Novelty and Diversity in Information Retrieval Evaluation

Charles L. A. Clarke Maheedhar Kolla Gordon V. Cormack Olga Vechtomova
Azin Ashkan Stefan Büttcher Ian MacKinnon
University of Waterloo

ABSTRACT

Evaluation measures act as objective functions to be optimized by information retrieval systems. Such objective functions must accurately reflect user requirements, particularly when tuning IR systems and learning ranking functions. Ambiguity in queries and redundancy in retrieved documents are poorly reflected by current evaluation measures. In this paper, we present a framework for evaluation that systematically rewards novelty and diversity. We develop this framework into a specific evaluation measure, based on cumulative gain. We demonstrate the feasibility of our approach using a test collection based on the TREC question answering track.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Experimentation, Human Factors, Measurement

1. INTRODUCTION

For a given query, an information retrieval system should respond with a ranked list that respects both the breadth of available information and any ambiguity inherent in the query. The query "jaguar" represents a standard example of an ambiguous query. In responding to this query, an IR system might best return a mixture of documents discussing the cars, the cats, and the classic Fender guitar. Taken together, these documents should provide a complete picture of all interpretations.

Ideally, the document ordering for this query would properly account for the interests of the overall user population. If cars were more popular than cats, it might be appropriate to devote the first few documents to them, before switching topics. The earlier documents might cover key aspects of each topic. Later documents would supplement this basic information, rather than redundantly repeating the same thing over and over again.

The creation of an IR system that systematically accounts for redundancy and ambiguity presents many challenges. Not the least of these challenges is the lack of a clear and meaningful objective function defining what the optimal response for a given query should be under such circumstances. The evaluation measures in widespread use — such as MAP, bpref [8] and nDCG [20] — assume that the relevance of each document can be judged in isolation, independently of other documents. Tuning IR systems to optimize these evaluation measures may produce unsatisfactory results when redundancy and ambiguity are considered.

The presence of duplicates and near-duplicates in a document collection represents an extreme version of the problem. Bernstein and Zobel [5] examined the impact of near-duplicates on the TREC GOV2 collection. This test collection comprises roughly a half-terabyte of Web pages taken from the `gov` domain, along with topics and corresponding judgments. More than 17% of documents within this collection are essentially duplicates of other documents. Returning identical versions of a relevant document may produce a high score on a standard evaluation measure, but would certainly be viewed unfavorably by a user. The expedient solution of deleting these duplicates merely sweeps the problem under the rug. Instead, the evaluation measure itself should directly accommodate the possibility of duplicates.

The problem is exacerbated by the application of machine learning techniques to IR systems [1, 9, 30]. These systems may learn their ranking functions from masses of relevance judgments and implicit user feedback. To be applicable in these environments, an evaluation measure must reflect genuine user requirements. While many machine learning techniques do not directly optimize an evaluation measure, it still must be possible to compute the measure rapidly and mechanically, without the need for additional judging, even when previously unseen documents are surfaced.

This paper builds on a thread of related ideas stretching back more than four decades [6, 11, 13, 17, 32]. Many of the central ideas presented in this paper have been expressed in various forms by this earlier work. Our aim is to codify these ideas into a coherent foundation that properly accounts for redundancy and ambiguity. The resulting framework allows us to make a precise distinction between *novelty* — the need to avoid redundancy — and *diversity* — the need to resolve ambiguity.

A second aim is to demonstrate the practical application of this framework to the construction of test collections and

SIGIR'08, July 20–24, 2008, Singapore.

Page title	URL
1. UPS Global Home	www.ups.com
2. UPS: Tracking Information	www.ups.com/tracking/tracking.html
3. Uninterruptible power supply - Wikipedia,...	en.wikipedia.org/wiki/Uninterruptible_power_supply
4. The UPS Store: Retail packing, shipping,...	www.theupsstore.com
5. University of Puget Sound :: Home	www.ups.edu

Table 1: Possible results for the Web query "UPS".

evaluation measures. Since graded relevance arises naturally from the framework, we base our proposed evaluation measure on the Normalized Discounted Cumulative Gain (nDCG) measure developed by Järvelin and Kekäläinen [20], which assumes graded relevance.

In addition, we describe a test collection exploring our proposal based on the TREC 2005/2006 question answering collections. We use this collection to examine the impact of pseudo-relevance feedback on novelty. Under traditional evaluation measures, pseudo-relevance feedback generally provides a significant performance gain. The opposite is true under our proposed measure.

