Query Polyrepresentation for Ranking Retrieval Systems without Relevance Judgments

Miles Efron
Graduate School of Information and Library Science
University of Illinois, 501 E. Daniel St.
Champaign, IL, 61820
mefron@gmail.com

School of Information
University of Texas
1 University Station D7000, Austin, TX 78712
megan@ischool.utexas.edu

June 16, 2009

Abstract

Ranking information retrieval (IR) systems with respect to their effectiveness is a crucial operation during IR evaluation, as well as during data fusion. This paper offers a novel method of approaching the system ranking problem, based on the widely studied idea of polyrepresentation. The principle of polyrepresentation suggests that a single information need can be represented by many query articulations–what we call *query aspects*. By skimming the top k (where k is small) documents retrieved by a single system for multiple query aspects, we collect a set of documents that are likely to be relevant to a given test topic. Labeling these skimmed documents as putatively relevant lets us build pseudo-relevance judgments without undue human intervention. We report experiments where using these pseudo-relevance judgments delivers a rank ordering of IR systems that correlates highly with rankings based on human relevance judgments.

1 Introduction

Ranking information retrieval (IR) systems is a core problem in the field of information retrieval. Given a set of N IR systems, system ranking refers to the problem of listing these systems in decreasing order of their effectiveness with respect to a body of test queries. System ranking plays a role in several areas of IR. In particular, system ranking forms a crucial part of IR evaluation. Likewise, IR based on multiple searches (so-called metasearch) often uses ranking information to weight individual systems during data fusion.

In this paper we propose a novel way to approach the ranking problem. Our approach is based on the idea of query *polyrepresentation*, the notion that test queries can be represented in a variety of ways. In this paper we rely on multiple query representations–what we call *query aspects*–to build pseudo-relevance judgments in support of system ranking.

This problem is important because the bottleneck in established system ranking methods (e.g. the Cranfield method) lies in obtaining relevance judgments–a list of which documents are relevant to each query. Relevance judgments are often called *qrels*. While documents and queries are relatively easy to gather, creating relevance judgments requires significant effort and resources. Thus recent years have seen increased interest in methods for evaluating IR systems without human-generated relevance judgments.

This paper presents a novel method for creating relevance judgments without human assessments. The idea is very simple. Given a query q and a corpus of documents D we want to create a set of qrels for q and D. To do this, we propose the simple idea of query aspects.

Query aspects are variations–re-articulations, generalizations, specializations, etc.–of a query. Though q consists of a particular set of terms, it is not unreasonable to think that the same information need $\mathcal{I}$ that generated q could also generate another query focusing on another aspect of the topic at hand. We call this variation a. The query aspect a might be an elaboration, rephrasing, specification, or generalization of q. We call q and a "aspects" of the same information need $\mathcal{I}$.

To build a set of qrels for q on a corpus of documents D we generate a small number m of query aspects. Using a single IR system, we run each of the query aspects as a query against the corpus, resulting in a collection of m retrieval results related to q. The union of the top k documents retrieved for each aspect constitutes a list of pseudo-qrels for q. From this point, evaluation proceeds in the usual fashion, merely substituting pseudo-qrels for human-judged qrels. Our core idea is that the set of query aspects articulates different facets of the underlying information need. Running each aspect as a query against D allows us to cast a wide net, collecting what is hopefully a variety of relevant documents.

Of course casting such a wide net will surely bring non-relevant documents into the derived pseudo-qrels. But we argue that the presence of these spurious "relevant" documents will not be detrimental for system ranking because they will be artifacts of individual query aspects, and thus are unlikely to be retrieved by the actual test query q. The empirical results presented in this paper suggest that this intuition is not misguided.

The proposed approach is appealing in two main senses. First, unlike most judgment-free evaluation methods, our technique requires only one system to create pseudo-relevance judgments; the method does not rely on a system pool. Thus a researcher could create a test collection to experiment with different parameterizations of a single "in-house" system. Second, our approach does not entirely eliminate human attention from performance evaluation. Rather we rely on judicious and easy guidance by people.

2 Related Work

To enable Cranfield-style evaluation, researchers rely on test collections that consist of a corpus of documents, a set of queries, and a set of relevance judgments–statements of which documents are relevant to each query, also called *qrels* [Cleverdon and Mills, 1963]. Under the auspices of NIST's Text REtrieval Conferences (TREC)[1] qrels are usually obtained by the method of *pooling*. Pooling involves collecting the top k documents retrieved for each query by many systems and presenting the union of this document set to human assessors for manual relevance judgment. Pooling has shown great robustness with respect to ranking systems [Voorhees, 1998, Voorhees, 2000, Buckley and Voorhees, 2000]. However, creating test collections by pooling requires a heavy investment of resources, a fact that has spawned new areas of IR research.

Using Cranfield-style evaluation when we have few if any relevance judgments is an active area in the IR literature. Recent work has seen excellent advances in methods of creating test collections with very few relevance judgments or in the absence of any sort of pooling [Carterette and Allan, 2005, Carterette et al., 2006, Sanderson and Joho, 2004]. Additionally, scholars have proposed novel performance metrics that account for common realities of IR evaluation such as changing document collections, incomplete relevance judgments, etc. [Buckley and Voorhees, 2004, Sakai, 2007].

A growing body of work concerns system evaluation when we have *no* relevance judgments. