WIDIT: Integrated Approach to HARD Topic Search

Kiduk Yang, Ning Yu, Hui Zhang, Shahrier Akram, and Ivan Record

School of Library and Information Science, Indiana University, Bloomington,
Indiana 47405, U.S.A.
{kiyang, nyu, hz3, sakram, irecord}@indiana.edu

1 Introduction

Web Information Discovery Tool (WIDIT) Laboratory at the Indiana University School of Library, whose basic approach to combine multiple methods as well as to leverage multiple sources of evidence, participated in 2005 Text Retrieval Conference's Hard track (HARD-2005) to investigate methods of effectively dealing with HARD topics by exploring a variety of query expansion strategies, the results of which were combined via an automatic fusion optimization process. We hypothesized that the "difficulty" of topics is often due to the lack of appropriate query terms and/or misguided emphasis on non-pivotal query terms by the system. Thus, our first-tier solution was to devise a wide range of query expansion methods that can not only enrich the query with useful term additions but also identify important query terms. Our automatic query expansion included such techniques as noun phrase extraction, synonym identification, definition term extraction, keyword extraction by overlapping sliding window, and Web query expansion. The results of automatic expansion were used in soliciting user feedback, which was utilized in a post-retrieval reranking process. The paper describes our participation in HARD-2005 and is organized as follows. Section 2 gives an overview of HARD track, section 3 describes the WIDIT approach to HARD-2005, and section 4 discusses the results and implications, followed by the concluding remarks in section 5.

2 HARD-2005 Overview

The goal of the TREC's HARD track in 2005 was to achieve "high accuracy retrieval from documents" (i.e. improved retrieval performance at top ranks) against a set of difficult topics using targeted interactions with the searcher[1]. The document collection used in HARD-2005 is the AQUAINT corpus from Linguistic Data Consortium (LDC), and the topic set for HARD-2005 consisted of 50 "difficult" topics, which were selected from a pool of past TREC topics that most systems performed poorly on. HARD participants were first to submit for each topic an automated retrieval result (i.e., Baseline Run) along with an HTML form with which to collect user feedback (i.e., Clarification Form). The Clarification Forms (CF) were then filled out by TREC assessors and the results were sent back to TREC participants to be utilized to improve retrieval performance in the subsequent submissions (Final Run).

[1] See TREC 2005 HARD Track Guidelines at http://ciir.cs.umass.edu/research/hard/guidelines.html

H.T. Ng et al. (Eds.): AIRS 2006, LNCS 4182, pp. 649 – 658, 2006.

The main question in HARD-2005, with its “difficult” topic set and CF mechanism, is how user feedback can improve retrieval performance of difficult topics. At a first glance, one may be inclined to suggest as a solution *relevance feedback*, which is a well-known user feedback mechanism in IR (Rocchio, 1971; Salton & Buckley, 1990). Since difficult topics tend to retrieve non-relevant documents at top ranks, however, it will be difficult for the user to evaluate enough retrieval results in a short time period to find relevant documents with which to deploy relevance feedback.

3 WIDIT Approach to HARD-2005

We hypothesized that the “difficulty” of topics is often due to the lack of appropriate query terms and/or misguided emphasis on non-pivotal query terms by the system. Thus, our first-tier solution was to devise a wide range of query expansion methods that can not only enrich the query with useful term additions but also identify important query terms. Our automatic query expansion included such techniques as noun phrase extraction, synonym identification, definition term extraction, keyword extraction by overlapping sliding window, and Web query expansion. The results of automatic expansion were used in soliciting user feedback, which was utilized in a post-retrieval reranking process.

For synonym identification, we integrated a sense disambiguation module into WIDIT’s synset identification module so that best synonym set can be selected according to the term context. To reduce the noise from word definitions, we applied the overlapping sliding window (OSW) method to multiple definitions harvested from web and extracted the overlapping terms. To extract noun phrase, we combined the results of multiple NLP taggers as well as applying the OSW method. OSW method was also applied to topic fields to identify important topic terms. The Web query expansion method was a slight modification of the PIRC approach (Grunfeld et al., 2004; Kwok et al., 2005).