2. RELATED WORK

In 1964, Goffman [17] recognized that the relevance of a document must be determined with respect to the documents appearing before it. This recognition was echoed by Boyce [6] in 1982, who wrote, "The most relevant document should be topical, novel... The change it makes in the knowledge state must then be reflected in the choice of document for the second position."

Several recent papers directly inspired our work. Carbonell and Goldstein [11] describe the *maximal marginal relevance* method, which attempts to maximize relevance while minimizing similarity to higher ranked documents. Zhai and colleagues [31, 32] develop and validate subtopic retrieval methods based on a risk minimization framework and introduce corresponding measures for subtopic recall and precision. Chen and Karger [13] describe a retrieval method incorporating negative feedback in which documents are assumed to be *not relevant* once they are included in the result list, with the goal of maximizing diversity.

Spärck Jones et al. [25] call for the creation of evaluation methodologies and test collections addressing the problem of query ambiguity. They stress that a response from an IR system must accommodate multiple user needs. A user study by Xu and Yin [29] suggests that "novelty seeking" is not equivalent to "diversity seeking", and that the novelty preferences of individual users are directed towards finding more information on specific subtopics of interest, rather than an undirected quest for any new information.

3. EXAMPLES

As motivation, and to provide running examples, we present retrieval results for two queries: one taken from a Web search context and the other from the TREC question answering track.

3.1 Web Search Example

Table 1 shows five of the top ten results for the Web query "UPS", as returned by a leading commercial search engine in late 2007. We have retained the ordering specified by the search engine, but have removed a few results to keep the example concise.

The ambiguity is obvious. A user entering this query might be tracking a package sent via the United Parcel Service, planning the purchase of an uninterruptible power supply, or searching for the home page of the University of Puget Sound. The correct expansion of the acronym depends on the intent of the user. This intent may correspond to any of the standard Web query types [7]: navigational (seeking a home page), informational (seeking information on power supplies), or transactional (seeking a form to track a package).

It is difficult to argue that any one of these five pages is more relevant than any other. For all of them, there is a group of users for which it is the best result. Under the topic development methodology used by TREC, and by similar experimental efforts, relevance judgments would depend on the details of the topic narrative, details which are hidden from the IR system. Depending on these hidden details, any of these documents might be judged relevant or non-relevant. Naturally, because of the glaring ambiguity, a topic based on this query would not be accepted for inclusion at TREC, allowing the problem to be avoided. Unfortunately, the problem cannot be avoided in practice.

One possible guide for ranking these pages is the relative sizes of the groups for which they would be relevant. At a guess, the group of users intending the United Parcel Service is substantially larger than the group intending the University of Puget Sound, even within Washington State. The number of users interested in uninterruptible power supplies may fall somewhere in between.

The ordering in Table 1 is consistent with this guess. But note that a page related to uninterruptible power supplies (#3) lies between two pages related to the United Parcel Service. This arrangement may be justified by assuming that users interested in uninterruptible power supplies form a plurality of the users still scanning the result list at that point. By the fifth result, users interested in the university may form a plurality. Thus, diversity in the results proceeds directly from the needs of the user population.

Assuming that Table 1 gives the best possible ranking (which it may not) it can be justified informally and intuitively. It should be possible for our evaluation measure to reflect this intuition, assigning the highest score to precisely this ranking.

3.2 Question Answering Example

Our second example is based on a topic taken from the TREC 2005 question answering task [28]. In this task, questions were grouped into series, with a single target associated with each of these series. Figure 1 gives the target and questions for topic 85: "Norwegian Cruise Lines (NCL)". The goal of a participating QA system was to provide exact

85: Norwegian Cruise Lines (NCL)
85.1: Name the ships of the NCL.
85.2: What cruise line attempted to take over NCL in 1999?
85.3: What is the name of the NCL's own private island?
85.4: How does NCL rank in size with other cruise lines?
85.5: Why did the Grand Cayman turn away a NCL ship?
85.6: Name so-called theme cruises promoted by NCL.

Figure 1: TREC 2005 question answering topic 85

answers to these questions, when given both the target and the question.

We view this topic in a different light, treating the target as a query, and the questions as representatives or examples of the information a user may be seeking. Table 2 presents the results of executing the target as a query using a typical implementation of the BM25 scoring formula [26]. The corpus is the same AQUAINT collection of newspaper articles used at TREC. The titles of the top ten documents are shown. For each article, the table indicates the questions answered by that article, according to the official TREC judgments. For the purpose of this example, we consider a document to answer question 85.1 if it lists the name of any NCL ship. The last column gives the total number of questions answered.

