The Correlation between Cluster Hypothesis Tests and the Effectiveness of Cluster-Based Retrieval

Fiana Raiber
fiana@tx.technion.ac.il

Oren Kurland
kurland@ie.technion.ac.il

Faculty of Industrial Engineering and Management, Technion
Haifa 32000, Israel

ABSTRACT

We present a study of the correlation between the extent to which the cluster hypothesis holds, as measured by various tests, and the relative effectiveness of cluster-based retrieval with respect to document-based retrieval. We show that the correlation can be affected by several factors, such as the size of the result list of the most highly ranked documents that is analyzed. We further show that some cluster hypothesis tests are often negatively correlated with one another. Moreover, in several settings, some of the tests are also negatively correlated with the relative effectiveness of cluster-based retrieval.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords: cluster hypothesis, cluster-based retrieval

1. INTRODUCTION

The cluster hypothesis states that "closely associated documents tend to be relevant to the same requests" [19]. The hypothesis plays a central role in information retrieval. Various tests were devised for estimating the extent to which the hypothesis holds [5, 20, 3, 17]. Furthermore, inspired by the hypothesis, document retrieval methods that utilize document clusters were proposed (e.g., [10, 11, 6, 7, 15]).

There are, however, only a few reports regarding the correlation between the cluster hypothesis tests and the relative effectiveness of cluster-based retrieval with respect to document-based retrieval [20, 3, 13]. Some of these are contradictory: while it was initially argued that Voorhees' nearest neighbor cluster hypothesis test is not correlated with retrieval effectiveness [20], it was later shown that this test is actually a good indicator for the effectiveness of a specific cluster-based retrieval method [13].

The aforementioned reports focused on a single cluster hypothesis test (the nearest neighbor test), used a specific retrieval method which is not state-of-the-art and were evaluated using small documents collections which were mostly composed of news articles. Here, we analyze the correlation between cluster hypothesis tests and the relative effectiveness of cluster-based retrieval with respect to document-based retrieval using a variety of tests, state-of-the-art retrieval methods and collections.

We found that (i) in contrast to some previously reported results [3], cluster hypothesis tests are in many cases either negatively correlated with one another or not correlated at all; (ii) cluster hypothesis tests are often negatively correlated or not correlated at all with the relative effectiveness of cluster-based retrieval methods; (iii) the correlation between the tests and the relative effectiveness of the retrieval methods is affected by the number of documents in the result list of top-retrieved documents that is analyzed; and, (iv) the type of the collection (i.e., Web vs. newswire) is a strong indicator for the effectiveness of cluster-based retrieval when applied over short retrieved document lists.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609533.

2. RELATED WORK

The correlation between cluster hypothesis tests was studied using small document collections, most of which were composed of news articles [3]. We, on the other hand, use a variety of both (small scale) newswire and (large scale) Web collections. The correlation between cluster hypothesis tests and the effectiveness of cluster-based retrieval methods was studied using only a single test — Voorhees' nearest neighbor test [20, 13]. Each study also focused on a different cluster-based retrieval method. This resulted in contradictory findings. In contrast, we use several cluster hypothesis tests and retrieval methods.

Document clusters can be created either in a query dependent manner, i.e., from the list of documents most highly ranked in response to a query [21] or in a query independent fashion from all the documents in a collection [5, 10]. In this paper we study the correlation between cluster hypothesis tests and the effectiveness of retrieval methods that utilize *query dependent* clusters [6, 7, 15]. The reason is threefold. First, these retrieval methods were shown to be highly effective. Second, we use for experiments large-scale document collections; clustering all the documents in these collections is computationally difficult. Third, the cluster hypothesis was shown to hold to a (much) larger extent when applied to relatively short retrieved lists than to longer ones or even to the entire corpus [18].

3. CLUSTER HYPOTHESIS TESTS AND CLUSTER-BASED RETRIEVAL

To study the correlation between tests measuring the extent to which the cluster hypothesis holds and the effective-

ness of cluster-based retrieval methods, we use several tests and (state-of-the-art) retrieval methods.