To produce the optimum baseline results, we merged various combinations of query formulation results and the query expansion results using a weighted sum fusion formula. The fusion weights were determined using previous year’s Robust data to train the system via an automatic fusion optimization process, where best performing systems in selected categories (e.g., short query, top 10 systems, etc.) were combined using average precision as fusion weights until the performance gain fell below a threshold.

We viewed the clarification form as both manual query expansion and relevance feedback mechanism. Our clarification form included query term synonyms, noun phrase, and best sentences from top documents of the baseline result. Since difficult topics tend to produce few relevant documents in top ranks, we clustered the results and selected the best sentence from each cluster to include in the CF. In addition to expanding the query with user-selected terms from the clarification form, we also utilized the user’s best sentence selection by boosting the rank of the documents in which selected sentences occurred.

WIDIT HARD system consists of five main modules: indexing, retrieval, fusion (i.e. result merging), reranking, and query expansion modules. The overview of WIDIT HARD system architecture is displayed in Figure 1.

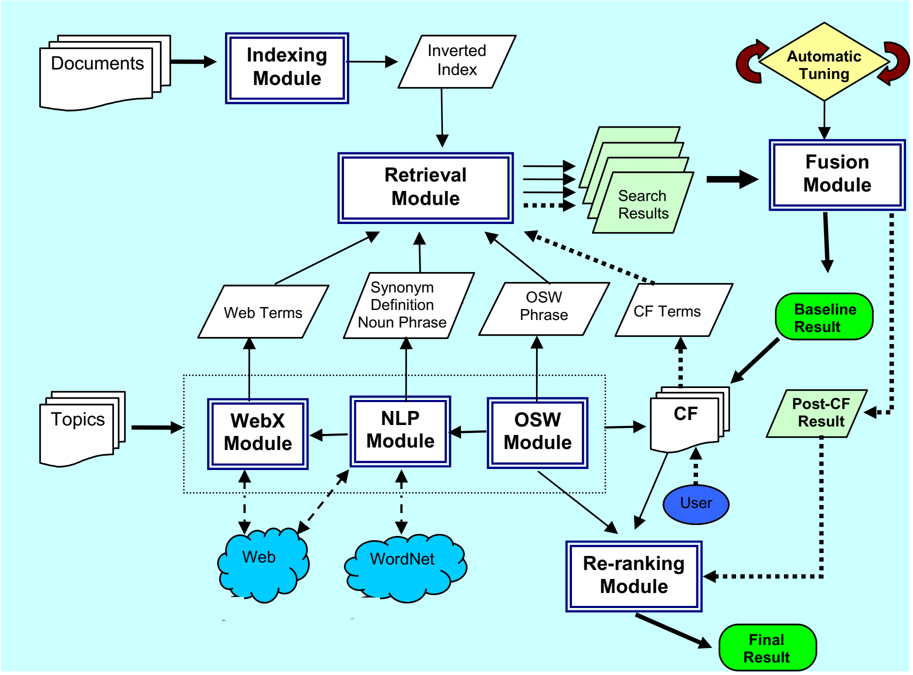

Fig. 1. WIDIT HARD System Architecture

3.1 Query Expansion

The query expansion module consists of three submodules: Web expansion (WebX) module expands the query with terms from Google search results; NLP module finds synonyms and definitions of query terms and identifies nouns and noun phrases in the query; Overlapping Sliding Window (OSW) module extract key phrases from the query.

3.1.1 OSW Module

The main function of OSW module is to identify important phrases. OSW method, which is based on the assumption that phrases appearing in multiple fields/sources tend to be important, works as follows:

i. Define window size and the number or maximum words allowed between words.
ii. Slide window from the first word in the source field/source. For each of the phrase it catches, look for the same/similar phrase in the search fields/sources.
iii. Produce the OSW phrases
iv. Change source field/source and repeat step 1 to 3 till all the fields/sources have been used.

OSW method was applied to topic descriptions and query term definitions to identify key phrases.

3.1.2 NLP Module

WIDIT's NLP module expands acronyms using a Web-harvested acronym list, identifies nouns and noun phrases using multiple taggers, and finds synonyms and definitions via querying the Web. Two main objectives of the NLP module refinement this year were to reduce noise in query expansion and to identify key (i.e. "important") phrases. For noise reduction, we integrated a sense disambiguation[2] into WIDIT's synset identification module so that best synonym set can be identified based on the term context, and refined the WIDIT's definition module by applying OSW to extract overlapping terms from the multiple Web-harvested definitions (*WordIQ*, *Dictionary.com*, *Google*, *Answers.com*). For key phrase identification, we used a combination of NLP tools as well as WordNet to identify 4 types of noun phrases: proper names, dictionary phrases, simple phrases, complex phrases.

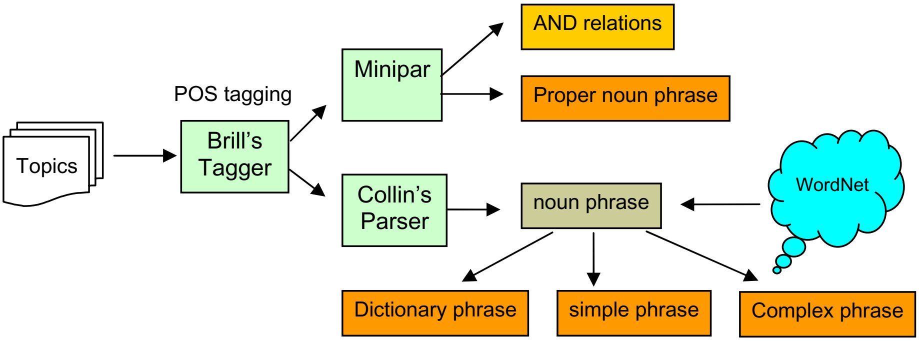

Fig. 2. Noun Phrase Identification Diagram

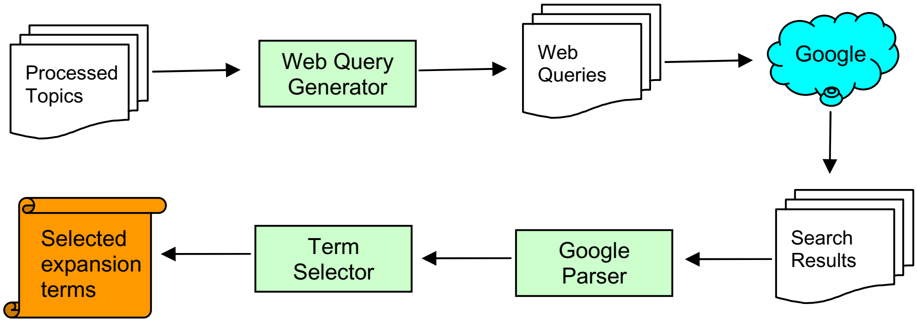

Fig. 3. WebX Module Architecture

3.1.3 WebX Module

The PIRCS group has demonstrated that web expansion is an "effective" method for improving the performance of weak (i.e. difficult) topics (Grunfeld et al., 2004; Kwok et al., 2005). WebX module, which is based on the PIRC approach, expands the query

[2] WordNet word sense disambiguation software developed by the Natural Language Processing Group at the University of Minnesota, Duluth.

with related terms harvested from Google search results. WebX module consists of Web query construction, Web search, search result parsing and term selection (Figure 3). The Web query generator constructs Web queries by selecting up to 5 most salient terms from the processed HARD topics (i.e., stopped and stemmed text, nouns, phrases). The queries are then sent to Google, and subsequent search results (the snippets and the body texts) are parsed to extract up to 60 terms per topic to be used as query expansion terms.