While these questions certainly do not cover all aspects of the topic, we might view them as reasonable representatives. From this viewpoint, we might base overall document relevance on these questions, treating the total number answered as a graded relevance value. Therefore, if we consider only the number of questions answered, one "ideal" ordering for the documents would be a-e-b-c-f-g-h-d-i-j, with those documents answering two questions placed before those answering one.

If we consider novelty, our ideal ordering would place document g third, ahead of other documents answering one question, since only document g answers question 85.3. Moreover, the ordering a-e-g covers all the questions, with the exception of question 85.5, which is not answered by any document. The other documents might then be considered non-relevant, since they add nothing new.

However, since these other documents likely contain aspects not covered by the questions, we should not just stop at the third document. In addition, the judgments may contain errors, or the document may not fully answer an indicated question. Given the information available, we might complete our ranking by considering the number of times each question is answered. Document b (answering 85.2) might be ranked after document g, followed by document f (answering 85.1), and then by documents c and h (answering these questions for a third time). The final ordering would be a-e-g-b-f-c-h-i-j.

4. EVALUATION FRAMEWORK

The probability ranking principle (PRP) forms the bedrock of information retrieval research [22, 24]. We state the principle as follows:

> *If an IR system's response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized.*

The PRP is often interpreted as a nascent retrieval algorithm: Estimate the probability of relevance for each document and sort. We take a different view, interpreting the PRP as the starting point for the definition of an objective function to be optimized by the IR system.

Let q be a query. This query is implicit and fixed throughout our discussion. Let u be the information need occasioning a user to formulate q, and let d be a document that may or may not be relevant to u. Let R be a binary random variable representing relevance. To apply the PRP, we must estimate

$$P(R = 1|u, d).$$

It has become common in the summarization and question answering communities to refer to *information nuggets*, and to assess summaries on the basis of the nuggets they contain [15]. Following this lead, we model our user's information need as a set of nuggets $u \subseteq \mathcal{N}$, where $\mathcal{N} = \{n_1, ...n_m\}$ is the space of possible nuggets. Similarly, the information present in a document is modeled as a set of nuggets $d \subseteq \mathcal{N}$.

We interpret the notion of a nugget broadly, extending its usual meaning to encompass any binary property of a document. As is typical in summarization and question answering, a nugget may represent a fact or similar piece of information. In our QA example, a nugget might represent an answer to a question. However, a nugget may also represent other binary properties, such as topicality. We may also use a nugget to indicate that a page is part of particular Web site or is the home page of a particular organization. In our Web search example, a nugget might represent a specific fact about uninterruptible power supplies, a form for tracking packages, or the university's home page. Thus, nuggets may be used to model navigational needs, as well as informational needs.

Following the practice at TREC and other evaluation forums [18], we consider a document relevant if it contains any relevant information. In other words, a particular document is relevant if it contains at least one nugget that is also contained in the user's information need.

$$P(R = 1|u, d) = P(\exists n_i \text{ such that } n_i \in u \cap d) \qquad (1)$$

For a particular nugget n_i, $P(n_i \in u)$ denotes the probability that the user's information need contains n_i, and $P(n_i \in d)$ denotes the probability that the document contains n_i. These probabilities may be estimated for user information needs separate from documents, and for documents separate from user information needs. The only connection is the set of nuggets associated with each.

Traditionally, the probabilities are estimated to be 0 or 1 for particular examples of u and d; that is $P(n_i \in u) = 1$ indicates that n_i is known to satisfy u; $P(n_i \in u) = 0$ indicates that n_i is known not to satisfy u. Similarly, $P(n_i \in d) = 1$ indicates that n_i is found in d, and vice versa. This traditional model overstates the certainty with which either of these quantities may be assessed. Taking a more relaxed view better models the true situation. Human assessors are known to be inconsistent in their judgments [27]. Relevance judgments inferred from implicit user feedback may not always be accurate [1, 2, 16, 21]. If a classifier is applied to augment manual judgments, we may take advantage of a probability supplied by the classifier itself [10].