Let $\mathcal{D}_{\text{init}}$ be an initial list of n documents retrieved in response to query q using some retrieval method. The retrieval method scores document d by $score(q, d)$. (Details of the scoring function used in our experiments are provided in Section 4.) All the cluster hypothesis tests and the retrieval methods that we consider operate on the documents in $\mathcal{D}_{\text{init}}$. In what follows we provide a short description of these tests and methods.

Cluster hypothesis tests. The first test that we study conceptually represents the **Overlap** test [5]. The test is based on the premise that, on average, the similarity between two relevant documents should be higher than the similarity between a relevant and a non-relevant document. Formally, let $R(\mathcal{D}_{\text{init}})$ be the set of relevant documents in $\mathcal{D}_{\text{init}}$ and $N(\mathcal{D}_{\text{init}})$ the set of non-relevant documents; n_R and n_N denote the number of documents in $R(\mathcal{D}_{\text{init}})$ and $N(\mathcal{D}_{\text{init}})$, respectively. The score assigned by the Overlap test to $\mathcal{D}_{\text{init}}$ is $\frac{\frac{1}{n_R(n_R-1)}\sum_{d_i,d_j\in R(\mathcal{D}_{\text{init}}),d_i\neq d_j}\left(sim(d_i,d_j)+sim(d_j,d_i)\right)}{\frac{1}{n_R n_N}\sum_{d_i\in R(\mathcal{D}_{\text{init}}),d_j\in N(\mathcal{D}_{\text{init}})}\left(sim(d_i,d_j)+sim(d_j,d_i)\right)}$; $sim(\cdot,\cdot)$ is an inter-text similarity measure described in Section 4.[1] This score is averaged over all the tested queries for which n_R and n_N are greater than 1 to produce the final test score.

We next consider Voorhees' Nearest Neighbor test (**NN**) [20]. For each relevant document d_i ($\in R(\mathcal{D}_{\text{init}})$) we count the number of relevant documents among d_i's $k-1$ nearest neighbors in $\mathcal{D}_{\text{init}}$; k is a free parameter. These counts are averaged over all the relevant documents retrieved for all the tested queries. The nearest neighbors of d_i are determined based on $sim(d_i, d_j)$.

The **Density** test [3] is defined here as the ratio between the average number of unique terms in the documents in $\mathcal{D}_{\text{init}}$ and the number of terms in the vocabulary. The underlying assumption is, as for the tests from above, that relevant documents are more similar to each other than they are to non-relevant documents. Now, if the number of terms that are shared by documents in the initial list is high, then presumably relevant documents could be more easily distinguished from non-relevant ones.

We also explore the Normalized Mean Reciprocal Distance test (**nMRD**) [17]. The test is based on using a complete relevant documents graph. Each vertex in the graph represents a different document in $R(\mathcal{D}_{\text{init}})$; each pair of vertices is connected with an edge. The edge weight represents the distance between the documents. The distance between documents d_i and d_j is defined as the rank of d_j in a ranking of all the documents $d' \in \mathcal{D}_{\text{init}}$ ($d' \neq d_i$) that is created using $sim(d_i, d')$; the rank of the highest ranked document is 1. The score assigned by the nMRD test to $\mathcal{D}_{\text{init}}$ is $\frac{1}{n_R\sum_{i=1}^{n_R}\frac{1}{\lfloor\log_2 i\rfloor+1}}\sum_{d_i,d_j\in R(\mathcal{D}_{\text{init}}),d_i\neq d_j}\frac{1}{spd(d_i,d_j)}$; $spd(d_i, d_j)$ is the shortest path distance between d_i and d_j in the graph. This score is averaged over all tested queries for which $n_R > 1$ to produce the final nMRD score.

Cluster-based document retrieval methods. Let $Cl(\mathcal{D}_{\text{init}})$ be the set of clusters created from the documents in $\mathcal{D}_{\text{init}}$ using some clustering algorithm. All the cluster-based retrieval methods that we consider re-rank the documents in $\mathcal{D}_{\text{init}}$ using information induced from clusters in $Cl(\mathcal{D}_{\text{init}})$.

The interpolation-f method (**Interpf** in short) [6] directly ranks the documents in $\mathcal{D}_{\text{init}}$. The score assigned to document d ($\in \mathcal{D}_{\text{init}}$) is $\lambda\frac{score(q,d)}{\sum_{d_i\in\mathcal{D}_{\text{init}}}score(q,d_i)} + (1-\lambda)\frac{\sum_{c\in Cl(\mathcal{D}_{\text{init}})}sim(q,c)sim(c,d)}{\sum_{d_i\in\mathcal{D}_{\text{init}}}\sum_{c\in Cl(\mathcal{D}_{\text{init}})}sim(q,c)sim(c,d_i)}$; λ is a free parameter.

The cluster-based retrieval methods that we consider next are based on a two steps procedure. First, the clusters in $Cl(\mathcal{D}_{\text{init}})$ are ranked based on their presumed relevance to the query. Then, the ranking of clusters is transformed to a ranking over the documents in $\mathcal{D}_{\text{init}}$ by replacing each cluster with its constituent documents (and omitting repeats).

The **AMean** and **GMean** methods [12, 16] rank the clusters based on the arithmetic and geometric mean of the original retrieval scores of the documents in a cluster, respectively. Specifically, AMean assigns cluster c with the score $\frac{1}{|c|}\sum_{d\in c}score(q,d)$ where $|c|$ is the number of documents in c. The score assigned to c by GMean is $\prod_{d\in c}score(q,d)^{\frac{1}{|c|}}$.

Another cluster ranking method that we use is **ClustRanker** [7]. ClustRanker assigns cluster c with the score $\lambda\frac{cent(c)sim(q,c)}{\sum_{c_i\in Cl(\mathcal{D}_{\text{init}})}cent(c_i)sim(q,c_i)} + (1-\lambda)\frac{\sum_{d\in c}score(q,d)sim(c,d)cent(d)}{\sum_{c_i\in Cl(\mathcal{D}_{\text{init}})}\sum_{d\in c_i}score(q,d)sim(c_i,d)cent(d)}$; $cent(d)$ and $cent(c)$ are estimates of the centrality of a document d in $\mathcal{D}_{\text{init}}$ and that of a cluster c in $Cl(\mathcal{D}_{\text{init}})$, respectively. These estimates are computed using a PageRank algorithm that utilizes inter-document and inter-cluster similarities [9, 7].

We also use the recently proposed state-of-the-art **ClustMRF** cluster ranking method [15]. ClustMRF uses Markov Random Fields which enable to integrate various types of cluster-relevance evidence.

4. EXPERIMENTAL SETUP

Experiments were conducted using the datasets specified in Table 1. WSJ, AP and ROBUST are small (mainly) newswire collections. WT10G is a small Web collection and GOV2 is a crawl of the .gov domain. CW09B is the Category B of the ClueWeb09 collection and CW09A is its Category A English part. We use two additional settings, CW09BF and CW09AF, for categories B and A [2], respectively. These settings are created by filtering out from the initial ranking documents that were assigned with a score below 50 and 70 by Waterloo's spam classifier for CW09B and CW09A, respectively. Thus, the initial lists, $\mathcal{D}_{\text{init}}$, used for these two settings presumably contain fewer spam documents.

corpus	# of docs	# of unique terms	data	queries
WSJ	173,252	186,689	Disks 1-2	151-200
AP	242,918	259,501	Disks 1-3	51-150
ROBUST	528,155	663,700	Disks 4-5 (-CR)	301-450, 600-700
WT10G	1,692,096	4,999,228	WT10g	451-550
GOV2	25,205,179	39,251,404	GOV2	701-850
CW09B CW09BF	50,220,423	87,262,413	ClueWeb09 Cat. B	1-200
CW09A CW09AF	503,903,810	507,500,897	ClueWeb09 Cat. A	1-200

Table 1: TREC data used for experiments.

[1] We use both $sim(d_i, d_j)$ and $sim(d_j, d_i)$ as the similarity measure that was used for experiments is asymmetric. Further details are provided in Section 4.

The Indri toolkit was used for experiments[2]. Titles of topics served for queries. We applied Krovetz stemming to documents and queries. Stopwords were removed only from queries using INQUERY's list [1].

