

Towards Automatic Facet Analysis and Need Negotiation: Lessons from Mediated Search*

Jimmy Lin, Philip Wu
The iSchool, College of Information Studies
University of Maryland
College Park, MD 20742, USA
{jimmylin,fwu}@umd.edu

Eileen Abels[†]
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104, USA
eabels@drexel.edu

Abstract

This work explores the hypothesis that interactions between a trained human search intermediary and an information seeker can inform the design of interactive IR systems. We discuss results from a controlled Wizard-of-Oz case study, set in the context of the TREC 2005 HARD track evaluation, in which a trained intermediary executed an integrated search and interaction strategy based on conceptual facet analysis and informed by need negotiation techniques common in reference interviews. Having a human “in the loop” yielded large improvements over fully-automated systems as measured by standard ranked-retrieval metrics, demonstrating the value of mediated search. We present a detailed analysis of the intermediary’s actions to gain a deeper understanding of what worked and why. One contribution is a taxonomy of clarification types informed both by empirical results and existing theories in library and information science. We discuss how these findings can guide the development of future systems. Overall, this work illustrates how studying human-information seeking processes can lead to better information retrieval applications.

1 Introduction

Searching for information is a highly complex and iterative activity (Swanson, 1977; Spink, 1997) that occurs within the context of broader information-seeking behaviors (Wilson, 1999), processes of cognition (Ingwersen, 1999), and attempts to navigate unfamiliar information spaces (Bates, 1991; Dervin, 1991; Pirolli and Card, 1999). As the user is the final arbiter of which information objects are examined and which are ignored, interaction between the user and the system is arguably *the* single most important element in the design of retrieval systems. Indeed, it is difficult to imagine information retrieval without a user, and even when retrieval technology serves as the basis for other applications (e.g., question answering or document summarization), a user remains firmly in the loop *somewhere*.

*Please cite as: Jimmy Lin, Philip Wu, and Eileen Abels. Towards Automatic Facet Analysis and Need Negotiation: Lessons from Mediated Search. *ACM Transactions on Information Systems*, 2008, in press. This is the pre-print version of an article to be published. This version was prepared December 7, 2008, and may have minor differences with the published article.

[†]This work was conducted while the author was at the University of Maryland.

Despite many elaborate theoretical models (Belkin et al., 1995; Ingwersen, 1996; Saracevic, 1997a) for interactive information retrieval, most systems today exhibit comparatively little diversity in their underlying interaction models. For the most part, interactivity is based on some variant of relevance feedback (Efthimiadis and Robertson, 1989; Salton and Buckley, 1990). Under this general framework, users are called upon to assess the relevance of system results, whether at the document, sentence, or word level, presented in various contexts (Kelly et al., 2005; Kelly and Fu, 2006). In the simplest case, relevance information can be exploited by the system through enriched queries. Within a language modeling framework, user feedback can be leveraged to refine models of relevance (Zhai and Lafferty, 2001).

Although relevance feedback has proven to be effective in user studies (Koenemann and Belkin, 1996; Efthimiadis, 2000) and in other controlled settings (Harman, 1988; Ruthven, 2003), there is evidence that users are reluctant to supply the necessary relevance judgments (Beaulieu et al., 1997; Belkin et al., 1999). Increased cognitive load has been hypothesized as a possible cause (Bruza et al., 2000), but there are several confounding factors that might prevent an accurate generalization, including the diversity of user tasks. Naturally, for simple fact-finding searches where only one answer instance is sought, relevance feedback is of little use. In the Web environment, there is scant published evidence that users employ “more like this” or “find similar documents” features implemented by many search engines, despite the demonstrated effectiveness of such capabilities in simulated environments (Wilbur and Coffee, 1994; Smucker and Allan, 2006; Lin and Smucker, 2008); but see (Lin and Wilbur, 2007; Lin et al., 2008 in press). Recent work in interactive question answering has shown that intelligence analysts are willing to engage in extended interactions with systems in order to solve complex analytical problems (Small et al., 2004; Harabagiu et al., 2005), as evidenced by successful large-scale user evaluations in the AQUAINT research program, e.g., (Wacholder et al., 2007). Given these various threads of work, it appears clear that relevance feedback is not a “one-size-fits-all” solution to interactive retrieval.

Building on previously published results (Lin et al., 2006), this article explores the hypothesis that interactions between a trained human search intermediary and an information seeker can inform the design of interactive IR systems. To better circumscribe the problem space, we focus on time-limited single-iteration interactions, in which the system has only one opportunity to solicit input from the user (what we define later as a clarification dialogue). Our exploration can be divided into two complementary parts:

- First, we discuss results from a controlled Wizard-of-Oz case study in which a trained intermediary engaged in search with an off-the-shelf IR system to address a series of information needs. The intermediary executed an integrated search and interaction strategy based on conceptual facet analysis and informed by need negotiation techniques common in reference interviews. Having a human “in the loop” yielded large improvements over fully-automated systems as measured by standard ranked-retrieval metrics. These results affirm the value of mediated search.
- Second, we present a detailed analysis of our intermediary’s search and interaction strategy to gain a deeper understanding of what worked and why. One contribution is a taxonomy of clarification questions informed both by empirical results and existing theories in library and information science. We are able to quantify the effectiveness of different clarification types using a simple linear regression model. Finally, we discuss how these findings can guide the development of future retrieval systems.

This work not only confirms our initial hypothesis, but actually supports a stronger claim: we argue that our strategy for mediated search provides a good model for designing interactive IR systems. In an attempt to bridge system-centered approaches (typically adopted by computer scientists) and user-centered approaches (typically adopted by researchers in library and information science), we discuss how studying human information-seeking processes can lead to better information retrieval applications.

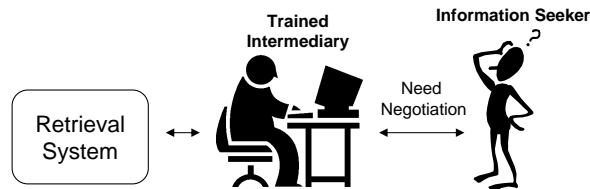


Figure 1: Illustration of the reference interview. The information seeker has access to the retrieval system only indirectly through the trained intermediary. Through the process of need negotiation, both the intermediary and the information seeker arrive at a better understanding of the need and strategies for achieving the desired objectives.

This article is organized in the following manner: We begin in Section 2 with an overview of need negotiation, the reference encounter, and interactions in mediated search. Section 3 discusses different IR evaluation methodologies, concluding in a description of the TREC 2005 HARD track, which encoded an interesting set of tradeoffs in evaluation design—halfway between user-centered and system-centered approaches. Section 4 describes how the HARD evaluation provided an experimental vehicle for this work. The section also details the strategy we developed for mediated search and the procedures followed by our intermediary. Section 5 discusses results from our case study and Section 6 provides more detailed analyses of what worked well and why. Section 7 proposes a taxonomy of clarifications that captures patterns of intent in intermediary–user exchanges. We then discuss implications for the design of interactive IR systems in Section 8 before concluding.

2 The Reference Encounter and Mediated Search

It is easy to forget in today’s digital world that information seeking does not necessarily involve computers. Before the widespread availability of online retrieval systems, mediated search was the norm—to address their needs, information seekers sought the assistance of trained intermediaries (e.g., reference librarians), the only population with access to computerized search systems.¹ The starting point of our work is the hypothesis that interactions in the context of mediated search can inform the design of interactive IR systems in non-mediated search.

In a library setting, the dialogue between an intermediary and a patron is commonly known as the reference interview, which Bopp and Smith (1995) define as “conversation between a member of the library reference staff and a library user for the purpose of clarifying the user’s needs and aiding the user in meeting those needs”. We illustrate this process in Figure 1. This complex communication begins with the patron describing the information sought after. Through a series of interactions, *both* parties arrive at a better understanding of the information need (Taylor, 1962; Knapp, 1978)—a process often referred to as need negotiation. These interactions help ensure that the right answer is found to the right question. In this work, we pose the following question: might the reference encounter provide a model for interactive IR systems?

Of course, face-to-face human communication is a highly complex activity full of subtle nuances that would be difficult to capture within an automated retrieval system. However, additional technological developments in the library setting make our research question more tractable. As users become more accustomed to searching online electronic resources, the face-to-face reference interview is gradually being replaced by other media: initially, the telephone, and now, email (Abels, 1996) and online chat (Francoeur, 2001). In particular, the last two modalities provide useful models, since it eliminates

¹Throughout this article, we use “information seeker” and “user” interchangeably to refer to the person with the information need (e.g., the patron in a library); we use “intermediary” to refer to the person who assists.

those communicative cues from a face-to-face setting that are the most difficult to computationally model (gestures, body language, facial expressions, etc.).

Our exploration is guided by the substantial body of work on reference encounters in the library science literature, e.g., (White, 1985; Taylor, 1968). For example, Talyor identifies five broad categories of exchanges in the need negotiation process:

- determination of the subject that the user is searching on;
- objective and motivation for the current search;
- personal characteristics of the user;
- relationship between the search statement and the file organization in the collection; and
- anticipated or acceptable answers.

It is not difficult to see how such information might be useful to an interactive IR system. Nevertheless, it is by no means obvious how automated systems might solicit such information from a user, represent it internally, and exploit it to delivery higher-quality results. This article presents a step toward this goal.

3 The Evolution of Evaluation Methodologies

This section discusses different methodologies that have been used to evaluate information retrieval systems, culminating in a description of the TREC HARD track, which provided the experimental vehicle for our work.

Despite the importance of interaction, the development of information retrieval systems for the past several decades has been primarily guided by a paradigm that marginalizes the user. Batch-style, system-centered evaluations in the Cranfield tradition (Cleverdon et al., 1968) assume a one-shot model of interaction and an impoverished model of the user. Such evaluations are exemplified by the *ad hoc* retrieval task in Text Retrieval Conferences (TRECs), annual evaluations organized by the U.S. National Institute of Standards and Technology (NIST), as well as similar tasks in other community evaluations such as CLEF, INEX, and NTCIR. Although test collections built from batch evaluations provide a fast and cost-effective method for quantifying the effectiveness of IR systems—typically through single-point metrics such as mean average precision—a number of empirical studies have shown that system gains as measured by these laboratory tools do not necessarily translate into improvements in users’ task performance (Hersh et al., 2000; Turpin and Hersh, 2001; Allan et al., 2005; Turpin and Scholer, 2006). The lack of realism associated with large-scale batch evaluations has been pointed out by many (Beaulieu et al., 1996; Saracevic, 1997b; Sparck Jones, 2000). The same researchers and many others have called upon the IR community to embrace research programs that renew a focus on the user. Nevertheless, the development and evaluation of interactive retrieval systems faces a number of challenges.

Carefully-orchestrated user studies that examine human search behavior in natural settings provide an effective approach for evaluating interactive retrieval systems. Indeed, for many years the IR community invested in an interactive track at TREC (Beaulieu et al., 1996; Hersh and Over, 2001; Over, 2001); more recently, in interactive evaluations at CLEF and INEX. However, compared to batch evaluations, the high-cost and time-consuming nature of these studies limit the speed at which hypotheses can be explored and the statistical significance of results. As a specific case study, consider the interactive tracks at TREC, whose setup is illustrated in Figure 2. They can be best described as coordinated user studies, where formal properties of study design were standardized across participants. However, within a shared framework teams were free to explore whatever research questions

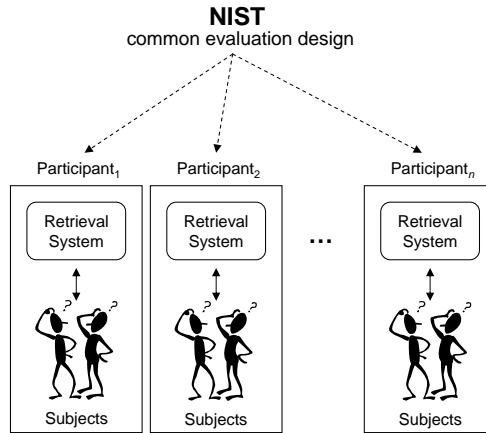


Figure 2: Illustration of the methodology employed by the TREC interactive tracks. Multiple independent participants at different sites adopted a common evaluation design, but were responsible for recruiting their own subjects and were free to explore different research questions.

interested them. Even with a common study design, results were often not comparable and sometimes even contradictory, which made it difficult to derive generalizations (see above references and TREC proceedings² from the years during which the interactive track was active). Ultimately, the interactive tracks ended because there was not a clear way forward. Ellen Voorhees (personal communication) elaborates: Interactivity with its inherent emphasis on the user is difficult to study in general, and more difficult to study in the context of TREC, which has an emphasis on building reusable resources. Paul Over (personal communication) provides a complementary perspective: Despite continued interest at NIST and among a small set of faithful participants, NIST ended the interactive track in order to devote more staff time to other tracks and to encourage exploration of interactive issues in other tracks (e.g., Web search, question answering, etc.).