To formulate an objective function over needs and documents, we assume the independence of $n_i \in u$ and $n_{j \neq i} \in u$;

Document Title	85.1	85.2	85.3	85.4	85.5	85.6	Total
a. Carnival Re-Enters Norway Bidding		X		X			2
b. NORWEGIAN CRUISE LINE SAYS OUTLOOK IS GOOD		X					1
c. Carnival, Star Increase NCL Stake		X					1
d. Carnival, Star Solidify Control							0
e. HOUSTON CRUISE INDUSTRY GETS BOOST WITH...	X					X	2
f. TRAVELERS WIN IN CRUISE TUG-OF-WAR	X						1
g. ARMCHAIR QUARTERBACKS NEED... THIS CRUISE			X				1
h. EUROPE, CHRISTMAS ON SALE	X						1
i. TRAVEL DEALS AND DISCOUNTS							0
j. HAVE IT YOUR WAY ON THIS SHIP							0

Table 2: Top ten documents returned for the query "Norwegian Cruise Lines (NCL)". The questions answered by each document are indicated.

also of $n_i \in d$ and $n_{j\neq i} \in d$. Under this assumption, Equation 1 may be rewritten

$$P(R=1|u,d) \;=\; 1-\prod_{i=1}^{m}(1-P(n_i \in u)\cdot P(n_i \in d)). \quad (2)$$

Next, we turn our attention to the problem of estimating $P(n_i \in u)$ and $P(n_i \in d)$. With respect to the user, we are making the strong assumption that a user's interest in one nugget is independent of other nuggets. We discuss the ramifications of this assumption further in Section 4.2. We then consider ranked lists, where the relevance of each subsequent element is conditioned on the preceding ones.

4.1 Relevance Judgments

To estimate $P(n_i \in d)$ we adopt a simple model inspired by the manual judgments typical of TREC tasks. We assume that a human assessor reads d and reaches a binary decision regarding each nugget: Is the nugget contained in the document or not?

Let $J(d,i)=1$ if the assessor has judged that d contains nugget n_i, and $J(d,i)=0$ if not. A possible estimate for $P(n_i \in d)$ is then:

$$P(n_i \in d) = \begin{cases} \alpha & \text{if } J(d,i)=1, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The value α is a constant with $0 < \alpha \leq 1$, which reflects the possibility of assessor error. This definition assumes that positive judgments may be erroneous, but that negative judgments are always correct. This definition is a crude approximation of reality, but is still a step beyond the traditional assumption of perfect accuracy. More sophisticated estimates are possible, but are left for future work. If we assume Equation 3, then Equation 2 becomes:

$$P(R=1|u,d) = 1-\prod_{i=1}^{m}\left(1-P(n_i \in u)\alpha J(d,i)\right). \quad (4)$$

4.2 Ambiguity and Diversity

In arguing for evaluation methodologies that address query ambiguity, Spärck Jones et al. [25] emphasize that queries "are linguistically ambiguous, not just in the classic sense of words with multiple senses present in a dictionary, but also ambiguous across place names, person names, acronyms, etc." A number of researchers have investigated the relationship between query ambiguity and query difficulty [4, 14, 19]. Cronen-Townsend et al. [14] view ambiguity as a property that is inherent to a query "with respect to the collection being searched". They develop and validate a *clarity score*, based on the K-L divergence between a query language model and the collection language model. Their query language model is constructed from the top-ranked documents returned by the query. The clarity score is intended to reflect the coherence of these documents: Are they about a single topic or a mixture of topics?

In part, ambiguity may be associated with dependencies between the nuggets, which are ignored by Equation 2. While a user interested in the Norwegian Cruise Lines may find any fact regarding the company useful, the same cannot be said for our Web search example. A user interested in the parcel service will be less interested in nuggets related to power supplies.

An evaluation measure intending to reward diversity must take these dependencies into account when estimating $P(n_i \in u)$. In the case of our Web search example, we identified three possible interpretations of the query. Assuming our intuition regarding the user population is correct, nuggets related to the parcel service must be assigned substantially higher probabilities than nuggets related to the university. Problems can occur if, for example, the number of nuggets representing a more obscure interpretation is substantially larger than the number of nuggets representing a more popular interpretation. Under these circumstances, a document containing many nuggets related to the more obscure interpretation may receive an inappropriately high probability of relevance. We leave for future work the question of whether these dependencies represent a problem in practice

Beyond these dependencies, our notion of ambiguity includes other forms of underspecified queries. A user typing "UPS" may be tracking a package more often than locating the UPS Store in Redmond, Washington. Navigational interpretations of a query, for `www.ups.com` and `www.ups.edu`, may be accommodated by assigning high probabilities to nuggets associated with these home pages.

Assigning meaningful probabilities requires knowledge of user preferences, which might be determined explicitly or implicitly from user behavior and feedback. In the absence of this knowledge, we might assume that nuggets are independent and equally likely to be relevant. Assuming $P(n_i \in u) = \gamma$ for all i, where γ is a constant, and substituting into Equation 4, gives

$$P(R=1|u,d) = 1-\prod_{i=1}^{m}\left(1-\gamma\alpha J(d,i)\right). \quad (5)$$

4.3 Redundancy and Novelty

To this point, we have worked with a single document only. Applying Equation 5 to each document allows us to determine the one to be ranked first. For the second and subsequent documents, we must view relevance in the light of the documents that rank higher.

