SAIC & University of Virginia at TREC 2005: HARD Track

Xiangyu Jin
University of Virginia
Charlottesville, VA 22903, USA
xiangyu@virginia.edu

James C. French
University of Virginia
Charlottesville, VA 22903, USA
french@virginia.edu

Jonathan Michel
Science Applications International Corporation(SAIC)
Charlottesville VA 22911, USA
Jonathan.D.Michel@saic.com

Abstract

SAIC (Science Applications International Corporation) and the University of Virginia collaborated to participate in the HARD (High Accuracy Retrieval from Documents) track of TREC 2005. Two clarification forms (CF) and 8 runs were submitted. The main focus of our work is to estimate the impact of incorrect user judgment on relevance feedback performance. The same set of documents are presented to the user to make judgments with different information shown on two CFs. Theoretically speaking if the user could make 100% accurate judgment, CF2 would perform much better than CF1. However, in practice, the user judgment accuracy is about 77.2% for CF1 and 65.4% for CF2. Thus results from feedback based on CF2 actually performs worse than CF1. This indicates possible unfairness when comparing relevance feedback techniques in a purely automatic evaluation environment where the user is assumed to be "perfect".

1 Introduction

Due to cost considerations, automatic machine judgment (judging according to the relevance information in pre-defined groundtruth) is usually employed for relevance feedback study instead of human subjects. By this method, the "user" is usually assumed to be perfect. Here perfect means the user never makes any mistake and her judgment is 100% accurate (consistent with the groundtruth). However, this is difficult to achieve in a practical retrieval environment. Due to practical constrains, the user interface can only deliver pieces of information to the user to judge. With such limited information, the user could make incorrect judgments.

Take document level relevance feedback as an example. Some documents are selected from the initial search results and presented to the user to judge their relevance. However, instead of presenting the full document to the user, we only show part of the information, such as title, keywords, abstract, etc. Since the abstract and title may not reveal the full content of the document, the user might judge the document differently from a purely automatic approach (where document are "judged" by the machine according to groundtruth directly). For example, we analyzed two clarification forms submitted by UIUC at HARD03 and found that the user judgement accuracy was around 80% for both clarification forms.

This imperfect user problem is usually neglected by current evaluation work for relevance feedback. The improvement is often exaggerated and comparison among relevance feedback approaches could be unfair. For example, we want to compare two relevance feedback approaches A and B. A is a classic document level feedback while B is a cluster-based feedback. The interface of A will show a list of documents for the user to judge but B will show some document groups for the user to judge. By assuming the user always make correct judgments, we draw a conclusion from an automatic evaluation that B's performance is much higher than A. However, in practice, the surrogates generated by B may be harder for the user to make correct judgments, hence it may result in even poorer performance.

Our major purpose for HARD05 is to analyze the impact of the user judgments on relevance feedback performance. HARD05 is very suitable for such task due to:

1. It is based on a large-scale document collection (1M).
2. Its topics are created or known to be difficult from prior TREC evaluations.
3. The CFs and the groundtruth of a specific topic are judged by the same assessor. We assume the assessors are domain experts and hence are consistent in making their judgments. By doing so, possible inconsistency caused by different human subjects is much reduced.

In our experiments, we first select 8 documents from the initial search results for each topic. Then we generate two different clarification forms for the same set of selected documents for a user to make judgments. These two CFs are based on techniques called feedback by "present" and feedback by "future", respectively. Since CF2 carries more information, it should perform better than CF1 if the user can make correct judgments. However, we find their performances in HARD05 is similar and later experiments find that CF2 performs even worse since its relevance questions are harder to answer.

In the rest of this paper, we describe the two feedback approaches in section 2. Section 3 gives our experimental settings. Section 4 shows and analyzes the results we submit for HARD05 and some supplemental runs. Finally, we conclude the work in section 5.

2 Feedback by "present" vs. feedback by "future"

The classic cluster hypothesis state: closely associated documents tend to be relevant to the same queries [1]. Therefore, if the given document is relevant, its neighboring documents in the retrieval space tend to be relevant as well. The feedback by "present" technique is based on this assumption. The document is only judged according to its content and has nothing to do with its local environment. If it is relevant, we tend to move the query region toward this area so that more relevant documents would be covered.