Different types of evaluations encode, whether implicitly or explicitly, tradeoffs between insightfulness, affordability, and repeatability. Cranfield-style experiments score high on repeatability and affordability (once the test collection has been created). However, these two strengths come at considerable expense to insightfulness—for example, it is difficult to interpret ranked-retrieval metrics such as mean average precision within real-world user tasks. Additional TREC assumptions such as independent binary relevance have also been criticized as being unrealistic. User studies, on the other hand, sacrifice affordability and rapid repeatability for insightfulness. Any experiment that involves users must contend with variability and the confounding factors that inevitably come with human subjects, in spite of the usual set of best practices employed in study design. However, batch experiments and user studies occupy but two points in the design space of IR evaluations. What other compromises might one strike between insightfulness, affordability, and repeatability?

The HARD (High Accuracy Retrieval of Documents) track was started in TREC 2003 as an attempt to reintroduce user issues back into TREC, of which interaction was one. It was originally conceived with a focus on three different ideas (Allan, 2003): richer specifications of information needs (i.e., context), finer-grained units of retrieval (i.e., passages), and limited interaction with the user (the so-called clarification dialogues). After discussions at the TREC workshop, the community collectively decided that clarification dialogues represented the most promising avenue in which to advance the state of the art, and they were retained as the sole focus for TREC 2005 (Allan, 2005). A clarification dialogue was operationally defined as a single iteration of user–system interaction. That is, the system is given one opportunity to solicit additional input from the user and, based on this feedback, the

²available at <http://trec.nist.gov/>.

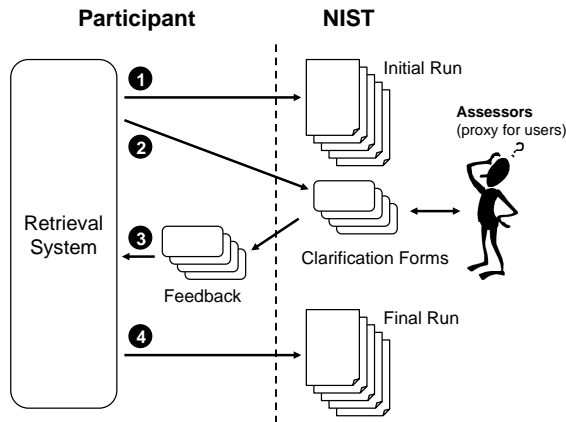


Figure 3: Illustration of the methodology employed in the TREC 2005 HARD track. NIST assessors served as proxies for users and engaged in one iteration of interaction (a clarification dialogue) with systems through HTML forms.

system should be able to retrieve more relevant documents. In other words, a clarification dialogue in the context of HARD represents what Spink calls an interactive feedback unit (Spink, 1997), which she argues to be a fundamental unit in IR interaction.

Figure 3 illustrates the setup of the TREC 2005 HARD track. At the core, interaction was designed on top of a standard *ad hoc* retrieval task, where the goal was to return a ranked list of documents. NIST assessors, in addition to providing relevance judgments, also served as proxies for users (i.e., information seekers). In the evaluation, participants first submitted initial (pre-interaction) runs to NIST (label 1) along with a set of forms that encapsulated the clarification dialogues (label 2). These clarification forms, which served as the vehicle for user–system interactions, were standard HTML forms that displayed output from the system and solicited input from the user. Thus, interaction was limited to that which could be conveyed on a Web page—check boxes, text input boxes, and the like (although Javascript was allowed and indeed exploited by some groups). This design allowed for asynchronous interactions that made evaluation of multiple systems practical. An important point to emphasize: clarification forms imposed only *technical* restrictions on the interactions—participants were free in designing whatever *content* they felt was appropriate. NIST developed the software infrastructure for managing and presenting these clarification forms to the assessors; three minutes were allotted for each topic (information need). Feedback was gathered via the CGI protocol (i.e., which check boxes were selected, what was typed in each text box, etc.) and returned to each participant (label 3 in Figure 3). Finally, participants submitted final (post-interaction) runs that exploited the assessor feedback (label 4). NIST evaluated all initial and final runs using the standard pooling methodology for *ad hoc* retrieval (Harman, 2005). By comparing the effectiveness of the initial and final runs, it was possible to quantify the impact of the interactions. For us, this evaluation design provided an experimental vehicle for assessing the impact of mediated search (see Section 4).

The HARD evaluation represented an interesting tradeoff point in the space of evaluation design—halfway between user-centered and system-centered approaches. The primary difference that distinguished HARD from previous TREC interactive tracks was the centralization of experimental subjects at NIST (using assessors as proxies for users). Since each assessor interacted with *all* systems, cross-site comparisons were possible (and with them, the possibility for broader generalizations). Clarification dialogues represented an attempt to reduce both the scope and duration of user–system interactions to allow practical implementation in large-scale evaluations: instead of arbitrarily complex interface controls, interactions were limited to elements that could appear on an HTML page; instead of arbitrarily long sessions, interactions were restricted in duration. The design, however, was not without

Topic 436**Title:** Railway Accidents**Description:** What are the causes of railway accidents throughout the world?**Narrative:** A relevant document provides data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.

Figure 4: Example of a HARD topic.

drawbacks: due to unavoidable constraints involved in coordinating the HARD track, documents were not assessed until approximately one month after the clarification questions had been answered. The implications for this lag are explored in Section 6.3.

Two salient characteristics of HARD clarification dialogues are their asynchronous nature and short duration. Although both stem from operational considerations in coordinating multi-site evaluations, they reflect trends in information-seeking environments. Users today are accustomed to a rapid back-and-forth style of interaction, which makes prolonged exchanges less realistic. The asynchronous nature of the interaction mirrors quite well certain scenarios commonplace today, for example, emailing a reference librarian with questions or submitting a question on a dedicated forum of experts.

Before describing our experiments, we provide a few more details on the test collection used in the evaluation: the TREC 2005 HARD track employed the AQUAINT collection of newswire text,³ consisting of English data drawn from three sources: the New York Times News Service, the Associated Press Worldstream News Service, and the Xinhua News Service (from the People’s Republic of China). The collection contains approximately one million articles totaling roughly three gigabytes.

As with other *ad hoc* retrieval tasks, the starting point of a search was a written description of the information need, or in TREC parlance, a *topic*. HARD topics followed the standard format consisting of a short title, a sentence-long description, and a more detailed narrative. Figure 4 shows a typical HARD topic. Instead of developing topics from scratch, the evaluation reused fifty “difficult” topics from previous *ad hoc* tasks (defined as topics that almost no system was able to handle well). However, since the topics were previously evaluated on a different collection, NIST first manually vetted them to insure that at least three relevant documents could be found in the new collection. The standard pooling methodology (Harman, 2005) was used to gather relevance judgments, which were then applied to evaluate all system output (i.e., ranked lists).

4 Mediated Search in TREC

The HARD evaluation in TREC provided an experimental vehicle for exploring the hypothesis that mediated search can inform the design of interactive IR systems. In place of an automated retrieval system, we deployed a “system” consisting of a trained intermediary⁴ who worked with the INQUERY system⁵ (see Figure 5). This in effect created a mediated search scenario—without the NIST assessors’ knowledge. Such an experimental design is commonly known as a “Wizard-of-Oz” setup (Kelley, 1984), where a subject is lead to believe that he or she is interacting with a machine, when in fact there is a human “behind the curtain”.

³LDC catalog number LDC2002T31

⁴Philip Wu (one of the co-authors), a Ph.D. student in the iSchool at Maryland.

⁵Version 3.1p1 for Solaris.

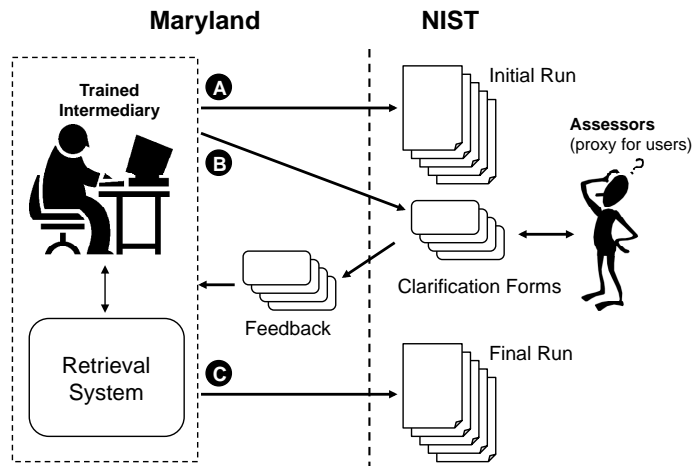


Figure 5: Illustration of the methodology employed by the University of Maryland in the TREC 2005 HARD track. The retrieval system was augmented with a trained intermediary in a Wizard-of-Oz setup.

Our first goal was to measure the effectiveness of mediated search and to quantify improvements that could be obtained by placing a human “in the loop”. To many IR researchers, it is not obvious that this should even be the case, considering that manual runs in previous TREC evaluations were not considerably better than fully-automatic runs; see, for example, (Voorhees and Harman, 1999).

However, the true value of our experiment resided in post-hoc analysis of assessor–intermediary exchanges and generalizations gleaned from those transactions that can inform the design of interactive IR systems. In anticipation of this analysis, we devised a systematic search and clarification strategy that the intermediary followed. This strategy consisted of three inter-connected parts:

- Initial search (label A in Figure 5): an approach to need analysis and query formulation based on conceptual facet analysis.
- Generation of clarification questions (label B in Figure 5): an approach to dialogue based on disambiguating relevance categories.
- Use of feedback (label C in Figure 5): an approach to refining category membership based on responses gathered from the assessor.

While we acknowledge substantial work that has been done on analyzing face-to-face reference encounters, e.g., (Swigger, 1985; Nordlie, 1999; Spink et al., 1996; Spink and Saracevic, 1997; Spink, 1997; White, 1998), we believe that this work is qualitatively different for several reasons. First, the TREC setup narrows the scope of inquiry, which naturally limits the realism of the problem, but allows us to make stronger conclusions. Second, since most of the previous studies were observational in nature, the researchers had little control over aspects of intermediary behavior, unlike in the current study. Third, by replacing unconstrained face-to-face human–human interaction with stylized clarification dialogues, it is more likely that study findings can be directly applied to automated systems. In Section 7.4, we return to these issues in detail.

Note that our experimental setup essentially amounts to a case study involving one trained intermediary. Due to the involved nature of the study design, we lacked the resources to employ multiple intermediaries. However, case studies represent a well-established tradition in qualitative research (Merriam, 1998; Yin, 2003), typically used to analyze complex events in a holistic fashion. According to Yin (2003), it is possible to derive generalizations, especially when a “previously developed theory is

used as a template with which to compare the empirical results of the case study” (p. 33). Our work is framed in the context of well-established research in library science and information retrieval; our results are consistent with the existing literature on information retrieval and knowledge about human information-seeking processes; our analyses extend and refine principles for designing information retrieval systems without contradicting them. Furthermore, there is nothing particularly remarkable about our study design: our intermediary represents a skilled searcher, typical of what one might expect with formal training in library science; the search system is a standard off-the-shelf ranked retrieval system that supports a wide range of query operators. The combination of theoretical grounding and generic study design helps ensure both the validity of our generalizations and the transferability of our results to other contexts.

Details of our mediated search strategy are organized as follows: Section 4.1 discusses our query formulation strategy; Section 4.2 presents our clarification strategy based on creating and shuffling document piles; Section 4.3 describes how these document piles are linearized into ranked lists for evaluation in TREC.

4.1 Query Formulation Strategy

Starting from the TREC topic statement, the intermediary used the “building blocks” strategy (Harter, 1986; Marchionini, 1995) to construct search queries that leveraged INQUERY’s rich query operators, which are capable of combining multiple term evidence using inference networks (Turtle and Croft, 1991). The building blocks strategy is based on conceptual facet analysis, and is frequently used by reference librarians. Our intermediary first identified conceptual facets from each HARD topic, leveraging text in all three fields of the topic statement (title, description, narrative). Each facet was then instantiated through a set of synonyms and related terms using INQUERY’s #syn operator (conceptually equivalent to the Boolean operator OR). Note that conceptual facet analysis explicitly distinguishes conceptual entities (e.g., “railway”) from terms that may represent the concept in documents (e.g., “tracks”, “rail”, “train”, etc.). The process allows a separation of information need analysis and query construction. As we discuss in Section 7, a search strategy based on conceptual facet analysis provides a basis for understanding the nature of clarification exchanges.