We use the nearest neighbor clustering algorithm to create the set of clusters $Cl(\mathcal{D}_{\text{init}})$ [4]. A cluster is created from each document $d_i \in \mathcal{D}_{\text{init}}$. The cluster contains d_i and the $k-1$ documents $d_j \in \mathcal{D}_{\text{init}}$ $(d_j \neq d_i)$ with the highest $sim(d_i, d_j)$. We set $k = 5$. Recall that k is also the number of nearest neighbors in the NN cluster hypothesis test. Using such small overlapping clusters was shown to be highly effective, with respect to other clustering schemes, for cluster-based retrieval [4, 11, 7, 14, 15].

The similarity between texts x and y, $sim(x, y)$, is defined as $\exp\left(-CE\left(p_x^{Dir[0]}(\cdot) \,\middle||\, p_y^{Dir[\mu]}(\cdot)\right)\right)$; CE is the cross entropy measure and $p_z^{Dir[\mu]}(\cdot)$ is the Dirichlet-smoothed (with the smoothing parameter μ) unigram language model induced from text z [8]. We set $\mu = 1000$ in our experiments [22]. This similarity measure was found to be highly effective, specifically for measuring inter-document similarities, with respect to other measures [9]. The measure is used to create $\mathcal{D}_{\text{init}}$ — i.e., $score(q, d) \stackrel{def}{=} sim(q, d)$ —[3] and to compute similarities between the query, documents and clusters. We represent a cluster by the concatenation of its constituent documents [10, 6, 8, 7]. Since we use unigram language models the similarity measure is not affected by the concatenation order.

To study the correlation between two cluster hypothesis tests, we rank the nine experimental settings (WSJ, AP, ROBUST, WT10G, GOV2 and the ClueWeb09 settings) based on the score assigned to them by each of the tests. Kendall's-τ correlation between the rankings of experimental settings is the estimate for the correlation between the tests. We note that Kendall's-τ is a *rank* correlation measure that does not depend on the actual scores assigned to the settings by the tests. Kendall's-τ ranges from -1 to $+1$ where -1 represents perfect negative correlation, $+1$ represents perfect positive correlation, and 0 means no correlation.

The correlation between a cluster hypothesis test and the relative effectiveness of a cluster-based retrieval method is also measured using Kendall's-τ. The experimental settings are ranked with respect to a cluster-based retrieval method by the performance improvement it posts over the original document-based ranking. Specifically, the ratio between the Mean Average Precision at cutoff n (MAP@n) of the ranking induced by the method and the MAP@n of the initial ranking is used; n is the number of documents in $\mathcal{D}_{\text{init}}$.

The free-parameter values of Interpf, ClustRanker and ClustMRF were set using 10-fold cross validation. Query IDs were used to create the folds; MAP@n served for the optimization criterion in the learning phase. The value of λ which is used in Interpf and ClustRanker is selected from $\{0, 0.1, \ldots, 1\}$. To compute the document and cluster centrality estimates in ClustRanker, the dumping factor and the number of nearest neighbors that are used in the PageRank algorithm were selected from $\{0.1, 0.2, \ldots, 0.9\}$ and $\{5, 10, 20, 30, 40, 50\}$, respectively. The implementation of ClustMRF follows that in [15].

[2]www.lemurproject.org/indri

[3]Thus, the initial ranking is induced by a standard language-model-based approach.

		Overlap	NN	Density	nMRD
$n = 50$	Overlap	1.000	−0.171	−0.778	0.278
	NN	−0.171	1.000	0.229	−0.229
	Density	−0.778	0.229	1.000	−0.056
	nMRD	0.278	−0.229	−0.056	1.000
$n = 100$	Overlap	1.000	−0.354	−0.833	−0.222
	NN	−0.354	1.000	0.412	0.000
	Density	−0.833	0.412	1.000	0.389
	nMRD	−0.222	0.000	0.389	1.000
$n = 250$	Overlap	1.000	−0.329	−0.611	−0.556
	NN	−0.329	1.000	0.569	0.329
	Density	−0.611	0.569	1.000	0.722
	nMRD	−0.556	0.329	0.722	1.000
$n = 500$	Overlap	1.000	−0.588	−0.778	−0.611
	NN	−0.588	1.000	0.650	0.402
	Density	−0.778	0.650	1.000	0.722
	nMRD	−0.611	0.402	0.722	1.000

Table 2: The correlation between cluster hypothesis tests (measured in terms of Kendall's-τ). n is the number of documents in $\mathcal{D}_{\text{init}}$.