However, the relevant and irrelevant documents are usually not perfectly clustered in the retrieval space. There are cases where highly relevant documents' neighbors are mostly irrelevant. Take topic 314 "Marine Vegetation" as an example, document NYT19990330.0106 (TREC ID) is highly relevant and document APW20000404.0176 is irrelevant. However, the latter one's nearest neighbors are much more "relevant" than the former one's. In this case, feedback by "present" approach would perform poorly, since moving the query region toward these areas would be disastrous. The user just blindly makes her choice, for she has no idea what might happen if the document is judged as "relevant" or "irrelevant". Feedback by "future" is a possible solution to this problem. In this technique, instead of showing the document itself to the user, an abstraction of the neighborhood of this document is shown to the user to make corresponding judgment. Note, this approach is different from cluster-based feedback. By cluster-based feedback, a group of documents is judged as a whole and all participates in the later refinement process. By feedback-by-future, although multiple documents' information is employed to assist the judgment, only one document will participate in the refinement. That is, the refinement algorithm for both strategies could be the same, while the same document could be judged differently under each strategy. Documents whose neighborhoods are composed mostly of relevant documents are judged as "relevant" this time. This approach lets the user anticipate what documents might be gotten by her action, as if the user can see the "future". Obviously, this requires more information to be delivered to the user. When the user effort is limited, the provided relevance questions are more difficult to answer.

3 Experimental Settings

3.1 Corpus, Queries, and Search Engine

In HARD05, the whole AQUAINT collection is used as the corpus. This document collection contains about 1M documents and on average each document contains 425 terms. All the documents are indexed with Porter stemming.

Terms from each TREC topic (include title+desc+narr) are ranked by their $tf * idf$ scores. Here tf means the frequency of terms in the topic and *idf* is the inverse document frequency of the same term from TREC-3 topics (topic 151-200, each topic is treated like a document). Here we use TREC-3's topic *idf* instead of HARD05's due to the requirement that we cannot use any information in other topics to improve current search. The purpose of using topic *idf* is to decrease the weight for high frequency common terms like “document,” “describe,” “relevant,” etc. The top 20 terms are selected to form a vector query, with their normalized score as the term weights.

We use a BM25 ranking retrieval system to generate the search result. The performance of this system is similar to Okapi in TREC-3 testbed [2].

3.2 Baseline Runs

We submit two baseline runs, SAICBASE1 and SAICBASE2. SAICBASE1 use previous settings to generate the search result. SAICBASE2 is a pseudo-relevance-feedback version of SAICBASE1, where 40 terms of each top document (30 documents gotten from SAICBASE1) are employed to expands the query.

3.3 Clarification Forms

In order to perform fair comparisons, clarification forms are generated from the same selected document set. Duplicated documents are filtered from the result list [1] and 8 documents are selected by method similar to the gapped selection described in [3]. The reason for such gapped selection is to increase the possible diversity in the candidate feedback documents.

Two clarification forms are provided. CF1 (SAIC1) lets the user make judgments based on a document’s content. Each document’s title, source, creation time, and its abstraction are shown for the user to make judgment. Possible judgment choice are “relevant,” “non-relevant,” and “perhaps.” The default judging for a document is “perhaps.” In order to let the user concentrate more on the current document, we use a dynamic interface to show the corresponding abstraction. The clarification form is split into two areas. The left area gives a list of 8 documents with their title, source, creation time, and relevance judgment choices shown. The right area gives the corresponding abstraction for the current document. When the user moves the mouse over any document in the left area, the right area will be dynamically updated with the current document’s abstraction. The abstraction is less than 70 terms and within 3 sentences. With key terms from the topic and top terms from this document emphasized in different color. This eye catching technique can direct the user to read the abstraction more efficiently. The emphasized terms give the user an intuitive feeling for how the retrieval system determined that the given document matches the topic. Figure 1 gives an example of CF1 on topic 307.

CF2 (SAIC2) lets the user make judgments based on a document’s neighborhood information. This is motivated by the fact that sometimes a document’s relevance information does not imply its neighbor’s relevance. As we described before, a highly relevant document may exist in an irrelevant document cluster. If it is issued as a positive feedback example, it might get a lot of irrelevant documents into the refined search result. CF2 is based on feedback by future technique and the user is directed to make judgment by “What might happen if I feed back this document?” Each document is abstracted by its top 20 terms and forms a query to search. The abstraction of the top 8 documents (including the original document) is shown. We still keep the abstraction less than 70 terms and within 3 sentences thus making the user judging effort similar to CF1. This time, instead of showing document title, source, and creation time on the left, the top five key terms from the document group are shown. Those terms that duplicate with topic key terms are excluded, so that the given five fresh terms will show us what is “more emphasized” inside this document group. Figure 2 gives an example of CF2 on topic 307.