The resulting sets of query terms (representing the facets) were then organized into a complete query that captured constraints and relationships in the information need. Typically, one of two INQUERY operators was used as the connective: the #sum operator, which combines evidence from multiple clauses and can be viewed as a “soft” AND (i.e., best match), or the #band operator, equivalent to a strict AND. The intermediary switched back and forth between strict and soft connectives in a manner similar to what Spink (1997) calls magnitude feedback—query modifications based on the size of the result set. If a soft connective did not appear precise enough, strict constraints were imposed. If strict constraints yielded too few hits, the intermediary switched back to soft operators.

Schematically, a complete building blocks query would look like the following (using strict Boolean operators, for illustrative purposes):

$$(A_1 \vee A_2 \vee A_3 \vee \dots) \wedge (B_1 \vee B_2 \vee B_3 \vee \dots) \wedge \dots$$

Three specific examples of INQUERY queries that were constructed with the above strategy are shown below:

Topic 436: Railway Accidents

#band(#syn(railway railroad locomotive trolley streetcar) #syn(incident injure dead kill))

Topic 354: Journalist Risks

#sum(#syn(journalist correspondent reporter) #syn(risk kill arrest hostage))

Topic 448: Ship Losses

#band(ship #syn(weather storm wind) #syn(loss missing disappear sink))

These queries represent a straightforward execution of the query formulation strategy outlined above. However, many topics required augmenting conceptual facet analysis with additional search strategies. In many cases, the initial queries created using the building blocks approach returned imprecise results. For those, the interactive scanning strategy (Hawkins and Wagers, 1982) was employed to refine the queries. Once a set of documents was returned by the initial search, the intermediary reviewed a small sample of the set (usually the top 10–15) to identify both relevant documents and reasons why irrelevant results were retrieved. These observations were then exploited to revise the initial query or to reformulate a new query—taking advantage of additional terms that may have appeared in the top hits and INQUERY’s full range of query operators (negation, proximity, etc.). This process was repeated until the intermediary was confident that a sufficient number of relevant documents had been gathered.

Consider the following two examples of query sequences issued to INQUERY:

Topic 330: Iran-Iraq Cooperation

Description: This query is looking for examples of cooperation or friendly ties between Iran and Iraq, or ways in which the two countries could be considered allies.

Q1: #sum(iran iraq cooperate collaborate ally tie relation)

Q2: #sum(#5(iran iraq) cooperate collaborate tie relation #5(return airplane))

Q3: #band(iran iraq #5(return airplane) #1(gulf war))

Q4: #sum(#10(iraq iran) #syn(cooperate collaborate) #phrase(border control) minority sanction)

Topic 345: Overseas Tobacco Sales

Description: Health studies primarily in the U.S. have caused reductions in tobacco sales here, but the economic impact has caused U.S. tobacco companies to look overseas for customers. What impact have the health and economic factors had overseas?

Q1: #sum(tobacco #phrase(#lit(U.S.) american company) market sale overseas health)

Q2: #band(tobacco #syn(increase decrease decline drop) #syn(health marketing) #syn(u.s. american) #syn(international overseas asia russia africa))

Q3: #band(#syn(tobacco cigarette) sale #syn(health advertise marketing) #syn(international overseas asia russia africa))

These two examples illustrate common query refinements, which may involve adding, replacing, or dropping query terms and restricting or adjusting proximity of terms (using the #*n* operator). For topic 330, Q1 returned many articles regarding diplomatic negotiations and the return of POW’s—the intermediary was unsure if such documents were relevant, and thus generated a clarification question (more on this later). A subsequent query retrieved more relevant documents (Q2), but a further attempt to narrow the results yielded only one hit (Q3). A backoff query (Q4) returned relevant documents that mostly overlapped with those from Q2, and hence the intermediary stopped at this point. A similar process of query refinement occurred with topic 345: Q2 was found to be overly restrictive since it returned only 10 hits, and hence the search was broadened.

In addition to examining top INQUERY results to find hints for query refinement, external sources such as search engines, online encyclopedias, and other electronic resources were used to help the intermediary understand the topic statement and identify good search terms. The use of external sources was especially helpful when the intermediary found it difficult to find relevant documents after several rounds of query reformulation. For example, after trying five different queries on topic 650

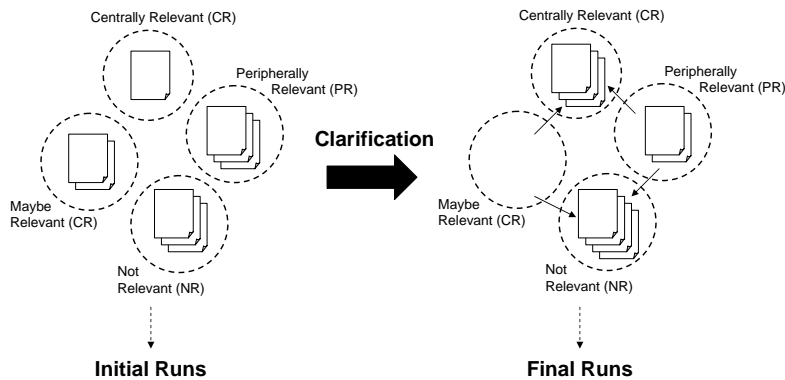


Figure 6: A schematic representation of our clarification strategy. The intermediary first assigned documents to one of four categories; the goal of the clarification dialogue was to refine category membership.

“Tax Evasion Indicted” without good results, the intermediary used a Web search engine to look for documents regarding major tax evasion cases; the Web documents were then examined with the aim of finding names of people and organizations that could be used in a new query. Another example was the use of Wikipedia to identify names of famous philosophers associated with stoicism, which was the central concept of topic 433 “Greek, Philosophy, Stoicism”.

This query formulation strategy provided a starting point for our clarification framework, which we describe next.

4.2 Clarification Strategy

We devised a framework for generating clarification questions and exploiting feedback based on the idea of creating and shuffling document piles. The intermediary manually classified documents into one of four relevance categories, and then took advantage of interaction to refine the category membership—this process is schematically shown in Figure 6. From these piles, we generated ranked lists that were submitted for evaluation (see Section 4.3).

After employing the search techniques described in Section 4.1, between 50 and 120 documents in the best query were manually examined for relevance. These judgments were made in addition to relevant documents that were gathered during the query refinement process. The amount of effort devoted to each topic varied according to its difficulty and the number of relevant results found. The intermediary assigned one of four judgments to each examined document:

- **Centrally relevant (CR):** based on the intermediary’s understanding of the information need, this document would be considered topically relevant.
- **Peripherally relevant (PR):** based on the intermediary’s understanding of the information need, this document would be considered relevant, but less so than documents marked centrally relevant (for example, a passing mention or a vague reference). Relevance is well-recognized in the information science literature as a graded property (Saracevic, 1975; Spink and Greisdorf, 2001; Sormunen, 2002), and distinguishing PR from CR documents represents a coarse attempt to capture this.
- **Maybe relevant (MR):** based on the intermediary’s understanding of the information need, this document may be relevant. Ambiguity in TREC topic statements often force the intermediary to make assumptions, draw inferences, etc. If a document would be considered relevant based

on a particular interpretation of the topic, this judgment is assigned. Note that MR documents are distinct from PR documents—MR documents may actually be centrally relevant, modulo assumptions by the intermediary.

- **Not relevant (NR)**: this document would not be considered relevant.

The generation of clarification questions, encapsulated in HTML forms sent to NIST (label 2 in Figure 3), was interwoven between the query formulation and relevance assessment processes. We conceived of clarification as a reshuffling of documents between the four piles created—that is, the intermediary aimed to refine the category membership of the documents. Clarification questions were explicitly created with one of two general goals:

- To move documents from the PR pile into either the CR or the NR pile. Although topical relevance is a graded quantity, TREC assessors are ultimately forced to make binary relevance judgments. Thus, there exists a “relevance threshold” that guides the user in making hard decisions about document-level relevance; these clarifications are aimed at a better understanding of this threshold.
- To move documents from the MR pile into either the CR or the NR pile. In searching, the intermediary makes judgments based on an interpretation of the information need; this often involves drawing inferences, making assumptions, etc. The purpose of these clarification requests is to verify the correctness of the interpretation.

Although we anticipated the creation of a taxonomy of clarifications based on post-hoc analysis (see Section 7), we consciously adopted an inductive, bottom-up approach. Thus, the intermediary formulated questions as appropriate, without reference to any pre-existing taxonomies, questions types, frames, templates, etc. However, all questions were constructed so that responses could be captured via check boxes—this ensured a consistent interaction pattern. In addition to topic-specific questions, all clarification forms included two generic questions (located at the end of the form): “Any additional search terms?” and “Any other comments?” Both were followed by a 70×4 character text box for free-formed input.

Figure 7 shows two complete examples of topic statements and the corresponding clarification forms—for topic 416 “Three Gorges Project” and topic 436 “Railway Accidents”. In the first topic, the second question was targeted at PR documents, while the other questions were targeted at MR documents. In the second topic, all questions were targeted at MR documents. As previously mentioned, most of the clarification questions reflected the contents of documents that the intermediary had examined. Take the last question of topic 436 as an example: although the topic statement explicitly mentioned only accidents involving vehicles and pedestrians, reports of accidents involving animals (e.g., elephants) were also found in the collection. These documents were marked MR, since it seemed reasonable to expand the scope of interest in this manner—the clarification dialogue provided the opportunity to confirm this assumption.

After receiving the clarification forms, NIST assessors interacted with them, providing feedback as requested. The results of these interactions, captured via CGI variable bindings, were then returned to the participant (label 2 in Figure 3). Based on responses to the clarification questions (i.e., whether or not the check boxes were marked), the intermediary reorganized the document piles. In many cases, enough information had been gathered to reclassify documents in the MR and PR piles. To give a concrete example, in the clarification question discussed above regarding the relevance of railway accidents involving animals, the NIST assessor marked the check box. Based on this evidence, all documents discussing animal-related accidents were moved from the MR to CR pile. In some cases, assessor feedback gave the intermediary ideas for new search queries, particularly for more difficult topics. Documents retrieved from these new queries were also sorted into the four categories.

Topic 416**Title:** Three Gorges Project**Description:** What is the status of The Three Gorges Project?**Narrative:** A relevant document will provide the projected date of completion of the project, its estimated total cost, or the estimated electrical output of the finished project. Discussions of the social, political, or ecological impact of the project are not relevant.**Clarification Questions**

1. Check if yes: Must a relevant article mention the date of completion and total cost and estimated electrical output? Leave unchecked if it is sufficient to discuss any one of these facets.
2. Check if yes: Is “early next century” an acceptable projected date of completion?
3. Check if yes: Would articles mentioning state bank loans or foreign investment be relevant?
4. Check if yes: Would articles discussing the cost (or completion date) of a subcomponent of the project be relevant? For example, “power transmission project” or “the first construction phrase”.

Topic 436**Title:** Railway Accidents**Description:** What are the causes of railway accidents throughout the world?**Narrative:** A relevant document provides data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.**Clarification Questions**

1. Check if yes: Would articles about events resulting from terrorist attack (bombs, etc.) or other intentional damage be relevant?
2. Check if yes: Is “derailment” considered a “cause”? Leave unchecked if the cause of the derailment must be further specified.
3. Check if yes: Would incidents involving animals be relevant (e.g., elephants killed by a train)?

Figure 7: Sample topics and clarification questions.

In our conception of the ideal interaction, the clarification forms would supply sufficient evidence for the intermediary to eliminate the PR and MR piles completely. In reality, however, documents remained in those piles even after assessor responses had been appropriately processed. The effectiveness of our clarification strategy is examined in Section 6.2.

4.3 Generating Runs

Although the TREC HARD guidelines called for run submissions in the form of ranked lists, we explicitly designed our clarification framework around relatively coarse-grained categories since we did not believe that the intermediary could effectively provide an absolute order for documents in terms of likelihood of relevance. To bridge sets (i.e., relevance categories) and ranked lists, we devised a number of simple techniques—applied to both the document piles before the clarification process and after the clarification process (these correspond to initial and final runs). As an aside, one might argue that the task of generating ranked lists from unordered sets is an artifact of the TREC model. In a real-world situation, there is no reason why systems cannot present multiple piles and leave the final actions to the discretion of the user.

A total of three pre-clarification (initial) runs were submitted to NIST for evaluation:

- **Run B1** (relevance feedback with piles), our main manual run, consisted of CR, PR, and MR documents (enumerated in that order). However, since this yielded far fewer than 1000 documents (the depth to which NIST evaluated the ranked lists), we “padded” results with a relevance feedback run. This was accomplished as follows: Based on *tf.idf* scores, 20 terms were selected from the documents marked centrally relevant. These terms were combined with terms from the topic title and description using INQUERY’s weighted sum operator (weight of 3.0 for title terms, 1.0 for all others). Duplicate documents (in the CR, PR, and MR piles) were removed, and the resulting list was appended to the manually gathered documents.
- **Run B2** (relevance feedback) was the relevance feedback run portion of run B1, i.e., the CR, PR, and MR documents were not prepended.
- **Run B3** (baseline) represented an automatic baseline. We submitted an INQUERY run that used terms from the title and description, with blind relevance feedback (top 20 *tf.idf* terms from top 10 hits).