5. EXPERIMENTAL RESULTS

The correlations between the cluster hypothesis tests are presented in Table 2 for different values of n. With the exception of the Overlap test, we can see that the correlation between all other pairs of tests increases with increasing values of n, but can be negative or zero for low values of n. The Overlap test is negatively correlated with all the other tests across almost all values of n.

A decent positive correlation is attained between Density and NN for $n \geq 100$. For $n \geq 250$ a decent positive correlation is also attained between nMRD and NN. While nMRD is a global test that considers the relations between all the documents in $\mathcal{D}_{\text{init}}$, NN is a more local test that only considers the relations between a document and its nearest neighbors.

For $n \in \{250, 500\}$, nMRD and Density are the most correlated tests. This finding is surprising since these tests are based on completely different properties of the initial list. While nMRD is based on directly measuring inter-document similarities, the Density test is based on the number of unique terms in the documents which presumably attests to the ability to differentiate between relevant and non-relevant documents.

Cluster-based document retrieval methods. We next study the correlation between the cluster hypothesis tests and the relative effectiveness of cluster-based retrieval methods. For reference, we report the correlation numbers with respect to a ranking of the experimental settings induced by the size of the corresponding collections (**Size**). The results are presented in Table 3.

We observe a negative correlation between the Density test and all five cluster-based retrieval methods for $n = 50$. This finding can be explained as follows. First, the size of the collection is positively correlated with the number of terms in the vocabulary. (Refer back to Table 1.) Now, by definition, this number is negatively correlated with Density. Second, the size of the collection is positively correlated with the effectiveness of cluster-based retrieval methods as observed in Table 3 for the Size correlations for $n = 50$. We note that here, the Web collections are larger than the newswire collections and are in general noisier. Thus, we conclude that the type of the collection, i.e., Web vs. newswire, can have

		Overlap	NN	Density	nMRD	Size
$n = 50$	Interpf	0.761	−0.029	−0.648	0.028	0.725
	AMean	0.111	0.000	−0.333	0.278	0.286
	GMean	0.333	−0.229	−0.444	−0.056	0.457
	ClustRanker	0.778	−0.057	−0.667	0.167	0.743
	ClustMRF	0.889	−0.171	−0.778	0.167	0.857
$n = 100$	Interpf	0.500	0.000	−0.333	−0.167	0.400
	AMean	−0.222	−0.118	0.056	0.556	−0.114
	GMean	−0.333	−0.059	0.167	0.333	−0.229
	ClustRanker	0.611	0.059	−0.444	−0.278	0.514
	ClustMRF	0.722	−0.295	−0.556	−0.056	0.629
$n = 250$	Interpf	−0.254	0.486	0.254	0.028	−0.203
	AMean	−0.611	0.150	0.333	0.500	−0.400
	GMean	−0.333	0.210	0.389	0.556	−0.400
	ClustRanker	−0.056	0.150	0.000	−0.056	0.057
	ClustMRF	0.500	−0.210	−0.333	−0.278	0.400
$n = 500$	Interpf	−0.111	0.155	0.111	0.167	−0.057
	AMean	−0.444	0.340	0.444	0.500	−0.457
	GMean	−0.444	0.279	0.444	0.611	−0.457
	ClustRanker	−0.222	0.402	0.333	0.278	−0.286
	ClustMRF	0.389	−0.155	−0.167	0.000	0.229

Table 3: The correlation between cluster hypothesis tests and the relative effectiveness of cluster-based retrieval methods. The Size "test" ranks experimental settings by the number of documents in the collections. n is the number of documents in $\mathcal{D}_{\text{init}}$.

an influence on the effectiveness of cluster-based retrieval methods for short retrieved lists.