Assume we have a relevance estimate for the first $k-1$ documents in a ranked list (d_1, ..., d_{k-1}) and are now considering the relevance of d_k, the document at rank k. Let the random variable associated with relevance at each rank be R_1,...,R_k. Our goal is to estimate $P(R_k = 1|u, d_1, ..., d_k)$.

We assume that if a specific nugget appears in these first $k-1$ documents, then a repetition in d_k will provide no additional benefit — that redundancy is to be avoided in favor of novelty. Thus, the probability that the user is still interested in the nugget depends on the contents of these documents

$$P(n_i \in u|d_1, ..., d_{k-1}) = P(n_i \in u) \prod_{j=1}^{k-1} P(n_i \notin d_j).$$

We now define

$$r_{i,k-1} = \sum_{j=1}^{k-1} J(d_j, i),$$

the number of documents ranked up to position $k-1$ that have been judged to contain nugget n_i. For convenience, we define $r_{i,0} = 0$. Thus,

$$\prod_{j=1}^{k-1} P(n_i \notin d_j) \;=\; (1-\alpha)^{r_{i,k-1}},$$

and in the place of Equation 5 we have,

$$P(R_k = 1|u, d_1, ..., d_k) \qquad (6)$$
$$= 1 - \prod_{i=1}^{m} (1 - \gamma\alpha J(d_k, i)(1-\alpha)^{r_{i,k-1}}).$$

5. CUMULATIVE GAIN MEASURES

We now apply the results of the previous section to compute gain vectors for use with the Normalized Discounted Cumulative Gain measure [20]. Over the past few years, nDCG has established itself as the standard evaluation measure when graded relevance values are available [1,3,9]. Since graded relevance values arise naturally from the framework in the previous section, application to nDCG seems appropriate.

The first step in the computation of nDCG is the creation of a *gain vector*. While we could calculate a gain vector directly from Equation 6, it is convenient to simplify the equation further, as follows:

$$\begin{aligned} P(R_k = 1|u, d_1, ...) &= 1 - \prod_{i=1}^{m} (1 - \gamma\alpha J(d_k, i)(1-\alpha)^{r_{i,k-1}}) \\ &= 1 - 1 + \gamma\alpha J(d_k, 1)(1-\alpha)^{r_{1,k-1}} + ... \\ &\approx \gamma\alpha \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}} \end{aligned}$$

Dropping the constant $\gamma\alpha$, which has no impact on relative values, we define the kth element of the gain vector G as

$$\mathrm{G}[k] \;=\; \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}. \qquad (7)$$

For our QA example, if we set $\alpha = 1/2$, the document ordering listed in Table 2 would give

$$\mathrm{G} \;=\; \langle 2, \tfrac{1}{2}, \tfrac{1}{4}, 0, 2, \tfrac{1}{2}, 1, \tfrac{1}{4}, ... \rangle.$$

Note that, if we set $\alpha = 0$ and use a single nugget indicating topicality, the gain vector in Equation 7 represents standard binary relevance.

The second step in the computation of nDCG is the calculation of the *cumulative gain vector*

$$\mathrm{CG}[k] \;=\; \sum_{j=1}^{k} G[j].$$

For our QA example,

$$\mathrm{CG} \;=\; \langle 2, 2\tfrac{1}{2}, 2\tfrac{3}{4}, 2\tfrac{3}{4}, 4\tfrac{3}{4}, 5\tfrac{1}{4}, 6\tfrac{1}{4}, 6\tfrac{1}{2}, ... \rangle.$$

Before computing the cumulative gain vector, a discount may be applied at each rank to penalize documents lower in the ranking, reflecting the additional user effort required to reach them. A typical discount is $\log_2(1+k)$, although other discount functions are possible and may better reflect user effort [20]. We define *discounted cumulative gain* as

$$\mathrm{DCG}[k] \;=\; \sum_{j=1}^{k} G[j] / \left(\log_2(1+j)\right).$$

For our QA example,

$$\mathrm{DCG} \;=\; \langle 2, 2.315, 2.440, ... \rangle.$$

The final step normalizes the discounted cumulative gain vector against an "ideal" gain vector. However, CG and DCG may also be used directly as evaluation measures. In a study based on Web search results, Al-Maskari et al. [3] provide evidence that CG and DCG correlate better with user satisfaction than nDCG. Nonetheless, we include the normalization step in the results reported by this paper, leaving the exploration of the unnormalized measures for future work.