We acknowledge that our techniques for linearizing the manually-assessed document piles was far from optimal. No special attempt was made to order documents in each of the categories; they were simply arranged in the order they were examined in the search process. We had no insight on the relative likelihood that PR and MR documents would be relevant to the assessor, and hence we simply decided on an arbitrary order. Finally, we had considered eliminating NR documents from the submitted runs, although preliminary experiments on old HARD topics demonstrated that this results in lower effectiveness since documents considered irrelevant by the intermediary could nevertheless be deemed relevant by the assessor.

A total of three post-clarification (final) runs were submitted to NIST for evaluation:

- **Run C1** (relevance feedback with updated piles), our main manual post-clarification run, followed exactly the same procedure as the creation of our pre-clarification run B1, except with the updated piles. In summary: CP, PR, and MR documents were prepended to a relevance feedback run (with duplicate removal).
- **Run C2** (topic and assessor-supplied terms) used title and description terms from the topic, along with search terms supplied by the assessor in the clarification forms. Query terms were

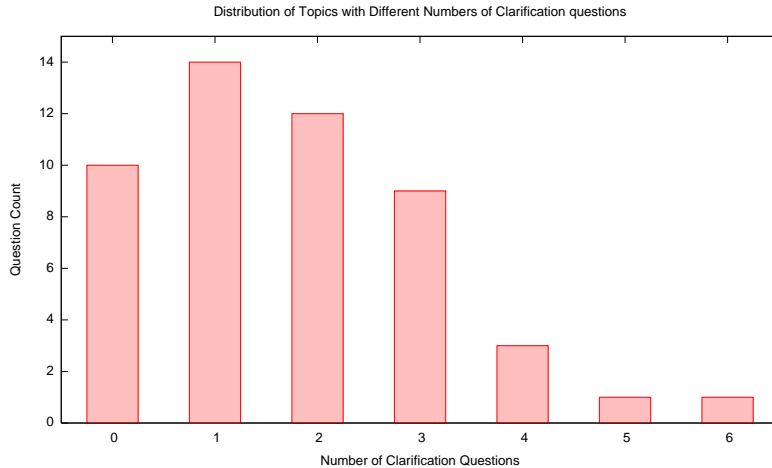


Figure 8: Histogram showing the distribution of topics in terms of the numbers of clarification questions.

combined using INQUERY’s weighted sum operator; a weight of 3.0 was given to title terms, and 1.0 to all other terms.

- **Run C3** (baseline with assessor-supplied terms) was created by augmenting run B3 with search terms supplied by the assessor in the clarification forms. The B3 run used title and description terms as the query, with blind relevance feedback (top 20 *tf.idf* terms from top 10 hits). Query terms were combined using INQUERY’s weighted sum operator; a weight of 3.0 was given to title terms, and 1.0 to all other terms.

The design of the initial and final runs explicitly isolated factors we wished to study (e.g., the effect of relevance feedback, the effect of clarification dialogues, etc.). In the next section, we present detailed comparisons between various runs and discuss the conclusions that can be drawn.

5 Results

This section presents results from our HARD experiments in TREC 2005: we provide descriptive statistics and also examine the effectiveness of our submitted runs in terms of standard ranked-retrieval metrics. Results confirm that our search and clarification strategy compares favorably to fully-automatic methods.

5.1 Descriptive Statistics

Our intermediary spent an average of 109 minutes per topic performing searches, assessing document relevance, and generating clarification questions (max 170, min 35, $\sigma = 29.7$). This time included analyzing the topic statement, formulating a “good” query, and performing the relevance judgments. For about half a dozen topics, the intermediary had difficulty generating a query that retrieved relevant documents; the advice of the co-authors was sought, but that time is not included in the figures above. We did not keep detailed time statistics for the process of exploiting clarification responses, but reassessing the documents took approximately ten to thirty minutes per topic. Note that there was about a two week gap between the time the clarification forms were submitted and the time assessor feedback was received.

Our intermediary generated a total of 88 clarification questions across 50 topics, for an average of 1.8 questions per topic ($\sigma = 1.41$). These figures do not include the two generic questions present for every

	B1	B2	B3	median	best	best auto
MAP	0.452	0.368	0.252	0.190	0.496	0.304
R-Prec	0.460	0.386	0.292	0.252	0.513	0.329
	C1	C2	C3	median	best	best auto
MAP	0.469	0.233	0.263	0.207	0.535	0.322
R-Prec	0.476	0.286	0.301	0.264	0.545	0.355

B1: rel feedback + piles **C1**: rel feedback + updated piles
B2: rel feedback **C2**: topic + assessor-supplied terms
B3: baseline **C3**: baseline + assessor-supplied terms

Table 1: Official results from the TREC 2005 HARD track (pre-clarification on top and post-clarification on bottom).

topic (“Any additional search terms?” and “Any other comments?”). Topic 341 “airport security” had the most clarification questions, with six. Ten topics had no clarification questions (beyond the two generic questions): the intermediary found them to be straightforward. Disregarding these ten topics, the average number of questions per topic jumps to 2.2 ($\sigma = 1.22$). The histogram in Figure 8 shows the distribution of topics in terms of the number of clarification questions they had.

For 35 of the topics, clarification responses included additional search terms supplied by the assessor. In 15 of the forms, clearly demarked phrases were entered. There was an average of 3.66 additional terms or phrases per topic ($\sigma = 3.31$), with a maximum of fourteen.

5.2 Run Effectiveness

We submitted a total of three pre-clarification and three post-clarification runs to the TREC 2005 HARD track (described in Section 4.3). In summary, they are:

- **B1**: relevance feedback with prepended piles.
- **B2**: relevance feedback.
- **B3**: baseline (blind relevance feedback).
- **C1**: Same run as **B1**, but with updated piles.
- **C2**: topic and assessor-supplied terms.
- **C3**: baseline with assessor-supplied terms.

Official results (mean average precision and R-precision) are shown in Table 1 for all submitted runs. A few other metrics are included for reference: the column marked “median” is the mean of the per-topic median score of all submitted runs, “best” is the mean of the best per-topic score of all submitted runs, and “best auto” is the highest-scoring automatic run. In total, 30 pre-clarification and 92 post-clarification runs were submitted by 16 groups. For 29 topics (out of 50 total), the B1 pre-clarification run achieved the highest average precision (across all submitted runs); for R-precision, 28 topics. For 20 topics, the C1 post-clarification run achieved the highest average precision; for R-precision, 17 topics.

A number of pairwise comparisons between the six submitted runs are presented in Table 2. The columns show relative differences in terms of mean average precision and R-precision. In all cases, the Wilcoxon signed-rank test was applied to determine the statistical significance of the differences:

Comparison	MAP	R-precision	Meaning
B1 vs. B2	+22.8% [▲]	+19.2% [▲]	effect of prepending CR, PR, MR piles
B2 vs. B3	+46.0% [▲]	+32.2% [▲]	effect of relevance feedback
B1 vs. B3	+79.4% [▲]	+57.5% [▲]	effect of human “in the loop”
C1 vs. B1	+3.8% [▲]	+3.5% [△]	effect of clarification dialogues
C3 vs. B3	+4.4% [△]	+3.1% [°]	effect of assessor-supplied terms
C2 vs. B3	−7.5% [°]	−2.1% [°]	assessor-supplied terms vs. blind relevance feedback terms

Table 2: Pairwise comparisons between various pre-clarification and post-clarification runs. All x vs. y comparisons indicate relative difference of x over y , i.e., positive if x is greater than y . Final column briefly explains each comparison.

significance at the 5% level is indicated with [△], at the 1% level, [▲]; results not statistically significant are marked with [°]. The final column briefly describes each comparison.

Table 2 is divided horizontally into three sections, which correspond to three general findings supported by our experimental results:

- **Mediated search is effective.** We confirm that placing a human in the loop can enhance retrieval effectiveness, as measured by standard ranked-retrieval metrics. Naturally, the potential benefits that can be derived will vary from intermediary to intermediary and will also vary based on the intermediary’s actions. On the whole, however, this case study illustrates the typical benefits of mediated search.

The blind relevance feedback (BRF) run B3 served as our fully-automatic baseline.⁶ Run B2, which incorporated (manual) relevance feedback (RF), significantly outperforms the baseline BRF run B3. We get a further boost in effectiveness over run B2 by prepending the intermediary’s CR, PR, and MR piles. Combining both methods (relevance feedback and prepending piles from the intermediary) yields a large, statistically-significant cumulative gain (B1 vs. B3).

These experiments demonstrate that the use of a trained intermediary can yield high payoffs. This is by no means an obvious finding, considering that manual runs in previous TREC evaluations were not considerably better than fully-automatic runs, e.g., (Voorhees and Harman, 1999). The crucial difference here is the element of interaction—previous manual runs for the most part consisted of single-shot retrieval with human-constructed queries. The setup is unrealistic in that an information seeker does not attempt to optimize a single ranked list from one interaction, but rather culls relevant information from multiple iterations.

- **Clarification dialogues improve effectiveness.** The interactions yield a small but statistically significant improvement in ranked-retrieval metrics. This conclusion is reached by comparing runs B1 and C1—the only difference between these two runs is the composition of the document piles, which isolates the effect of the HARD interaction. We see that clarification dialogues yielded a 3.8% gain in MAP and 3.5% gain in R-precision. Wilcoxon signed-rank tests show that the difference in MAP is significant at the 1% level and the difference in R-precision is significant at the 5% level. Figure 9 shows the effects of the clarification dialogues on average precision on a per-topic basis. Each pair of closely-spaced bars represents one topic: the left bar shows the range of the median to best scores before clarification; the right bar, after clarification. Boxes indicate

⁶Blind relevance feedback has been found to consistently improve IR, generally independent of topic difficulty (Carpineto et al., 2001). The fact that B3 is a competitive baseline is confirmed by results from the TREC 2005 Robust track, which used the same topics—the top scoring systems all took advantage of BRF (Voorhees, 2005).

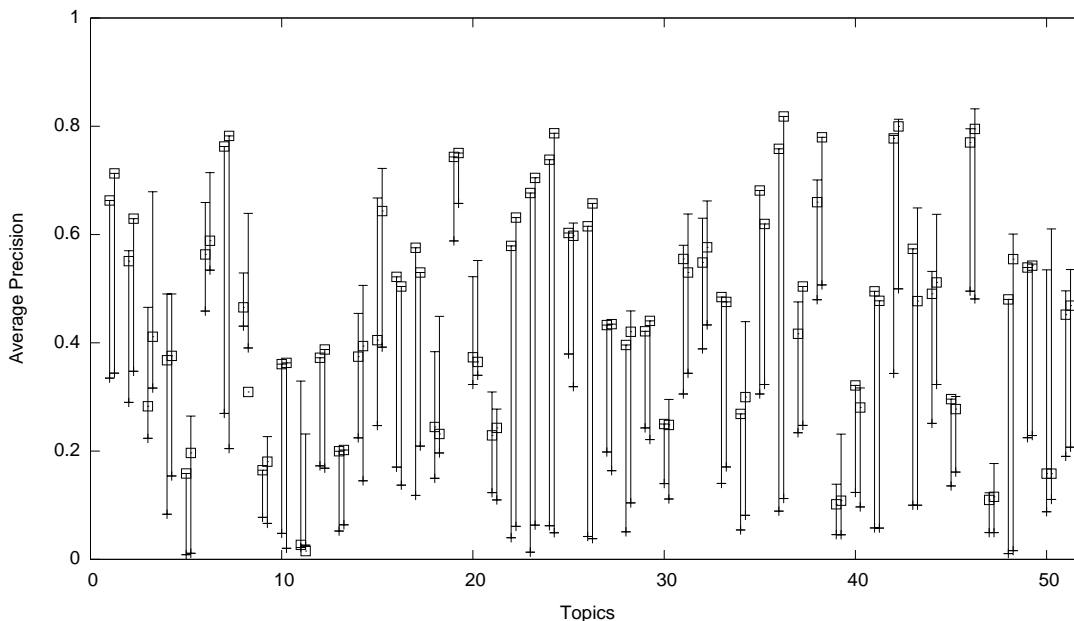


Figure 9: Comparison of average precision on a per-topic basis. Each pair of bars represents the median/best score range, before and after clarification. B1 (pre-clarification) and C1 (post-clarification) scores are marked with boxes. The rightmost set of bars represents the mean across all topics.

the average precision of runs B1 and C1, respectively. The rightmost set of bars represents the mean across all topics. Note that for some topics, average precision actually decreased after the clarification dialogue. See Section 6.3 for an analysis of factors affecting clarification effectiveness and discussion of this somewhat surprising result.