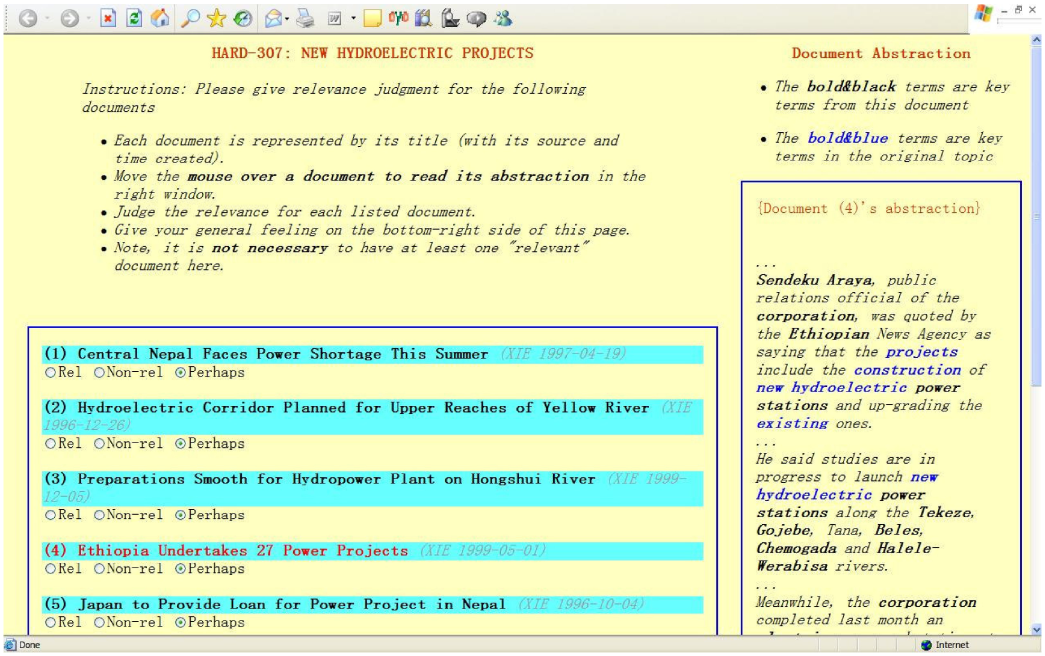
Moreover, on both CFs, we offer a choice of user’s feeling about the initial retrieval’s effect. Unfortunately, this choice seems not quite useful since most topics are selected as “some are relevant”.

3.4 Final Runs

The system can acquire the relevance judgment from CF1 or CF2. We also merge the refined search results of CF1 and CF2, with equal weighted mid-rank merge [4]. Thus we have three possible relevance judgment source: CF1, CF2, and CF1&2.

The classic Rocchio method [5] is used to refine the search result. Weights for initial query, positive documents, negative documents are set to be 1.0, 0.8, and 0.6, respectively. We also introduce a technique called conservative feedback, which modifies the terms’ weights in the initial query conservatively. We

[1] We found there are many pair-wise duplicated documents. They are almost exactly the same in content since one document is merely an editor approved version of the other.



HARD-307: NEW HYDROELECTRIC PROJECTS

Instructions: Please give relevance judgment for the following documents

- *Each document is represented by its title (with its source and time created).*
- *Move the* ***mouse over a document to read its abstraction*** *in the right window.*
- *Judge the relevance for each listed document.*
- *Give your general feeling on the bottom-right side of this page.*
- *Note, it is* ***not necessary*** *to have at least one "relevant" document here.*

(1) Central Nepal Faces Power Shortage This Summer *(XIE 1997-04-19)*
○Rel ○Non-rel ●Perhaps

(2) Hydroelectric Corridor Planned for Upper Reaches of Yellow River *(XIE 1996-12-26)*
○Rel ○Non-rel ●Perhaps

(3) Preparations Smooth for Hydropower Plant on Hongshui River *(XIE 1999-12-05)*
○Rel ○Non-rel ●Perhaps

