these results, it is believed that *df* can be used to determine term specificity most accurately when the terms are very specific.

The last experiment observed the effect of co-occurrence information on the probability of identifying a more general term of given pairs. The result showed 15% to 25% improvement in accuracy of the identification with the co-occurrence information. More improvement was added to document frequency obtained from a local document set (i.e. Top500 docs) than global set (i.e. Google). Although *df* from a larger collection was found to be more accurate for the identification (due, it is assumed, to the bigger sample size of word occurrences), the result indicated the co-occurrence information can be useful where df was only obtained from a small set of documents.

When the *df* is used as a means of generating a concept hierarchy, therefore, the statistics obtained from a larger collection is likely to provide a better result than a smaller set. A wider range of vocabulary will also be found in the larger collection. However, there will be the case where it is infeasible to access to the data set as large as the index of a major search engine's collection. In such a case, the co-occurrence information was shown to be useful for improving the accuracy of ordering concepts based on specificity.

References

Anick, P.G. & Tipirneni, S. (1999). The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. In: Hearst, M., Gey, F. & Tong, R. (eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 153-161. ACM, Berkeley, CA.

Barker, F.H., Veal, D.C. & Wyatt, B.K. (1972). Towards automatic profile construction. *Journal of Documentation*, **28** (1), 44-55.

Caraballo, S.A. & Charniak, E. (1999). Determining the specificity of nouns from text. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for retrieval. In: Belkin, N.J., Ingwersen, P. & Pejtersen, A.M. (eds.), *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 89-97. ACM, Copenhagen, Denmark

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago, IL: University of Chicago Press.

Miller, G.A. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, **3** (4), 245-264.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, **38** (11), 39-41.

Peat, H.J. & Willet, P. (1991). The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. *Journal of the American society for Information Science*, **42** (5), 378-383.

Qiu, Y. & Frei, H.P. (1993). Concept Based Query Expansion. In: Korfhage, R., Rasmussen, E.M. & Willett, P. (eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, Pittsburgh, PA. pp. 160-169. ACM, Pittsburgh, PA.

Ryu, P.M., Choi, K.S (2004) Measuring the Specificity of Terms for Automatic Hierarchy Construction: *KAIST, Korea ECAI-2004 Workshop on Ontology Learning and Population*

Salton, G. (1968). *Automatic Information Organization and Retrieval.* New York: McGraw-Hill.

Sanderson, M. & Croft, B. (1999). Deriving Concept Hierarchies from Text. In: Hearst, M., Gey, G. & Tong, R. (eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 206-213. ACM, Berkeley, CA.

Sanderson, M. & Van Rijsbergen, C.J. (1999). The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems*, **17** (4), 440-465.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, **28** (1), 11-21.

Voorhees, E.M. & Harman, D.K. (eds.) (1998). *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*. Gaithersburg, MD: NIST.

Weinberg, B.H. & Cunningham, J.A. (1985). The Relationship Between Term Specificity in MeSH and Online Postings in MEDLINE. *Bulletin of the Medican Library Association*, **73** (4), 365-372.

Xu, J. & Croft, B.W. (1996). Query Expansion Using Local and Global Document Analysis. In: Frei, H.-P., Harman, D., Schäuble, P. & Wilkinson, R. (eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland. pp. 4-11. ACM, Zurich, Switzerland.