- **The effects of assessor-supplied query terms are relatively minor.** Two comparisons suggest that directly eliciting query terms from assessors without any contextual aids is not an effective interaction technique. Consider the difference between runs C3 (baseline + assessor-supplied terms) and B3 (baseline): they are similar except that the C3 queries have been augmented with assessor-supplied query terms. This yields significantly better MAP at the 5% level, but the difference is not statistically significant in terms of R-precision.

Consider the difference between C2 (topic + assessor-supplied terms) and B3 (baseline): since the B3 run uses blind relevance feedback, we are essentially comparing the effectiveness of assessor-supplied terms and automatically-selected blind relevance feedback terms. We find no statistically significant differences in either MAP or R-precision.

6 Analysis

The second part of our work begins with a detailed analysis of our interaction strategy and experimental results, where we attempt to gain a deeper understanding of what worked and why. Our major findings are summarized below:

- We find a weak positive correlation between average precision for a topic and the prevalence of relevant documents for that topic.

- We confirm the effectiveness of our clarification strategy by examining the composition of the document piles before and after clarification. We see an increase in the number of documents that are relevant, as determined by the NIST assessors.
- Based on a failure analysis of instances in which the clarification dialogue decreased effectiveness, we discovered cases where assessors’ feedback was inconsistent with their actual relevance criteria.

In what follows, we elaborate on each finding in turn. These analyses lay the groundwork for our taxonomy of clarification questions (Section 7) and discussions of implications for system design (Section 8).

6.1 Effect of Size of Relevant Document Set

We are interested in relationships that may exist between the number of known relevant documents for a topic and average precision for that topic. This is an interesting question for several reasons: since our intermediary examined far less than the 1000 hits per topic in our submitted ranked lists, effectiveness on topics with large numbers of relevant documents might be poor due to recall-related problems. Previous studies have demonstrated that humans are not very good at estimating recall (Blair and Maron, 1985). Since mediated search is often precision-focused, our intermediary might excel in topics with fewer relevant documents.

To better understand these effects, we focused on the manual run C1 (relevance feedback with updated document piles prepended) and the automatic run B3 (baseline blind relevance feedback). For both runs, we created scatter plots relating the size of the relevant documents set to average precision; these are shown in Figure 10. Regression lines are superimposed on the plots. In both cases, there is a weak positive correlation between average precision and number of relevant documents. In other words, for both the manual and automatic runs, topics with more relevant documents appear to be “easier”. In the case of the manual run, our technique for generating ranked lists appears to address possible recall concerns. We note a stronger relation between number of relevant documents and average precision in the automatic run B3 than in the manual run C1 (R^2 of 0.185 *vs.* 0.077). That is, more variance in terms of average precision is explained by the prevalence of relevant documents in the automatic run than in the manual run.

In the Figure 10 scatter plot for the manual C1 run, topics without clarification questions are shown as solid squares—these represent topics that our intermediary found to be straightforward and required no clarification. We might expect these topics to be easier and average precision to be higher, but this does not turn out to be the case. Mean average precision over the 40 topics with clarification questions was 0.465, compared to 0.399 for those without clarification exchanges. Of course, these 40 topics received the benefit of assessor feedback, but we get similar results from B1 (the pre-clarification manual run). This suggests that clarity in the topic formulation (i.e., how precisely the information need is defined) has little to do with the prevalence of relevant documents.

6.2 Overlap Between Assessor and Intermediary Judgments

How effective was our clarification strategy based on “shuffling piles”? The size of each pile, before and after clarification, is shown in Table 3. The “avg” column shows the average size of each set across all 50 topics; the “max” and “min” columns show the maximum and minimum size of the piles. The column named “empty” shows the number of topics for which that pile was empty. For example, before clarification, only three topics had no documents in the MR pile. After the clarification dialogue, this number increased to forty-seven; that is, for all but three topics, the intermediary completely resolved the MR documents. Both the CR and NR piles increased in size at the expense of the MR piles as a result of the clarification process, but there was little change in the size of the PR piles. These results

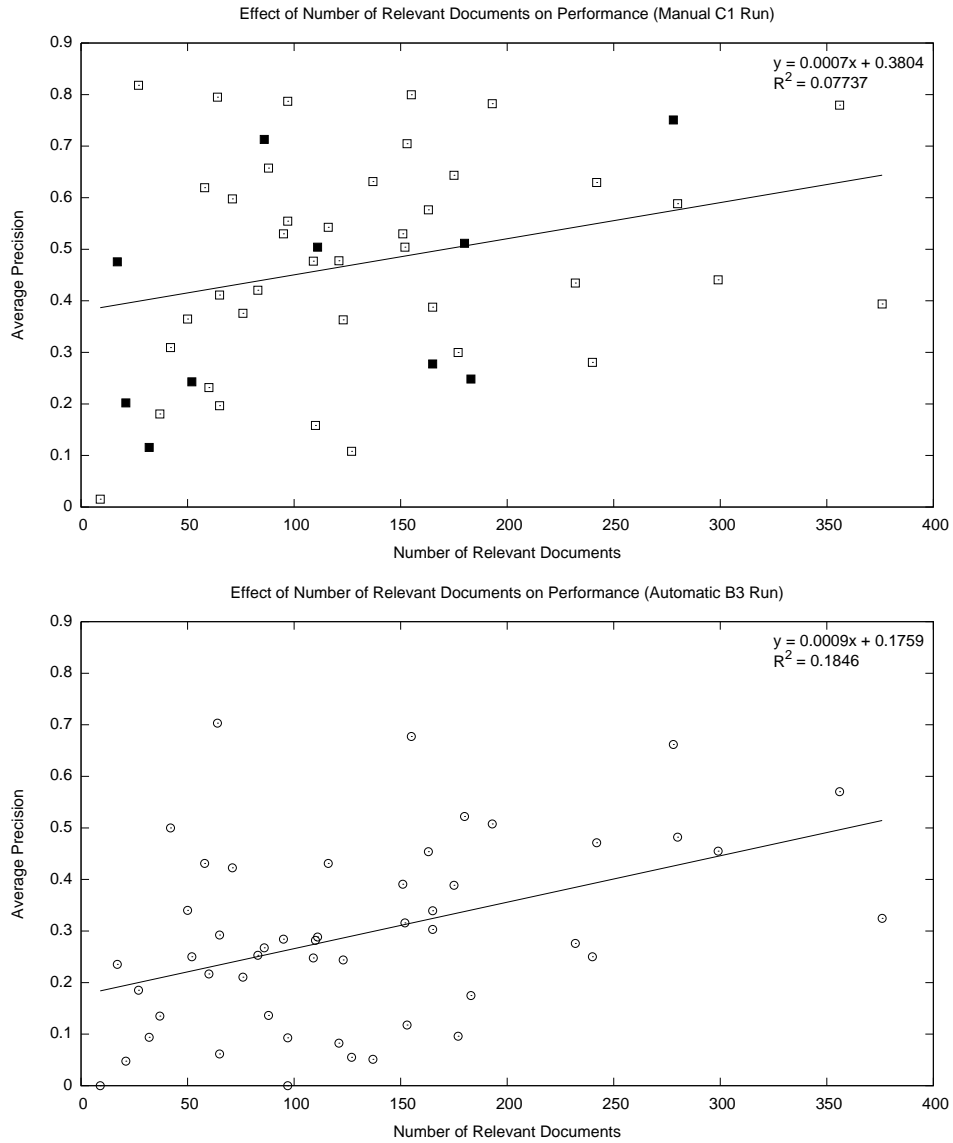


Figure 10: Scatter plots relating the number of known relevant documents to per-topic average precision: manual C1 run (top) and automatic B3 run (bottom). In the manual C1 run, topics without clarification questions are shown in solid squares.

	Pre-clarification				Post-clarification			
	avg	max	min	empty	avg	max	min	empty
CR	31.0	74	0	1	38.4	96	0	1
PR	3.7	39	0	27	3.8	30	0	22
MR	12.2	61	0	3	1.1	35	0	47
NR	28.9	63	3	0	34.0	67	0	1

Table 3: Sizes of the document piles created by the intermediary, before and after the clarification dialogue. (avg=average size per topic, max=maximum size, min=minimum size, empty=number of topics with zero documents)

	Pre-clarification			Post-clarification		
	avg size	avg rel	p	avg size	avg rel	p
CR	31.0	24.4	0.767	38.4	29.3	0.732
PR	3.7	1.2	0.210	3.8	1.2	0.259
MR	12.2	6.8	0.472	1.1	0.4	0.011
NR	28.9	2.3	0.090	34.0	4.4	0.135

Table 4: Overlap between assessors’ and the intermediary’s relevance judgments. The column p indicates the likelihood that a document was considered relevant by the assessor.

show that our clarification strategy was helpful in verifying the intermediary’s assumptions. However, we made little headway in sorting through the PR documents (i.e., determining a relevance threshold).

Although Table 3 shows the sizes of the piles, it does not actually reveal the quality of the documents in those piles. What is the probability that the assessor considered the document relevant, given that the intermediary placed it in the CR pile before and after clarification? And for the PR, MR, and NR categories? This information is shown in Table 4. The column marked “avg size” shows the average size of the piles, duplicated from the “avg” column in Table 3. The “avg rel” column shows the number of those documents that were judged relevant by the assessors. Finally, the column marked “ p ” expresses the relationship between the two columns as a probability. For reference, an average of 131.2 documents were judged relevant by the NIST assessors (based on pooling)—overall, we can see that the intermediary only found a small fraction of those. It appears that our clarification strategy increased the total number of relevant documents found, but at the expense of accuracy—that is, more relevant documents made their way into the NR piles, and more irrelevant documents made their way into the CR piles. While our clarification strategy greatly reduced the size of the MR pile, the size of the PR pile actually increased slightly. Clarification exchanges targeted at the relevance threshold did not appear to be effective (see Section 7.3).

6.3 Factors Affecting Clarification Effectiveness

In our final analysis, we conducted a topic-by-topic breakdown of the differences between runs B1 and C1 to better understand ways in which clarifications helped or hurt. As a reminder, the two runs only differed in the composition of the piles used to generate the ranked lists—the C1 piles benefited from assessor feedback, whereas the B1 piles didn’t. Thus, this comparison isolates the effects of the clarification dialogue.

We began by arbitrarily dividing the topics into five bins, according to the relative differences between pre- and post-clarification average precision: $\delta \geq 0.10$ (“helped a lot”), $0.05 \leq \delta < 0.10$ (“helped a bit”), $-0.05 \leq \delta < 0.05$ (“didn’t make much of a difference”), $-0.10 \leq \delta < -0.05$ (“hurt a

Topics	All	Subset A	Subset B
# q’s	50	40	32
pre-clarification MAP	0.452	0.465	0.482
post-clarification MAP	0.469 (+3.8%)	0.485 (+4.3%)	0.519 (+7.7%)
$\delta[+0.10, +\infty]$ “helped a lot”	8	8	8
$\delta[+0.05, +0.10)$ “helped a bit”	12	9	9
$\delta[-0.05, +0.05)$ “not much difference”	22	16	14
$\delta[-0.10, -0.05)$ “hurt a bit”	4	3	1
$\delta[-\infty, -0.10)$ “hurt a lot”	4	4	0

Table 5: Effects of clarification on different subsets of topics (subset A does not include topics without clarification questions; topics that exhibited the “inconsistent assessor” phenomenon are further excluded in subset B).

bit”), and $\delta < -0.10$ (“hurt a lot”). As can be seen in Table 5, eight topics fell in the last two bins, where clarification decreased average precision by at least 5%.

We narrowed our examination to topics for which there were clarification questions (forty topics). This is shown as subset A in Table 5. Considering this reduced set of topics, we observe a gain of 4.3% in terms of MAP (significant at the 1% level). We then manually examined each topic in order to better understand ways in which the clarification dialogue helped or hurt.

For many topics, it was easy to see why clarification dialogues improved effectiveness. A better understanding of the information need helped the intermediary make better relevance judgments. The most dramatic example of this was with topic 362 “human smuggling”, where average precision jumped from 0.405 to 0.643, a gain of 59%. The topic called for reports about incidents of human smuggling for monetary gain. The clarification questions confirmed that the element of monetary gain must be present, and that summaries of smuggling rings and smuggling statistics were not relevant.

Somewhat distressing were seven topics in which the clarification dialogue resulted in a decrease in average precision of at least 5%. For example, the average precision for topic 336 “black bear attacks” dropped 34% (from 0.466 to 0.309). To the clarification question “Does a document need to mention frequency of attacks **and** cause of attacks **and** method of control to be considered relevant?”, the assessor answered “yes”, indicating that documents with missing facets were not relevant. Our intermediary adjusted the composition of the document piles based on this feedback. However, post-hoc analysis of the final relevance judgments revealed that many documents missing the abovementioned facets were nevertheless marked relevant. In other words, the assessor’s answer to the clarification question did not match the actual criteria used in the assessment! We have dubbed this the “inconsistent assessor” phenomenon.