3.2 Fusion

The fusion module combines the multiple sets of search results after retrieval time. In our earlier study (Yang, 2002b), similarity merge approach proved ineffective when combining content- and link-based results, so we devised three variations of the weighted sum fusion formula, which were shown to be more effective in combining fusion components that are dissimilar (Yang, 2002a). Equation (4) describes the simple *Weight Sum* (WS) formula, which sums the normalized system scores multiplied by system contribution weights.:

$$FS_{WS} = \sum(w_i * NS_i), \tag{4}$$

where: FS = fusion score of a document,
w_i = weight of system i,
NS_i = normalized score of a document by system i,
$= (S_i - S_{min}) / (S_{max} - S_{min})$

One of the main challenges in using the weighted fusion formula lies in determination of the optimum weights for each system (w_i). Last year, we devised a novel man-machine hybrid approach called the *Dynamic Tuning* to tune the fusion formula (Yang, Yu, & Lee, 2005; Yang & Yu, 2005). This year, we devised another alternative fusion weight determination method called *Automatic Fusion Optimization by Category* (AFOC). AFOC involves iterations of fusion runs (i.e., result merging), where best performing systems in selected categories (e.g., short query, top 10 systems, etc.) are combined using average precision as fusion weights until the performance gain falls below a threshold. The current AFOC implementation does not guarantee true optimization since the process will stop when a local optimum is encountered. Figure 4 illustrates the automatic fusion optimization process.

3.3 Clairfication Form

The main objective of our CF design strategy was to obtain accurate feedback that validates and supplements the system efforts for dealing with difficult topics (e.g. query expansion). Consequently, our Clarification Forms served as manual query expansion and relevance feedback mechanism, which included such components as the selection from candidate expansion terms and phrase, the validation of BoolAnd relations. In addition to displaying important phrases and best sentences from top 200 retrieved documents[3], our CF included synonym sets, definition terms, and query term relations with the use of JavaScript to make the interaction more friendly and

[3] Since weak topics tend to retrieve non-relevant documents at top ranks, we clustered the top 200 documents and selected the best sentence from each cluster.

efficient. The CF terms selected by the user was used to create a CF-expanded query. Phrases, BoolAnd terms and relevant documents identified in CF were used by the reranking module to boosts the ranks of documents with important phrases and relevant documents identified by the user.