(4) Ethiopia Undertakes 27 Power Projects *(XIE 1999-05-01)*
○Rel ○Non-rel ●Perhaps

(5) Japan to Provide Loan for Power Project in Nepal *(XIE 1996-10-04)*
○Rel ○Non-rel ●Perhaps

Document Abstraction

- *The* ***bold&black*** *terms are key terms from this document*
- *The* ***bold&blue*** *terms are key terms in the original topic*

{Document (4)'s abstraction}

...
Sendeku Araya*, public relations official of the* ***corporation****, was quoted by the* ***Ethiopian*** *News Agency as saying that the* ***projects*** *include the* ***construction*** *of* ***new hydroelectric power stations*** *and up-grading the* ***existing*** *ones.*
...
He said studies are in progress to launch ***new hydroelectric power stations*** *along the* ***Tekeze****,* ***Gojebe****, Tana,* ***Beles****,* ***Chemogada*** *and* ***Halele-Werabisa*** *rivers.*
...
Meanwhile, the ***corporation*** *completed last month an*

Figure 1: Example of CF1 on topic 307.

Table 1: Final Run Settings

Relevance Judgment Source	CF1	CF2	CF1&2
Standard Rocchio method	SAICFINAL1	SAICFINAL2	SAICFINAL5
Conservative Rocchio method	SAICFINAL3	SAICFINAL4	SAICFINAL6

use a conservative factor vary from 0 to 1 (the default value is 0.2) to control the weight modification on initial query terms, where 0 makes no modification on the initial query terms and 1 is the standard Rocchio method. This is motivated by the fact that TREC topics are carefully designed and the initial query contains many good terms. We do not want later negative feedback examples affect the selection of these useful terms. Take topic 303 "HUBBLE TELESCOPE ACHIEVEMENTS" as an example. The top search results are mostly about repairing the Hubble telescope using space shuttles and hence are not considered relevant. If they are issued as negative feedback examples and standard Rocchio method is used, we will not only degrade the weight for "astronaut" and "space shuttle", but also "Hubble" and "telescope" as well. By conservative method we only decrease the importance of "astronaut" and "space shuttle" but not much for "Hubble" and "telescope". Moreover, it is still possible that a term receives negative weight. We simply discard these negative weighted terms from the query.

We submitted 6 final runs, named SAICFINAL1 to SAICFINAL6. The detailed settings are listed in Table 1.