In fact, examining all eleven topics where clarification dialogues caused a drop in average precision revealed eight cases of the “inconsistent assessor” phenomenon. For these topics, the feedback received was misleading and contradicted the assessors’ relevance criteria as reflected in the final judgments. Results of removing these topics from subset A are shown in Table 5 as subset B. On these topics, clarification dialogues yielded an increase of 7.7% in mean average precision (significant at the 1% level). The table shows that topics in the worst-performing bin (average precision decrease greater than 10%) can all be attributed to this cause.

For the three other topics in which post-clarification average precision was lower, assessor feedback actually clouded the intermediary’s understanding of the information need, mainly due to poorly-formulated clarification questions. One question, for example, asked whether or not “details” were necessary. This being a vague term, the intermediary and assessor ultimately had different notions of

what “details” meant. Note that even with these less-than-optimal exchanges, only one topic had in a drop in average precision greater than 5%.

What is the cause of this “inconsistent assessor” phenomenon? Ruling out intentional misinformation, there are at least two sets of possibilities: one points to a methodological flaw, while the other to shifting relevance criteria unavoidable in information seeking.

Due to real-world constraints involved in coordinating the HARD track, documents were not assessed until approximately one month after the clarification questions had been answered (in order to allow ample time for participants to process assessor feedback and to prepare final runs). During this time, assessors may have already forgotten their original answers: instability in relevance criteria over long periods of time could be the source of observed inconsistencies. This was exacerbated by the fact that the 2005 HARD topics were reused from previous evaluations, which meant the information needs were not the assessors’ own.

Research in information science, however, suggests that inconsistencies in assessors’ notions of relevance may be an inescapable fact of real-world information-seeking behavior. The TREC evaluation methodology assumes a static information need against which documents are evaluated for relevance, when, in truth, information needs are themselves constantly shifting and evolving as assessors learn more about the subject (Bates, 1991; Taylor, 1962). Therefore, the mere act of participating in the clarification dialogue may have altered the assessors’ perception of their underlying needs. Furthermore, since our clarification questions were created based on documents reviewed by the intermediary, we were already circumscribing the bounds of the relevance space and subtly influencing the feedback process. Most of our clarification questions could be considered “leading”, which may influence the assessor to respond in a calculated manner that runs counter to the underlying need. Thus, “neutral” questioning is preferred in reference interviews so that the questions posed do not lead to biased responses (Dervin and Dewdney, 1986).

In addition, previous work has shown that relevance criteria are affected by examined documents (Florance and Marchionini, 1995; Sormunen, 2002). Since HARD assessors did not have access to the documents during the clarification dialogue, the inconsistency between their responses and their judgments is perhaps not surprising. In the case of the “black bear attacks” topic, since no single document contained all three facets, the assessor must have relaxed the constraint. These inconsistencies also point out another flaw in TREC evaluations: the assumption of independent relevance judgments (i.e., that each document is examined in isolation). Although an operational necessity for pooled evaluations, the assumption is clearly not true, since what assessors read in one document will affect their judgment of other documents.

7 A Taxonomy of Clarifications

One major contribution of our study is a generalization of intermediary-initiated clarification dialogues into a resource that can guide the development of future interactive retrieval systems. We proceed in four steps: Section 7.1 presents an initial taxonomy of clarification questions that captures patterns of intent in the observed exchanges. In Section 7.2, it is shown that multiple, independent coders are able to agree on categorization of the clarification questions, supporting the validity of the taxonomy. Section 7.3 describes an attempt to use the taxonomic categories to model the effectiveness of different clarification types. Finally, in Section 7.4 we refine the taxonomy of clarification types based on existing literature and discuss related work.

7.1 Clarification Types: An Initial Attempt

It is important to remember that our intermediary was not specifically instructed to apply any existing theory or taxonomy when crafting the clarification questions, other than the general idea of

Type	Topic	Example Clarification Question	Freq.
RT	(404) Ireland, Peace Talks	Would a general reference to violence without specifying particular acts be relevant?	28 (32%)
ACF	(344) Abuses of E-Mail	Does an article need to discuss both cases of email abuse and steps taken to prevent abuse to be relevant?	9 (10%)
EC	(344) Abuses of E-Mail	Would email hoaxes be considered “abuse”?	20 (23%)
CRC	(336) Black Bear Attacks	Would other species of bears (brown bear, grizzly bear...) be of interest?	12 (14%)
RTA	(341) Airport Security	Would articles about tightened security policy on airport employees be relevant?	16 (18%)
AS	(362) Human Smuggling	Would a summary of a smuggling ring be relevant?	3 (3%)

Table 6: A taxonomy of clarification questions, with examples and frequencies of occurrence as coded by the first author.

categorizing documents with respect to relevance categories. We wished to let structure, if any, emerge naturally from the data. During post-hoc analysis of the clarification dialogues, we did notice a number of patterns in the intent of the clarification questions. The first author undertook the task of inducing a taxonomy by iteratively grouping similar items. Formation of the different types occurred simultaneously with the coding of the clarification questions.

As previously discussed, we view clarification dialogues as an opportunity to better understand a user’s information need so that PR (peripherally relevant) and MR (maybe relevant) documents can be sorted into either the CR (centrally relevant) or NR (not relevant) piles. Questions targeted at the PR documents form a coherent class:

- **To determine the relevance threshold (RT).** Although relevance is a graded property, the realities of the TREC evaluation framework force users to make binary relevance judgments. Thus, each user develops a “relevance threshold” that maps a continuous scale into a binary decision. Clarification questions of this type attempt to better understand this threshold.

Other clarification questions fall into five categories, discussed below. An example of each is shown in Table 6.

- **To determine the relationship between ambiguously conjoined facets (ACF).** In most cases, information needs are composed of multiple conceptual facets. Often, the relationship between these facets is unclear, e.g., does a document need to contain all of the facets to be considered relevant?

We discovered that clarification types can be schematically illustrated in terms of structured queries—that is, a clarification typically leads to a refined query that better captures the information need. For ACF, this situation can be shown in the following, where we adopt the convention of underlining the target of clarification:

$$(A_1 \vee A_2 \vee A_3 \vee \dots) \underline{?} (B_1 \vee B_2 \vee B_3 \vee \dots) \dots$$

For the sake of illustration, let us consider the simplest case with strict Boolean queries.⁷ An ACF question asks: should facets be connected by AND or OR? In other words, these exchanges

⁷In reality, this is an over-simplification since the intermediary had access to INQUERY’s rich set of query operators and thus had more options at his disposal.

attempt to resolve ambiguous relationships between facets in the topic statements (conjunction vs. disjunction). We note that in theory many other types of relationships may hold between facets, for example, temporal precedence, cause and effect, or logical entailment. However, the difficulty is that many of these relationships cannot be readily expressed in terms of query operators, and require more sophisticated linguistic processing.

- **To determine the relevance of an example concept (EC).** Is a particular concept present in one or more documents an example of a concept referenced in the topic statement? For example, topic 347 concerns wildlife extinction: it was unclear whether documents about plants would be considered relevant by the assessor. The intermediary therefore formulated a clarification question to better understand the assessor’s notion of “wildlife”. This situation can be schematically illustrated with a Boolean query, where the underlined portion of the query is the target of clarification:

$$(A_1 \vee A_2 \vee A_3 \dots \vee \underline{A_n \vee A_{n+1} \vee A_{n+2} \dots}) \wedge (B_1 \vee B_2 \vee B_3 \vee \dots) \wedge \dots$$

The building blocks search strategy involves identifying conceptual facets and instantiating those concepts with actual query terms. EC clarification questions inquire about other instantiations of those concepts.

An interesting subclass of this type concerns so-called “meta-terms”, such as pros/cons, advantages/disadvantages, etc. For the most part, they make poor query terms, and need to be operationalized in a particular context.

- **To determine the relevance of a closely-related concept (CRC).** Does the user’s interest in a particular concept A extend to a closely-related concept A' ? A and A' may be ontologically related via hypernymy, hyponymy, antonymy, etc. This situation can be schematically shown as follows:

$$((A_1 \vee A_2 \vee A_3 \dots) \vee \underline{(A'_1 \vee A'_2 \vee A'_3)}) \wedge (B_1 \vee B_2 \vee B_3 \vee \dots) \wedge \dots$$

Once again, the underlined portion of the Boolean query is the target of clarification. Although CRC and EC questions may be similar in terms of query structure, we view them as conceptually distinct. Whereas EC involves manipulation of the manner in which a facet is expressed in a query, CRC questions inquire about the relevance of a distinct (but conceptually-related facet).

- **To determine the relevance of related topical aspects (RTA).** Is the user interested in facets that are conceptually related, but not directly requested? Topics often focus on a specific aspect of a larger concept; these questions ascertain whether users might consider other aspects of the larger concept relevant. This situation can be schematically shown as follows:

$$((A_1 \vee A_2 \vee A_3 \dots) \vee \underline{(X_1 \vee X_2 \vee X_3)}) \wedge (B_1 \vee B_2 \vee B_3 \vee \dots) \wedge \dots$$

As a concrete example, for a topic about airport security (implicitly focusing on passengers), our intermediary constructed a question that inquired about the relevance of security as it pertained to airport employees. Although the schematic representations of RTA and CRC questions are similar, we see one important difference. For CRC questions, the concept under consideration is related to a concept in the information need by an explicit ontological relation (is-a, part-whole, etc.); the connections for RTA questions are usually less direct.

- **To determine the acceptability of summaries (AS).** If the topic description indicates interest in specific instances (of events, for example), would the user be interested in a general summary or overview (e.g., aggregate statistics)?

	JL	EA	PW
JL	1.000	0.766	0.723
EA		1.000	0.692
PW			1.000

Table 7: κ values quantifying agreement among all coders.

Condition	Count
All agree	62 (70%)
Two agree	15 (17%)
All disagree	11 (13%)
Total	88

Table 8: Instances of agreement and disagreement among all coders.

Returning to the example in Figure 7: for topic 416 “Three Gorges Project”, the first author classified the clarification questions as ACF, RT, RTA, CRC. For topic 436 “Railway Accidents”, the types RTA, EC, EC were assigned by the first author.

In addition to examples, the distribution of clarification types across the topics (the forty that contained explicit clarification requests) is shown in Table 6 (as coded by the first author). It can be seen that RT questions were the most prevalent, followed by EC questions. At the other end of the spectrum, only three AS questions were observed.

7.2 Inter-Coder Agreement

Following the initial analysis by the first author, all clarification questions were subsequently coded by the two other authors independently to validate the proposed taxonomy. A description of the categories was provided as a guide.

In general, little difficulty was encountered in the coding process, as the description of all the types were easy to understand. Agreement among the coders was quantified by the κ statistic (Carletta, 1996), shown in Table 7. The κ statistic is informative in that it corrects for chance, i.e., differences in the prevalence of the clarification types. According to the literature, the values of κ obtained in our study indicate substantial agreement. Actual counts of instances where the coders agreed or disagreed are shown in Table 8. It can be seen that all three coders were in agreement 70% of the time, and 87% of the time at least two coders assigned the same type.

Based on the description of the clarification types, one might hypothesize that CRC (closely-related concepts), EC (example concept), and RTA (related topical aspects) questions are easily confusable, since they have similar query representations. In addition, they all share in a focus on individual facets. Indeed, this prediction is borne out: examining the 26 distinct cases where there was disagreement among the coders, eleven of them involved confusion between CRC, EC, and RTA. In six more cases, two of the three assessors assigned either CRC, EC, or RTA (while the third selected a different category).

In summary, not only are we able to generalize intermediary-assessor exchanges into an taxonomy of clarifications, but substantial agreement among three independent coders supports the validity of these types. Furthermore, disagreements in many cases are explained by the type semantics. These results build toward our goal of generalized principles for guiding the design of future retrieval systems.

Type	Freq.	β	p -value
Relevance Threshold (RT)	32%	0.025	0.19
Ambiguously Conjoined Facets (ACF)	7%	0.010	0.85
Example Concept (EC)	25%	0.034	0.12
Closely-Related Concept (CRC)	15%	0.106	< 0.01
Related Topical Aspect (RTA)	16%	~ 0	0.99
Acceptability of Summaries (AS)	4%	0.214	<< 0.01

Table 9: Linear regression with counts of each clarification type in a topic as independent variables and relative average precision gain as the dependent variable. The column marked Freq. shows the prevalence of each type. This regression model achieves an R^2 value of 0.66 (adjusted R^2 of 0.56).