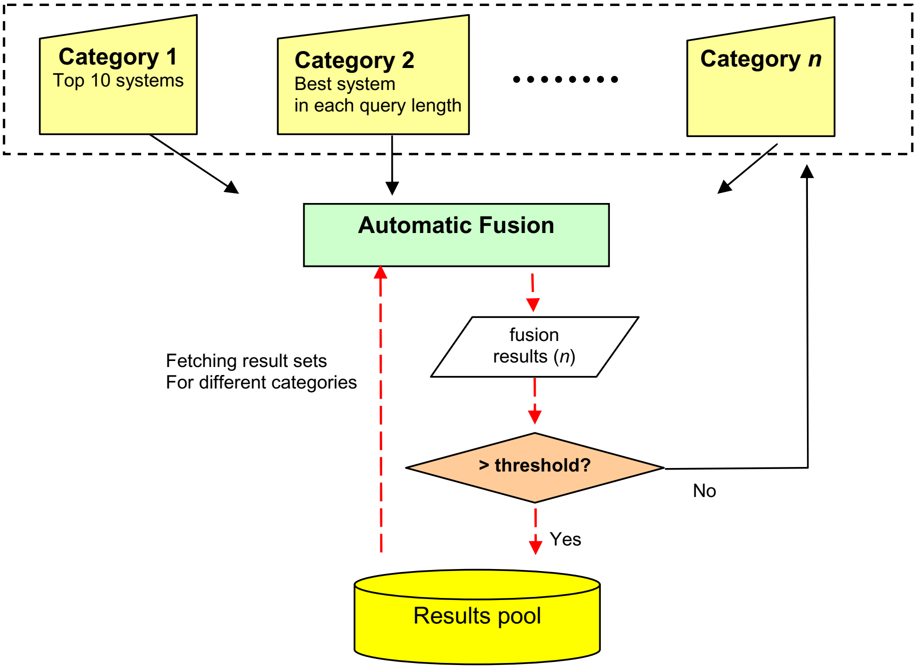

Fig. 4. Automatic Fusion Optimization by Category

3.4 Reranking

The objective of reranking is to float low ranking relevant documents to the top ranks based on post-retrieval analysis of reranking factors. After identifying reranking factors such as *OSW terms*, *CF terms*, and *CF-reldocs*, which are relevant documents identified in CF form, we computed the reranking factor scores (*rf_sc*) for top *k* documents and boosted the ranks of documents with *rf_sc* above a threshold score above a fixed rank R using the following formula:

$$\text{doc\_score} = rf\_sc + \text{doc\_score@rankR} \quad (5)$$

Although reranking does not retrieve any new relevant documents (i.e. no recall improvement), it can produce high precision improvement via post-retrieval compensation (e.g. phrase matching) or force rank-boosting to accommodate trusted information (e.g. *CF-reldocs*).

4 Results

Web query expansion (*WebX*) was the most effective method of all the query expansion methods. Figure 5, which shows Web query expansion results by query length, plots the retrieval performance gain (indicated by the bar above the zero line) or loss (indicated by the bar below the zero line) of various *WebX* methods over non-expansion query results. As can be seen in the figure, *WebX* showed most gain in performance for short queries (i.e. title) but had an adverse effect for longer queries (i.e. description) except when using the rotating window approach (blue, green, and yellow bars).

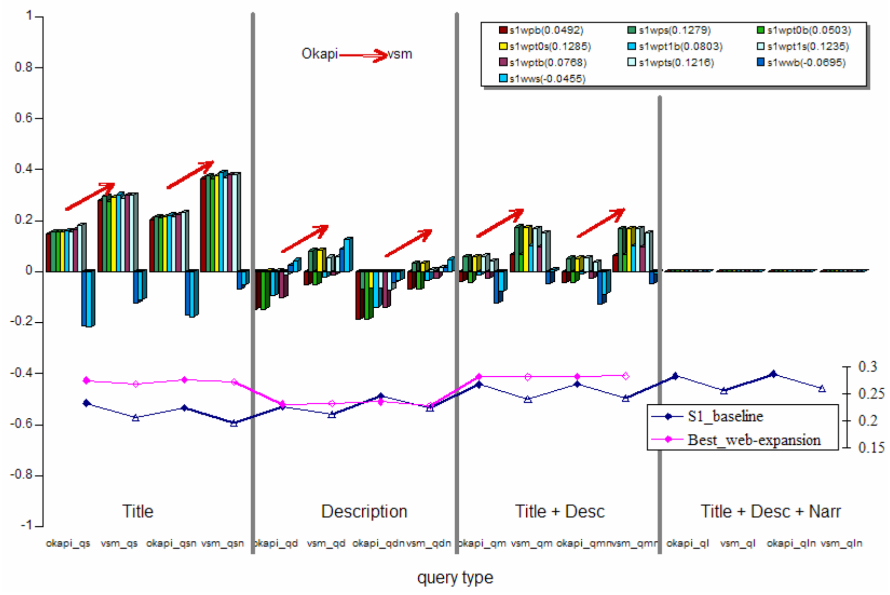

Fig. 5. Web Query Expansion Effect

Among the non-*WebX* query expansion methods, *Proper Noun Phrases*, *OSW Phrases*, and *CF Terms* helped retrieval performance for longer queries, although the rate of performance gains fall much below *WebX* methods (Figure 6). It is interesting to note that the effect of query expansion is influenced by the query length in an opposite manner between *WebX* and non-*WebX* methods. Without query expansion, longer queries usually outperform the shorter queries. With query expansion, however, query length has opposite effect on *WebX* and non-*WebX* methods (i.e., WebX methods works well with short queries, whereas non-WebX methods works better with longer queries). The composite effects of query expansion and query length suggest that *WebX* should be applied to short queries, which contain less noise that can be exaggerated by Web expansion, and non- *WebX* should be applied to longer queries, which contain more information that query expansion methods can leverage.