5.1 Computing Ideal Gain

The ideal ordering is the ordering that maximizes cumulative gain at all levels. In Section 3.2 we presented the intuition behind the ideal ordering for the documents in Table 2. For these documents, the ideal ordering is a-e-g-b-f-c-h-i-j. The associated ideal gain vector is

$$\mathrm{G}' \;=\; \langle 2, 2, 1, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}, ... \rangle.$$

The ideal cumulative gain vector is

$$\mathrm{CG}' \;=\; \langle 2, 4, 5, 5\tfrac{1}{2}, 6, 6\tfrac{1}{4}, 6\tfrac{1}{2}, ... \rangle,$$

and the ideal discounted cumulative gain vector is

$$\mathrm{DCG}' \;=\; \langle 2, 3.262, 3.762, ... \rangle.$$

In theory, the computation of the ideal gain vector is NP-complete. Given the definition of gain in Equation 7, minimal vertex covering may be reduced to computing an ideal gain vector. To transform vertex covering, we map each vertex into a document. Each edge corresponds to a nugget, with each nugget occurring in exactly two documents. Computing the ideal gain vector with $\alpha = 1$ provides the minimal vertex covering.

In practice, we have found it sufficient to compute (an approximation to) the ideal gain vector using a greedy approach [13, 31]. At each step, we select the document with the highest gain value, breaking ties arbitrarily. If we never encounter ties, this approach will compute the ideal gain vector. If ties occur, the gain vector may not be optimal. In the unusual event that a retrieval system outperforms this approximation, it would be credited with an ideal result.

5.2 α-nDCG

As final step in the computation of nDCG we normalize discounted cumulative gain by the ideal discounted cumulative gain vector

$$\text{nDCG}[k] \;=\; \frac{\text{DCG}[k]}{\text{DCG}'[k]}.$$

For our QA example,

$$\text{nDCG} \;=\; \langle 1, 0.710, 0.649, ... \rangle.$$

As is typical for IR evaluation measures, nDCG is computed over a set of queries by taking the arithmetic mean of the nDCG values for the individual queries. nDCG is typically reported at various retrieval depths, similar to precision and recall.

Our version of nDCG rewards novelty through the gain value defined in Equation 7. Otherwise it adheres to a standard definition of nDCG. To distinguish our version of nDCG, we refer to it as α-nDCG, emphasizing the role of the parameter α in computing the gain vector. When $\alpha = 0$, the α-nDCG measure corresponds to standard nDCG with the number of matching nuggets used as the graded relevance value.

6. EXPERIMENTS

The theory in the previous sections assumes that together the nuggets provide complete coverage of all information related to all interpretations of the query, potentially thousands or millions of nuggets. In practice, we may have to limit ourselves to a much smaller number, particularly if the topic creation and judging process is largely manual.

In this section we explore the creation of a test collection based on the preceding theory. We take as our starting point test collections from the TREC 2005 and 2006 question answering tracks. While these collections were built for an entirely different purpose, they do provide the basic structure of our desired collection. In this paper, we report only results using the TREC 2006 test collection. We used the TREC 2005 QA test collection for exploratory work; we do not report results using that collection, but they are consistent with the results from the 2006 collection.

The TREC 2006 collection comprises 75 question series, each based around a single target, similar to the example from 2005 given in Figure 1. The target from each series was treated as a query. The questions formed the basis for nuggets, with one or more nuggets associated with most questions. When creating nuggets, we omitted the last question in each series, which is a catch-all "OTHER" question asking for any other information the system could provide. Apart from the list questions, such as 85.1, a single nugget was associated with each of the remaining questions. For list questions, a nugget was associated with each possible answer to the question (unlike the example in Table 2). This procedure resulted in a query set with an average of 17.12 nuggets per query. The maximum number of nuggets per query is 56; the minimum is 7.

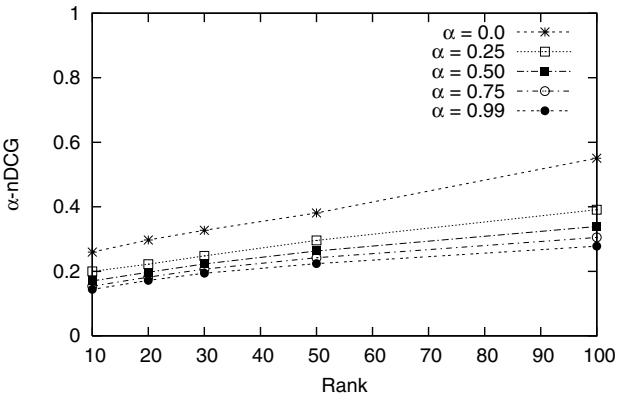


Figure 2: α-nDCG for reversed ideal gain.