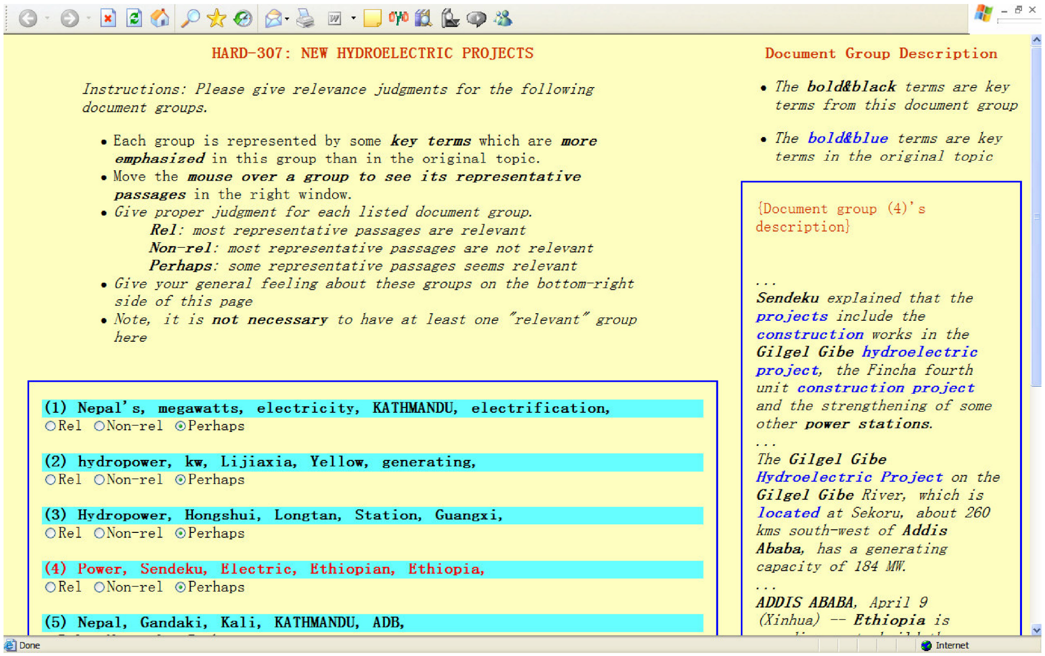
4 Result Analysis

4.1 Relevance Judging Accuracy

Our first attempt is to analyze how accurately the assessor makes judgments. For CF1, the assessor's judgment is evaluated against the groundtruth. For CF2, it is trickier since we do not have an absolute rule for what is "relevant." We use the following method to build up a "groundtruth" for CF2.

1. For a given topic, get the initial search's performance (R-precision is used) p_{init}.

2. Issue each candidate feedback document as a query (extract top 20 terms to form the query) to search in the document collection. Get its performance as p_{seed}.



HARD-307: NEW HYDROELECTRIC PROJECTS

Instructions: Please give relevance judgments for the following document groups.

- Each group is represented by some ***key terms*** which are ***more emphasized*** in this group than in the original topic.
- Move the ***mouse over a group to see its representative passages*** in the right window.
- *Give proper judgment for each listed document group.*
  Rel*: most representative passages are relevant*
  Non-rel*: most representative passages are not relevant*
  Perhaps*: some representative passages seems relevant*
- *Give your general feeling about these groups on the bottom-right side of this page*
- *Note, it is* ***not necessary*** *to have at least one "relevant" group here*

(1) Nepal's, megawatts, electricity, KATHMANDU, electrification,
○Rel ○Non-rel ⦿Perhaps

(2) hydropower, kw, Lijiaxia, Yellow, generating,
○Rel ○Non-rel ⦿Perhaps

(3) Hydropower, Hongshui, Longtan, Station, Guangxi,
○Rel ○Non-rel ⦿Perhaps

(4) Power, Sendeku, Electric, Ethiopian, Ethiopia,
○Rel ○Non-rel ⦿Perhaps

(5) Nepal, Gandaki, Kali, KATHMANDU, ADB,

Document Group Description

- *The* ***bold&black*** *terms are key terms from this document group*
- *The* ***bold&blue*** *terms are key terms in the original topic*

{Document group (4)'s description}

...
Sendeku *explained that the* ***projects*** *include the* ***construction*** *works in the* ***Gilgel Gibe hydroelectric project****, the Fincha fourth unit* ***construction project*** *and the strengthening of some other* ***power stations****.*
...
The ***Gilgel Gibe Hydroelectric Project*** *on the* ***Gilgel Gibe*** *River, which is* ***located*** *at Sekoru, about 260 kms south-west of* ***Addis Ababa****, has a generating capacity of 184 MW.*
...
ADDIS ABABA*, April 9 (Xinhua) --* ***Ethiopia*** *is*

Figure 2: Example of CF2 on topic 307.

3. If $p_{seed}/p_{init} > \theta$ (where θ is a threshold), this candidate feedback document would be treated as "relevant", otherwise it would be treated as "irrelevant".

We expect the two groundtruth have similar numbers of relevant and irrelevant documents (for the 400 candidate feedback documents in total). This is because if both CFs are judged by oracles (who know what the correct answer is), we would feed back a similar number of relevant and irrelevant documents so their performance can be compared fairly. That is, given the same number of documents, both select similar number of relevant and irrelevant documents to feed back to determine which one could do a better job. Here we set the threshold θ to be 0.5, so that CF2 has similar number of relevant and irrelevant documents (182/218) to CF1 (189/211) of all 50 topics. Table 2 gives the judging accuracy for two clarification forms. Excluding the non-judged, the judging precision is about 77.2% (similar to ILUC at HARD03) for CF1 and 65.4% for CF2. Obviously, CF2's judging accuracy is lower.