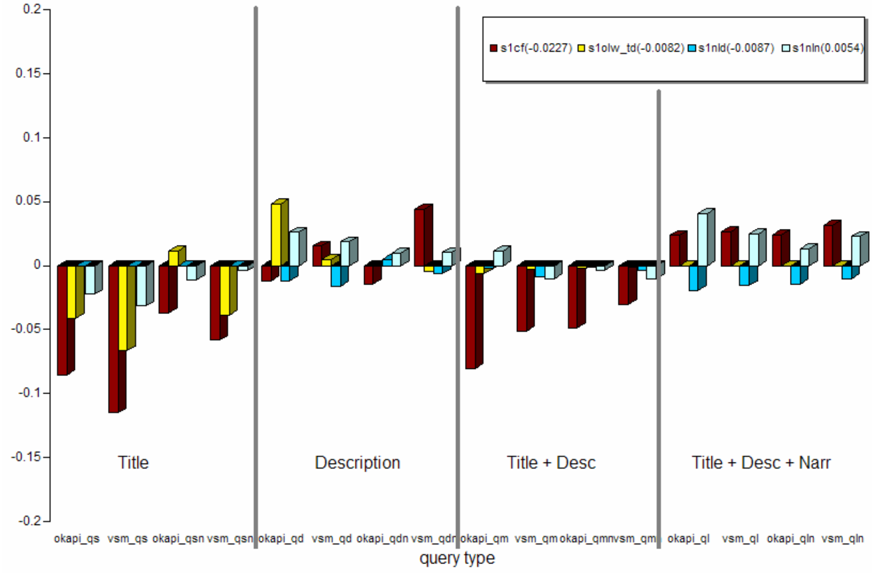

Fig. 6. Non-Web Query Expansion Effect

Fusion (i.e. result merging) improved the retrieval performance across the board with almost 50% improvement in mean average precision for short queries, showing that *Automatic Fusion Optimization by Category* is a viable method to streamline the process of combining numerous result sets in an efficient manner (Table 1). We attribute the lower fusion performance gain by MRP to the fact that *AFOC* used MAP in tuning the fusion formula.

Table 1. Fusion Effect

	Mean Avg. Precision	**Mean R-Precision**
Baseline Title Run	0.1694	0.2416
Baseline Description Run	0.1698	0.2395
Baseline Fusion Run	0.2324	0.2961
Final Title Run	0.2513 (+48%)	0.3020 (+25%)
Final Description Run	0.2062 (+21%)	0.3020 (+10%)
Final Fusion Run	0.2918 (+25%)	0.3442 (+16%)

Figure 7 shows the effect of reranking factors. The main reranking factors in Figure 7 are *OSW phrases* (*O* in run labels: e.g., R*O*DXX), *CF terms* (C in run labels: e.g., R*C*DXX), and *CF-reldocs* (D in run labels: e.g., RO*D*XX, RC*D*XX). Examination of top reranking systems suggests that CF-relevant documents has the most positive effect on retrieval, followed by *OSW* and *CF Terms*.