Another observation that we make based on Table 3 is that for $n = 50$ the correlation attained for the nMRD and NN tests is often lower than that attained for Overlap and Size. The relatively high positive correlation attained for the Overlap and the Size tests for $n = 50$ suggests that these tests are very strong indicators for the relative effectiveness of cluster-based retrieval with respect to document-based retrieval when applied to short retrieved lists. For larger values of n, NN and nMRD, as well as Density, start to post more positive correlations while the reverse holds for Overlap and Size.

We can also see that none of the retrieval methods is correlated only positively or only negatively with *all* the tests for any fixed value of n. In addition, only in a few cases a test is either only positively or only negatively correlated with all the retrieval methods for a fixed value of n. Thus, we conclude that the correlation between the effectiveness of a retrieval method and a cluster hypothesis test can substantially vary (both positively and negatively) across retrieval methods and tests.

6. CONCLUSIONS

We studied the correlation between cluster hypothesis tests and cluster-based retrieval effectiveness. We showed that the correlation between the two depends on the specific tests and methods that are used, and on the number of documents in the result list that is analyzed. We also showed that the type of the collection, i.e., Web or newswire, can be a stronger indicator for the relative effectiveness of cluster-based retrieval with respect to document-based retrieval, for short retrieved lists, than tests designed for estimating the extent to which the cluster hypothesis holds.

Acknowledgments We thank the reviewers for their comments. This work has been supported by and carried out at the Technion-Microsoft Electronic Commerce Research Center. This work has also been supported in part by Microsoft Research through its Ph.D. Scholarship Program.

7. REFERENCES

[1] J. Allan, M. E. Connell, W. B. Croft, F.-F. Feng, D. Fisher, and X. Li. INQUERY and TREC-9. In *Proceedings of TREC-9*, pages 551–562, 2000.

[2] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Informaltiom Retrieval Journal*, 14(5):441–465, 2011.

[3] A. El-Hamdouchi and P. Willett. Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*, 13:361–365, 1987.

[4] A. Griffiths, H. C. Luckhurst, and P. Willett. Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37(1):3–11, 1986.

[5] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971.

[6] O. Kurland. Re-ranking search results using language models of query-specific clusters. *Journal of Information Retrieval*, 12(4):437–460, August 2009.

[7] O. Kurland and E. Krikon. The opposite of smoothing: A language model approach to ranking query-specific document clusters. *Journal of Artificial Intelligence Research*, 41:367–395, 2011.

[8] O. Kurland and L. Lee. Clusters, language models, and ad hoc information retrieval. *ACM Transactions on information systems*, 27(3), 2009.

[9] O. Kurland and L. Lee. PageRank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on information systems*, 28(4):18, 2010.

[10] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of SIGIR*, pages 186–193, 2004.

[11] X. Liu and W. B. Croft. Experiments on retrieval of optimal clusters. Technical Report IR-478, Center for Intelligent Information Retrieval (CIIR), University of Massachusetts, 2006.

[12] X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *Proceedings of ECIR*, pages 454–462, 2008.

[13] S.-H. Na, I.-S. Kang, and J.-H. Lee. Revisit of nearest neighbor test for direct evaluation of inter-document similarities. In *Proceedings of ECIR*, pages 674–678, 2008.

[14] F. Raiber and O. Kurland. Exploring the cluster hypothesis, and cluster-based retrieval, over the web. In *Proceedings of CIKM*, pages 2507–2510, 2012.

[15] F. Raiber and O. Kurland. Ranking document clusters using markov random fields. In *Proceedings of SIGIR*, pages 333–342, 2013.

[16] J. Seo and W. B. Croft. Geometric representations for multiple documents. In *Proceedings of SIGIR*, pages 251–258, 2010.

[17] M. D. Smucker and J. Allan. A new measure of the cluster hypothesis. In *Proceedings of ICTIR*, pages 281–288, 2009.

[18] A. Tombros, R. Villa, and C. van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management*, 38(4):559–582, 2002.

[19] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, second edition, 1979.

[20] E. M. Voorhees. The cluster hypothesis revisited. In *Proceedings of SIGIR*, pages 188–196, 1985.

[21] P. Willett. Query specific automatic document classification. *International Forum on Information and Documentation*, 10(2):28–32, 1985.

[22] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334–342, 2001.