Table 2: Judging Correctness

CFs	Correct Judged	Incorrect Judged	Non-judged
CF1	0.5925	0.175	0.2325
CF2	0.525	0.2775	0.1975

Our second attempt is to analyze by different clarification forms to see how differently the user judgments would be. We make this analysis in two steps. First, we analyze the difference if both CFs are judged by oracles. Second, we analyze the difference if both CFs are judged by human subjects. Here, oracle has two possible judging results, relevant or irrelevant. Human subjects have three possible judging results, relevant, irrelevant, or non-judged. Table 3 and Table 4 gives the judgment difference of the two CFs between an oracle and human subjects. From Table 3, the judging consistent rate is 72.3%. This reveals the cluster hypothesis is not absolutely correct. Many documents which are considered relevant by CF1 actually do not have many relevant neighbors. From Table 4, human subjects judgment agrees 77.5% between the two CFs (only for the judged documents).

4.2 Results for HARD Runs

Compare with other participants, our baseline run is pretty good among all submissions. Pseudo-relevance-feedback shows some improvement in SAICBASE2. Unfortunately, all our final runs fail to

Table 3: Judging Difference (Oracle)

	Rel (CF1)	non-Rel (CF1)
Rel (CF2)	130	52
non-Rel (CF2)	59	159

Table 4: Judging Difference (Human Subject)

	Rel (CF1)	non-Rel (CF1)
Rel (CF2)	71	21
non-Rel (CF2)	38	132

show any improvement over our baseline. And their relations are not clear, too. A possible reason is the hardness of HARD makes many selected feedback documents are actually irrelevant, thus extensive negative feedback exist in the refinement process. Although negative feedback helps some topic such as 303, it degrades the performance in general. At present time we do not have a good solution to handle negative feedback. Later we exclude negative feedback and only employ positive feedback in our experiments. The results are shown in next section.

Table 5: HARD Runs Results

Run Name	R-Precision	Avg-Precision
Baseline (averaged median of all runs)	0.2518	0.1901
Final (averaged median of all runs)	0.2639	0.2070
SAICBASE1	0.3021	0.2435
SAICBASE2	0.3152	0.2876
SAICFINAL1	0.2826	0.2415
SAICFINAL2	0.2657	0.2265
SAICFINAL3	0.2881	0.2488
SAICFINAL4	0.2806	0.2391
SAICFINAL5	0.2636	0.2343
SAICFINAL6	0.2814	0.2469

4.3 Supplemental Results

In this experiment, we only feedback positive examples but exclude negative ones. We compared the system performance under oracle judgment and human judgment for two CFs. We also compare it with a "blind" strategy where pseudo-relevance-feedback is employed. The results are quite exciting.

First, we confirm that the reason why we do not have improved search in HARD final runs is that we do not handle negative feedback examples well. Under the new settings, performance after relevance feedback is much improved.

Second, we see that the oracle can do a much better job than human. For both CF1 and CF2, the oracle refined search has much higher performance than human refined search. This indicates the 23% judging error rate matters a lot. If automatic evaluation is used, the performance improvement would be exaggerated. More interestingly, we find that even feeding them back blindly (treat them all as relevant) would receive similar performance as human subject's judgment.

Third, we find that although CF2 performs better than CF1 from automatic evaluation (the oracle), it actually performs worse in practice. This is because CF2 is much harder for the user to make correct selection and its error rate reaches 35%. This further confirm our worry that unfair comparisons may exist in current literatures for relevance feedback study.

Table 6: Supplemental Results

Run Name	R-Precision	Avg-Precision
CF1 Oracle	0.3637	0.3510
CF2 Oracle	0.3945	0.3667
CF1 Human	0.3266	0.2947
CF2 Human	0.3128	0.2790
Blind	0.3255	0.2907

5 Conclusion

We find that real user cannot achieve perfect judging consistency in relevance feedback task. This problem is the bottleneck for practical application of relevance feedback techniques. Further analysis is needed for better understanding of the data.

References

[1] CJ Rijsbergern. *Information Retrieval (2nd Edition).* Butterworths, London, 1979.

[2] X Jin, JC French, and J Michel. Query formulation for information retrieval by intelligence analysts. In *Tech. report CS-2005-12, Dept. of Computer Science, Univ. of Virginia*, 2005.

[3] X Shen and C Zhai. Active feedback in ad hoc information retrieval. In *SIGIR '05*, pages 59–66, 2005.

[4] J French, J Watson, X Jin, and W Martin. Using multiple image representations to improve the quality of content-based image retrieval. In *Tech. report CS-2003-10, Dept. of Computer Science, Univ. of Virginia*, 2003.

[5] JJ Rocchio. Relevance feedback in information retrieval. In G Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, 1971.