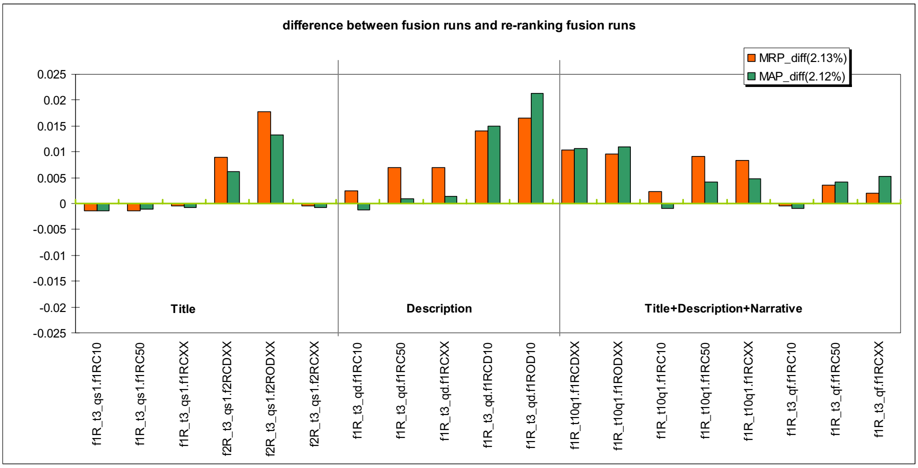

Fig. 7. Reranking Effect

5 Concluding Remarks

We investigated an integrated approach to HARD topic search. To address two major characteristics of difficult topics, we first devised a variety of automatic query expansion methods designed specifically to identify important query terms and phrases as well as to discover new query terms. The automatic query expansion methods, which generated candidate query expansion terms while improving baseline search results, seeded the content of the clarification form, where the system approach to HARD topic search was validated by the user and fed back into the system to further improve the retrieval performance. In keeping with the WIDIT philosophy of fusion, which is capture in the statement "the whole is greater than sum of its parts", we examined the effects of fusion throughout our investigation and found combining of not only data (e.g., short and long queries) and methods (e.g., query expansion methods) but also the system and the user (e.g., construction and utilization of the clarification form) to be quite beneficial. Furthermore, the study revealed the different behaviors by Web-based and non-Web-based query expansion methods when compounded with query length, which suggests the desirability of a flexible system that dynamically adapt its strategies to given situations over a static system. Finally, we devised an effective automatic fusion optimization process that can be deployed in situations where there are too many fusion components to be combined and tuned manually.

References

Buckley, C., Salton, G., & Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. *Proceeding of the 3rd Text Rerieval Conference (TREC-3)*, 1-19.

Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC 5. *Proceeding of the 5th Text REtrieval Conference (TREC-5)*, 105-118.

Fox, E. A., & Shaw, J. A. (1995). Combination of multiple searches. *Proceeding of the3rd Text Rerieval Conference (TREC-3)*, 105-108.

Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures & algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Grunfeld, L., Kwok, K.L., Dinstl, N., & Deng, P. (2004). TREC 2003 Robust, HARD, and QA track experiments using PIRCS. *Proceedings of the 12th Text Retrieval Conference*, 510-521.

Harman, D. & Buckley, C. (2004). The NRRC Reliable Information Access (RIA) workshop. *Proceedings of the 27th Annual International ACM SIGIR Conference*, 528-529.

Krovetz, R. (1993). Viewing morphology as an inference process. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191-203.

Kwok, K.L., Grunfeld, L., Sun, H.L., & Deng, P. (2005) TREC2004 robust track experiments using PIRCS. *Proceedings of the* 13th *Text REtrieval Conference (TREC 2004).*

Robertson, S. E. & Walker, S. (1994). Some simple approximations to the 2-Poisson model for probabilistic weighted retrieval. *Proceedings of the* 17th *ACM SIGIR Conference on Research and Development in Information Retrieval*, 232-241

Rocchio, J. J., Jr. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The Smart System-- experments in automatic document processing*, 313-323. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41,* 288-297.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

Yang, K. (2002a). Combining Text-, Link-, and Classification-based Retrieval Methods to Enhance Information Discovery on the Web. (*Doctoral Dissertation*. University of North Carolina).

Yang, K. (2002b). Combining Text- and Link-based Retrieval Methods for Web IR. *Proceedings of the 10th Text Rerieval Conference (TREC2001)*, 609-618.

Yang, K., & Yu, N. (2005). WIDIT: Fusion-based Approach to Web Search Optimization. *Asian Information Retrieval Symposium 2005*.

Yang, K, Yu, N., & Lee, Y (2005). Dynamic Tuning for Fusion: Harnessing Human Intelligence to Optimize System Performance. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics.*