7.3 Effectiveness of Different Clarification Types

One advantage of the HARD experimental setup is the ability to isolate factors that may impact ranked-retrieval effectiveness. In particular, since we submitted both pre-clarification and post-clarification runs, it was possible to quantify the effectiveness of the clarification dialogue (see Section 5.2). Given the taxonomy just presented, we can take the analysis one step further: it is possible to construct a fine-grained model of effectiveness at the level of individual question types. An interactive retrieval system could then apply such a model for dialogue planning and management.

We constructed a simple linear regression model to capture the relationship between different clarification types and retrieval effectiveness, as measured by average precision before and after the clarification dialogue. To start, we used the original categories shown in Table 6 (as determined by the first author). The numbers of clarification questions in each category served as the independent variables (predictors) and the relative difference in average precision served as the dependent variable.⁸ The intercept of the regression model was fixed to zero, since intuitively asking no questions should yield no score difference. We used the 32 topics denoted as subset B in Table 5, which does not contain topics without clarification questions or topics that exhibit the “inconsistent assessor phenomenon”.

Overall, our regression model was statistically significant, with an R^2 value of 0.66 (adjusted R^2 of 0.56). Regression coefficients for each variable are shown in Table 9, along with their p -values. The frequency of each clarification type is also shown; note these values are slightly different than the figures in Table 6 since the regression model was built using a subset of topics. Positive values for all regression coefficients confirm our expectation that asking clarification questions correlates positively with increased average precision (although to different degrees). Because the number of topics used in this analysis was relatively small, these results should be taken as indicative, not conclusive.

Of all clarification categories, AS (acceptability of summaries) was found to be the most significant predictor of improvements in average precision and has the largest regression coefficient. This suggests that, when appropriate, AS questions are highly effective—a result that makes sense since the answer to such a question would determine the relevance of a potentially large number of documents. It is interesting to note that, even in TREC, which focuses on topical relevance, non-topical factors such as this also affects retrieval effectiveness. Furthermore, this finding illustrates how the answer to a single question can have a large impact on the composition of the result set. Effective interactions need not be complex—they simply must be to the point.

Clarifications that inquire about closely-related concepts (CRC) were also found to be a statistically significant predictor of average precision gain. The other two similar types, which also aim to clarify individual concepts—EC (example concept) and RTA (related topical aspect)—were not found to be

⁸We also tried using the presence or absence of each clarification type as the independent variables, although the model fit was not as good.

Topical	
Relevance Threshold	RT
Conceptual Facets	EC, CRC, RTA
Relationship Between Facets	ACF
Non-Topical	
Acceptability of Summaries	AS

Table 10: A refined taxonomy of clarifications, illustrating correspondences to categories in Table 9.

statistically significant predictors.

According to our regression model, the ACF (ambiguously conjoined facets) type was not a statistically significant predictor of average precision gain. Recall that, in our clarification strategy, feedback in these cases informed the intermediary whether to treat the facets as conjunctive or disjunctive. The discussion in Section 6.3 shows that the interaction between relevance criteria and retrieved documents is far more complex than this simply dichotomy.

Finally, we discovered that RT (relevance threshold) questions were not particularly helpful, despite their prevalence. While it is important to map users’ scale of relevance to the intermediary’s relevance categories, our questions appeared to be ineffective. This result is consistent with the findings presented in Table 4. The size of the PR set did not change much after the clarification dialogue (in fact, it even increased slightly)—suggesting that answers to RT questions were not useful in helping the intermediary resolve documents in the PR pile. Future work might, for example, test different formats of RT questions such as employing Likert scales or asking for explicit rankings.

We conclude the present discussion with a caveat: it is important to keep in mind that the types of clarification questions possible are dictated by the nature of the information need. For example, although AS (acceptability of summary) questions were found to be effective, they are certainly not applicable for every topic. Thus, the model described here can only serve as a guide and must be adapted to individual circumstances.

7.4 Refinement and Discussion

Conceptual facet analysis provides a framework with which to understand the nature and purpose of clarification exchanges, and how subsequent queries can be modified to better capture users’ information needs. We found that elements of the building blocks strategy can be used to refine and extend our taxonomy of clarifications by providing theoretical top-down guidance. Interpreting empirical evidence in the context of well-established work in library and information science further helps to validate generalizations from our study.

One possible refinement is shown in Table 10, which subsumes the types in Table 6. Drawing insights from White (1985), we broadly classified clarification interactions as either topical or non-topical. Furthermore, we organized topical clarifications into three major categories: relevance threshold, conceptual facets, and relationship between facets.

- The first, relevance threshold, aims to help the system map between different scales of relevance. Although binary judgments in TREC are an artifact of the evaluation setup, there are realistic cases in which graded quantities must be collapsed into binary decisions. Ultimately, a system must determine whether or not to return a document; a user must decide whether or not to examine a search result in greater detail. Relevance threshold clarifications help a system convert graded scales into binary categories.
- The second interaction type aims to clarify the conceptual facets present in an information

need. This encompasses what we previously called example concept (EC), closed-related concepts (CRC), and related topical aspects (RTA). Interactions falling under this broad category focus on one particular facet of the information need. In general, most of the clarification questions asked in our experiments involved *broadening* the scope of the concept, although in principle it is possible for clarification questions to achieve the opposite effect, i.e., to narrow the scope. For example, in response to a topic about pandas, the intermediary might confirm that the user was referring to the giant panda, and not the red panda, its lesser-known cousin.

- The third class of topical clarification involves the relationship between facets, of which the previously-identified ACF type is one example. In the building blocks strategy, identification and analysis of conceptual facets are followed by the instantiation of search queries. One important consideration is the relationship between facets and how they should be expressed in terms of available query operators. A simple example would be to select between disjunctive and conjunctive interpretations of the facet structure (as in ACF). However, more nuanced relationships between facets are certainly possible: the user, for example, might specifically be interested in causality or temporal precedence (which may require deeper linguistic analysis). Furthermore, there is nothing to prevent users from having information needs that would translate into nested structures, such as “the mating and hunting habits of black and grizzly bears”. In these situations, there may exist a need to better understand the user’s intentions.

Clarification exchanges that focus on non-topical characteristics of the information need comprise another category of interaction. Although we have only noted one such type in our experiments (AS, acceptability of summaries), this most likely stems from the TREC framework, which is primarily concerned with topical relevance. However, studies of need negotiation and reference interviews could predict many other possibilities: inquiring about temporal and source constraints in the publication, suitability for different audiences (layman vs. expert, for example), etc. Since our experiments were conducted in the context of TREC and inherits its limitations, this work does not provide much guidance to further refine subcategories of non-topical clarification interactions.

To examine the predictive power of this refined clarification taxonomy, we grouped EC, CRC, and RTA interactions into one category and built a separate linear regression model. This grouping reflects the idea that those three categories represent concept-level clarifications, and aligns the independent variables with the categories in Table 10. The resulting model achieves an R^2 value of 0.59 (adjusted R^2 of 0.51)—although the fit is not as good as the original model, it still captures a substantial amount of variance. More importantly, however, the combined category of concept clarifications was found to be a statistically significant predictor of average precision gain. This suggests that concept clarifications as a whole form the basis of effective interactions.

We end this section with a discussion of related work, focusing particularly on previous studies that have analyzed and categorized elicitations in face-to-face reference encounters, e.g., (Nordlie, 1999; Spink et al., 1996; Spink, 1997; Swigger, 1985; White, 1998). How are these different from our study? We see several important distinctions:

First, the TREC setup narrows the scope of inquiry, which limits the realism of the problems examined, but increases computational tractability. Instead of the full range of reference exchanges that may occur within a lengthy information-seeking scenario spanning multiple iterations, the HARD track focuses on one (relatively brief) interactive feedback cycle. It is worth noting that in the asynchronous communication channels modeled by HARD (e.g., email), a small number of interaction cycles better captures real-world constraints (e.g., user patience). Nevertheless, this compromise in study design paves the way for large-scale, multi-system evaluations. More importantly, the HARD setup allows researchers to quantify pre- and post-clarification effectiveness, which is something that previous studies have been unable to do. Within the HARD framework, we were able to compare our techniques against other systems and we were also able to quantify the effectiveness of different interaction types.

Ultimately, the controlled setup of TREC experiments potentially allows researchers to make stronger conclusions. The downside, of course, is that we are restricted by the TREC methodology and unable to study factors not directly modeled by the setup, e.g., non-topical aspects of relevance.

The second point relates to the first: since most previous studies were observational in nature, they had little control over aspects of intermediary behavior. For example, two previous studies (Spink et al., 1996; Spink, 1997) were based on analyses of forty transcribed searches with professional intermediaries in an academic library environment. We can only speculate about the effectiveness of a “baseline” that does not involve the intermediary (i.e., had there been an unattended search terminal in the library) or alternative results had the intermediary taken different actions. Not only did our HARD experiments support comparisons between pre- and post-clarification results, but we also had control over the strategy employed by the intermediary, which allowed us to actively shape the interactions.

Finally, by replacing unconstrained face-to-face reference exchanges with stylized asynchronous clarification dialogues mediated by Web forms, we increase the applicability of findings to automated systems. The HARD setup eliminates the subtle nuances of human communication, which are merely distractions for our purpose since there is little hope that automated systems can capture the richness of human–human interactions (e.g., active listening), at least in the foreseeable future. For all intents and purposes, NIST assessors in the HARD track believed that they were interacting with machines, even though a human intermediary was involved in generating the submitted runs. Generalizing from these “disembodied” interactions actually yields a more realistic guide for the design of future interactive retrieval systems.

8 Mediated Search as a Retrieval Model

The starting point of this study is the hypothesis that interactions between a trained human search intermediary and an information seeker can inform the design of interactive IR systems. In the end, we believe that our work supports a stronger claim: that our strategy for mediated search provides a good model for designing interactive IR systems. In this section, we explain how many IR techniques can be understood in terms of conceptual facet analysis and our proposed taxonomy of clarifications. We believe that this general framework can serve as a guide for designers of future systems. This discussion builds on four specific findings and associated design recommendations:

1. Conceptual facet analysis is a general yet powerful strategy for decomposing complex information needs. Systems should provide some type of support for this method of analysis.
2. The relations between concepts in an information need are often complex. Systems should provide mechanisms for explicitly specifying these relations and selecting documents in which they hold.
3. The mappings between concepts and query terms are often complex. Systems should provide mechanisms for managing these mappings.
4. Information derived from users’ statements of need may not represent precisely what they are looking for.

Frame-based retrieval strategies represent computational implementations of conceptual facet analysis, since frames essentially capture fixed facet structures. Although early work in IR (Croft and Lewis, 1987; Smith et al., 1989) along these lines have not been successful, recent attempts in question answering (Harabagiu et al., 2000; Small et al., 2004; Demner-Fushman and Lin, 2007) have demonstrated the effectiveness of frame-based retrieval. Recent work on structured queries (Bilotti et al., 2007) lends additional support to the effectiveness of automatic facet analysis, since there is often a correspondence between query clauses and facets.

Much recent work in IR acknowledges the need to go beyond “bag of words” and explicitly capture relations between terms, named entities, concepts, etc. This echoes our second point. Examples include work on term dependence models, both those that are linguistically motivated (Gao et al., 2004) and those that are not (Metzler and Croft, 2005). In the context factoid question answering, matching of questions with candidate answers at the level of syntactic relations has been successful (Cui et al., 2005): these techniques handle the special case of facet relations that manifest linguistically. Alternatively, facet relationships can be visualized in user interfaces, e.g., TileBars (Hearst, 1995) and the interface described by Veerasamay and Belkin (1996).

Most work in interactive IR has focused on query formulation and overcoming imprecise descriptions of need, which correspond to the third and fourth points. One effective strategy has been to provide users with explicit control over query terms (Koenemann and Belkin, 1996), but it would be preferable to operate at the level of concepts, in conjunction with the related work discussed above. The integration of named-entity detection with information retrieval in the context of factoid question answering (Prager, 2007) provides a successful example of managing concept-term mappings. Factoid question answering systems primarily deal with the subcase where concepts are represented by entities such as people, organizations, dates, etc. Named-entity recognizers allow systems to process text at the level of semantic types, abstracting away from considerable variation in the expression of those types. As an interesting extension, named entities themselves can form the basis of effective interactions (Small et al., 2004; Toda et al., 2007).

Document clustering techniques (Hearst and Pedersen, 1996; Leuski and Allan, 2000) represent another approach to overcoming imprecise need descriptions and the mismatches between concepts and query terms. Clusters naturally suggest related terms, but users often find it difficult to understand what a cluster is “about”. We argue that approaches based on well-defined semantic categories are much more effective, e.g., (Dumais et al., 2001; Demner-Fushman and Lin, 2006; Kules et al., 2006). In particular, faceted browsing techniques (Yee et al., 2003) have the additional advantage in providing facet analysis, thus supporting an integrated mechanism for need decomposition and interaction.