Official TREC judgments are available for the question series [15]. We processed these official judgments into a total of 3,243 tuples, where each tuple specifies a query, a document and a nugget contained within it. Unfortunately, it is well known that official TREC judgments for QA tasks are incomplete [23] and our preliminary exploration of the 2005 collection confirmed this view. To complete the judging, we relied on a set of patterns distributed as part of the QA test collections. These patterns are designed to identify potential answers in unjudged documents, which then may be manually confirmed for QA evaluation purposes [23].

When run against the documents with official judgments, these patterns give a recall of 99% and a precision of 36%. While it is not reasonable to expect that an arbitrary document matching a pattern has a 36% chance of containing the corresponding nugget, it may be reasonable to assume that a document surfaced through a retrieval process does have this chance of containing the nugget. Consistent with the preceding theory, we might then set $\alpha = 0.36$.

In the following experiments, we base our judging on a combination of the official judgments and the patterns. We expect that any re-usable test collection focused on novelty and diversity will include an automatic judging component.

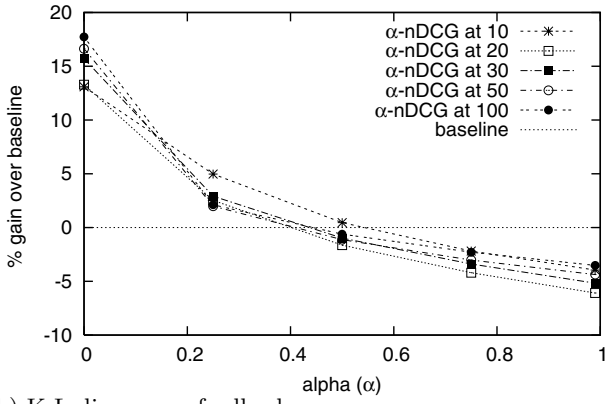
6.1 Reverse Ideal Gain

Under many traditional IR evaluation measures, such as MAP and bpref, an "ideal" retrieval result has all the relevant documents ranked ahead of all non-relevant documents. Unlike these measures, which consider only binary relevance, α-nDCG rewards both diversity and novelty (when $\alpha > 0$). Results that would score a perfect 1 under traditional evaluation measures may score considerably lower under α-nDCG.

To explore the extremes, we consider the effects of reordering relevant documents. In Section 5.1 we discussed the computation of an ideal gain vector. It is also possible to compute what we call a *reversed ideal gain vector*. This vector is constructed using a similar greedy algorithm, but attempts to minimize the α-nDCG score of the relevant documents — those that are judged to contain one or more nuggets.

Figure 2 plots α-nDCG values for the reversed ideal gain vector, for various values of α. In this graph, the ideal gain

a) Standard Okapi feedback:



b) K-L divergence feedback:

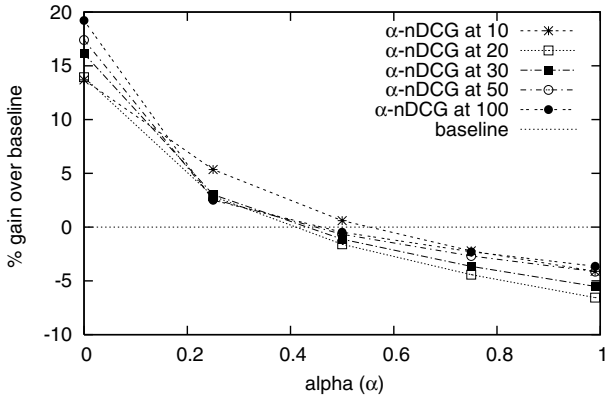


Figure 3: Impact of pseudo-relevance feedback on BM25 runs.

vector corresponds to a horizontal line at 1. As α increases, the gap between the two vectors also increases. From the standpoint of classic binary relevance, both the ideal gain vector and this reversed ideal gain vector are equivalent, with all relevant documents ranked first.

6.2 Pseudo-relevance feedback

The work of Chen and Karger [13] and Amati et al. [4] suggested to us that pseudo-relevance feedback may have a negative impact on novelty. Our collection provides an opportunity to test this hypothesis.

We executed the queries over the AQUAINT corpus, generating three runs. One run was generated by a version of the standard BM25 scoring formula. The other two runs represent variants of pseudo-relevance feedback: the first generated through standard Okapi-style feedback and the second generated through the K-L divergence feedback method described by Carpineto at al. [12].

Figure 3 shows the impact of pseudo-relevance feedback over the baseline BM25 run for both forms of feedback. When $\alpha = 0$ the α-nDCG measure is equivalent to standard nDCG with the number of matching nuggets used as the graded relevance value. Under this measure, which does not reward novelty, both variants of pseudo-relevance feedback produce typical performance improvements over the baseline BM25 run. At most ranks, these improvements are significant at the 95% level using a two-sided paired t-test.

As α increases, rewarding novelty, the situation changes. At $\alpha = 0.5$, there is no measured improvement over the baseline. At higher values, the curves for the pseudo-relevance feedback lie below the baseline, although this decrease is not significant.

7. CONCLUDING DISCUSSION

Our goal is to define a workable evaluation framework for information retrieval that accounts for novelty and diversity in a sound fashion. In our framework, documents are linked to relevance through informational nuggets, which represent properties of documents at one end and components of an information need at the other. The relationship between relevance and documents is captured by Equation 2, which lies at the core of our work. Its subsequent development and application to nDCG represents only one possible path. Other paths remain open.

Serious criticism could be applied to many links in our chain of assumptions. In particular, our assumption of independence between nuggets is invalid when the query has multiple unrelated interpretations. Moreover, we take a narrow view of relevance: that a document is relevant if and only if it contains a previously unreported nugget useful to the user. In some cases, repetition of information may also be useful, perhaps by increasing the user's confidence in its correctness. While the value of repetition is tacitly recognized in Equation 3, by giving credit to repeated nuggets, it may be beneficial to explicitly include it in the model.

The relatively small number nuggets used to operationalize our framework might be the subject of further concern, since many aspects will be unrepresented by these nuggets. This concern might be partially addressed by recognizing that the presence of a nugget in a document suggests the presence of information related to that nugget but not directly represented by it. For example, if the name of NCL's own private island appears in a document, answering question 85.3 in Figure 1, other information about the island may also appear. If the nugget appears again in later documents, different information may accompany it. Again, Equation 3 tacitly recognizes this possibility, but does not explicitly model it.

Despite these concerns, we believe we have made substantial progress towards our goal. Unusual features of our approach include recognition of judging error and the ability to incorporate a user model. While our experiments are limited in scope, they do demonstrate the feasibility of constructing evaluation measures and test collections under the framework.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–26, Seattle, August 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting Web search result preferences. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, Seattle, August 2006.

[3] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 773–774, 2007.

[4] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *26th European Conference on IR Research*, pages 127–137, Sunderland, UK, 2004.

[5] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *14th ACM International Conference on Information and Knowledge Management*, pages 736–743, 2005.

[6] B. Boyce. Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, 18(3):105–109, 1982.

[7] A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2):3–10, 2002.

[8] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

[9] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *22nd International Conference on Machine Learning*, pages 89–96, Bonn, Germany, 2005.

[10] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 63–70, 2007.

[11] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[12] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

[13] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 429–436, 2006.

[14] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[15] H. T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track. In *15th Text REtrieval Conference*, Gaithersburg, Maryland, 2006.

[16] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *16th International World Wide Web Conference*, 2007.

[17] W. Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2:73–78, 1964.

[18] D. K. Harman. The TREC test collections. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 2, pages 21–52. The MIT Press, 2005.

[19] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31:585–594, 2006.

[20] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[21] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, Salvador, Brazil, August 2005.

[22] J. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In W. B. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, chapter 1, pages 1–10. Kluwer Academic Publishers, 2003.

[23] J. Lin and B. Katz. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology*, 57(7):851–861, 2006.

[24] S. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977.

[25] K. Spärck Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: Implications for retrieval tests. *SIGIR Forum*, 41(2):8–17, 2007.

[26] K. Spärck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments - Part 1. *Information Processing & Management*, 36(6):779–808, 2000.

[27] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, 1998.

[28] E. M. Voorhees and H. T. Dang. Overview of the TREC 2005 question answering track. In *14th Text REtrieval Conference*, Gaithersburg, Maryland, 2005.

[29] Y. Xu and Hainan Yin. Novelty and topicality in interactive information retrieval. *Journal of the American Society for Information Science and Technology*, 59(2):201–215, 2008.

[30] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–278, 2007.

[31] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–17, 2003.

[32] C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing & Management*, 42:31–55, 2006.