Overall, our model of mediated search is valuable in providing an integrated view of both retrieval algorithms and interaction techniques, drawing connections between work in different areas. These ideas hopefully provide guidance for designers of future interactive retrieval applications.

9 Conclusion

This work provides support for the hypothesis that interactions between a trained human search intermediary and an information seeker can inform the design of interactive IR systems. Our argument begins with a demonstration of the value of mediated search, using the TREC 2005 HARD track as an experimental vehicle. We found that placing a human in the loop yields significant increases in terms of standard ranked-retrieval metrics—a result that is by no means obvious. However, this observation is of little value without an analysis of what worked and why, so that we may learn from the experiments. The second part of this article focuses on analysis of our search and interaction strategy to arrive at a series of generalizations and recommendations for system design. One concrete product is a taxonomy of clarification questions, along with a characterization of the effectiveness of different clarification types. Furthermore, we argue that our strategy for mediated search provides a good model for designing interactive IR systems. By demonstrating that it is possible to build better information retrieval applications by studying human information-seeking processes, we hope to bridge system-centered and user-centered perspectives to the same problem.

10 Acknowledgments

This work has been supported in part by DARPA cooperative agreement N66001-00-2-8910 and contract HR0011-06-2-0001 (GALE). We would like to thank James Allan for organizing the HARD track; Doug Oard for various engaging discussions; Sheri Massey for help on issues related to qualitative methodology; and three anonymous reviewers whose challenging comments have helped us immensely in improving this article. The first author would like to thank Esther and Kiri for their loving support.

References

- Eileen G. Abels. 1996. The e-mail reference interview. *Reference Quarterly*, 35(3):345–358.
- James Allan, Ben Carterette, and Joshua Lewis. 2005. When will information retrieval be “good enough”? User effectiveness as a function of retrieval accuracy. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 433–440, Salvador, Brazil.
- James Allan. 2003. HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 24–37, Gaithersburg, Maryland.
- James Allan. 2005. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 51–67, Gaithersburg, Maryland.
- Marcia J. Bates. 1991. The Berry-Picking search: User interface design. In M. Dillon, editor, *Interfaces for Information Retrieval and Online Systems: The State of the Art*, pages 51–61. Greenwood Press, New Jersey.
- Micheline Beaulieu, Stephen Robertson, and Edie Rasmussen. 1996. Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1):85–94.
- Micheline Beaulieu, Thien Do, Alex Payne, and Susan Jones. 1997. ENQUIRE Okapi Project. British Library Research and Innovation Report 17.
- Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395.
- Nicholas J. Belkin, Colleen Cool, Judith Head, Judy Jeng, Diane Kelly, Shin jeng Lin, Lynne Lobash, Soyeon Park, Pamela A. Savage-Knepshield, and Cynthia Sikora. 1999. Relevance feedback versus Local Context Analysis as term suggestion devices: Rutgers’ TREC-8 interactive track experience. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 565–573, Gaithersburg, Maryland.
- Matthew W. Bilotti, Paul Ogilvie, Jamie Callan, and Eric Nyberg. 2007. Structured retrieval for question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 351–358, Amsterdam, The Netherlands.
- David C. Blair and M. E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299.

- Richard E. Bopp and Linda C. Smith. 1995. *Reference and Information Services: An Introduction*. Libraries Unlimited, Englewood, Colorado, 2nd edition.
- Peter Bruza, Robert McArthur, and Simon Dennis. 2000. Interactive Internet search: Keyword, directory and query reformulation mechanisms compared. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 280–287, Athens, Greece.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27.
- Cyril W. Cleverdon, Jack Mills, and E. Michael Keen. 1968. Factors determining the performance of indexing systems. Two volumes, ASLIB Cranfield Research Project, Cranfield, England.
- Bruce Croft and David D. Lewis. 1987. An approach to natural language processing for document retrieval. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1987)*, pages 26–32, New Orleans, Louisiana.
- Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 400–407, Salvador, Brazil.
- Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 841–848, Sydney, Australia.
- Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.
- Brenda Dervin and Patricia Dewdney. 1986. Neutral questioning: A new approach to the reference interview. *Reference Quarterly*, 25(4):506–513.
- Brenda Dervin. 1991. From the mind’s eye of the user: The sense-making qualitative-quantitative methodology. In J. D. Glazier and R. R. Powell, editors, *Qualitative Research in Information Management*, pages 61–84. Libraries Unlimited, Inc., Englewood, Colorado.
- Susan Dumais, Edward Cutrell, and Hao Chen. 2001. Optimizing search by showing results in context. In *Proceedings of SIGCHI 2001 Conference on Human Factors in Computing Systems (CHI 2001)*, pages 277–284, Seattle, Washington.
- Efthimis N. Efthimiadis and Stephen E. Robertson. 1989. Feedback and interaction in information retrieval. In Charles Oppenheim, editor, *Perspectives in Information Management*, pages 257–272. Butterworth, London, England.
- Efthimis N. Efthimiadis. 2000. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003.

- Valerie Florance and Gary Marchionini. 1995. Information processing in the context of medical care. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995)*, pages 158–163, Seattle, Washington.
- Stephen Francoeur. 2001. An analytical survey of chat reference services. *Reference Services Review*, 29(3):189–204.
- Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence language model for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 170–177, Sheffield, United Kingdom.
- Sanda Harabagiu, Marius Paşca, and Steven Maiorano. 2000. Experiments with open-domain textual question answering. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 292–298, Saarbrücken, Germany.
- Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan. 2005. Experiments with interactive question-answering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 205–214, Ann Arbor, Michigan.
- Donna K. Harman. 1988. Towards interactive query expansion. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1988)*, pages 321–331, Grenoble, France.
- Donna K. Harman. 2005. The TREC test collections. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 21–52. MIT Press, Cambridge, Massachusetts.
- Stephen Harter. 1986. *Online Information Retrieval: Concepts, Principles, and Techniques*. Academic Press, San Diego, California.
- Donald T. Hawkins and Robert Wagers. 1982. Online bibliographic search strategy development. *Online*, 6(3):12–19.
- Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*, pages 76–84, Zürich, Switzerland.
- Marti A. Hearst. 1995. TileBars: a visualization of term distribution information in full text information access. In *Proceedings of SIGCHI 1995 Conference on Human Factors in Computing Systems (CHI 1995)*, pages 59–66, Denver, Colorado.
- William R. Hersh and Paul Over. 2001. Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management*, 37(3):365–367.
- William R. Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. 2000. Do batch and user evaluations give the same results? In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 17–24, Athens, Greece.
- Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50.
- Peter Ingwersen. 1999. Cognitive information retrieval. *Annual Review of Information Science and Technology*, 34(1):3–52.

- John F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Office Information Systems*, 2(1):26–41.
- Diane Kelly and Xin Fu. 2006. Elicitation of term relevance feedback: An investigation of term source and context. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 453–460, Seattle, Washington.
- Diane Kelly, Vijay Deepak Dollu, and Xin Fu. 2005. The loquacious user: A document-independent source of terms for query expansion. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 457–464, Salvador, Brazil.
- Sara D. Knapp. 1978. The reference interview in the computer-based setting. *Reference Quarterly*, 17(4):320–324.
- Jürgen Koenemann and Nicholas J. Belkin. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of SIGCHI 1996 Conference on Human Factors in Computing Systems (CHI 1996)*, pages 205–212, Vancouver, British Columbia, Canada.
- Bill Kules, Jack Kustanowitz, and Ben Shneiderman. 2006. Categorizing Web search results into meaningful and stable categories using fast-feature techniques. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2006)*, pages 210–219, Chapel Hill, North Carolina.
- Anton Leuski and James Allan. 2000. Strategy-based interactive cluster visualization for information retrieval. *International Journal on Digital Libraries*, 3(2):170–184.
- Jimmy Lin and Mark D. Smucker. 2008. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore.
- Jimmy Lin and W. John Wilbur. 2007. PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8:423.
- Jimmy Lin, Philip Wu, Dina Demner-Fushman, and Eileen Abels. 2006. Exploring the limits of single-iteration clarification dialogs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 469–476, Seattle, Washington.
- Jimmy Lin, Michael DiCuccio, Vahan Grigoryan, and W. John Wilbur. 2008, in press. Navigating information spaces: A case study of related article search in pubmed. *Information Processing and Management*.
- Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, England.
- Sharan B. Merriam. 1998. *Qualitative Research and Case Study Applications in Education*. Jossey-Bass, San Francisco, California.
- Donald Metzler and W. Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 472–479, Salvador, Brazil.

- Ragnar Nordlie. 1999. “User revelation”—a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 11–18, Berkeley, California.
- Paul Over. 2001. The TREC interactive track: An annotated bibliography. *Information Processing and Management*, 37(3):369–381.
- Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychology Review*, 106(4):643–675.
- John Prager. 2007. Open-domain question-answering. *Foundations and Trends in Information Retrieval*, 1(2):91–231.
- Ian Ruthven. 2003. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 213–220, Toronto, Canada.
- Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297.
- Tefko Saracevic. 1975. Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343.
- Tefko Saracevic. 1997a. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of 60th Annual Meeting of the American Society for Information Science and Technology*.
- Tefko Saracevic. 1997b. Users lost: Reflections on the past, future, and limits of information science. *SIGIR Forum*, 31(2):16–27.
- Sharon Small, Tomek Strzalkowski, Ting Liu, Sean Ryan, Robert Salkin, Nobuyuki Shimizu, Paul Kantor, Diane Kelly, Robert Rittman, and Nina Wacholder. 2004. HITIQA: towards analytical question answering. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1291–1297, Geneva, Switzerland.
- Philip J. Smith, Steven J. Shute, and Mark H. Chignell. 1989. In search of knowledge-based search tactics. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1989)*, pages 3–10, Cambridge, Massachusetts.
- Mark Smucker and James Allan. 2006. Find-Similar: Similarity browsing as a search tool. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 461–468, Seattle, Washington.
- Eero Sormunen. 2002. Liberal relevance criteria of TREC—counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 324–330, Tampere, Finland.
- Karen Sparck Jones. 2000. Further reflections on TREC. *Information Processing and Management*, 36(1):37–85.
- Amanda Spink and Howard Greisdorf. 2001. Regions and levels: Mapping and measuring users’ relevance judgments. *Journal of the American Society for Information Science and Technology*, 52(2):161–173.

- Amanda Spink and Tefko Saracevic. 1997. Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science*, 48(8):741–761.
- Amanda Spink, Abby Goodrum, David Robins, and Mei Mei Wu. 1996. Elicitations during information retrieval: Implications for IR system design. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*, pages 120–127, Zürich, Switzerland.
- Amanda Spink. 1997. Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science*, 48(5):382–394.
- Don R. Swanson. 1977. Information retrieval as a trial-and-error process. *Library Quarterly*, 47(2):128–148.
- Keith Swigger. 1985. Questions in library and information science. *Library and Information Science Research*, 7:369–383.
- Robert S. Taylor. 1962. The process of asking questions. *American Documentation*, 13(4):391–396.
- Robert S. Taylor. 1968. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29:178–194.
- Hiroyuki Toda, Ryoji Kataoka, and Masahiro Oku. 2007. The usefulness of dynamically categorizing search results. *International Journal of Human–Computer Interaction*, 23(1–2):3–23.
- Andrew Turpin and William R. Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 225–231, New Orleans, Louisiana.
- Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 11–18, Seattle, Washington.
- Howard Turtle and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222.
- Aravindan Veerasamy and Nicholas J. Belkin. 1996. Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*, pages 85–92, Zürich, Switzerland.
- Ellen M. Voorhees and Donna K. Harman. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–23, Gaithersburg, Maryland.
- Ellen M. Voorhees. 2005. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 81–89, Gaithersburg, Maryland.
- Nina Wacholder, Diane Kelly, Paul Kantor, Robert Rittman, Ying Sun, Bing Bai, Sharon Small, Boris Yamrom, and Tomek Strzalkowski. 2007. A model for quantitative evaluation of an end-to-end question-answering system. *Journal of the American Society for Information Science and Technology*, 58(8):1082–1099.
- Marilyn D. White. 1985. Evaluation of the reference interview. *Reference Quarterly*, 25(1):76–83.

- Marilyn D. White. 1998. Questions in reference interviews. *Journal of Documentation*, 54(4):443–465.
- W. John Wilbur and Leona Coffee. 1994. The effectiveness of document neighboring in search enhancement. *Information Processing and Management*, 30(2):253–266.
- Tom D. Wilson. 1999. Models in information behaviour research. *Journal of Documentation*, 55(3):249–270.
- Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of SIGCHI 2003 Conference on Human Factors in Computing Systems (CHI 2003)*, pages 401–408, Ft. Lauderdale, Florida.
- Robert K. Yin. 2003. *Case Study Research: Design and Methods*. Sage Publications, Newbury Park, California.
- ChengXiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pages 403–410, Atlanta, Georgia.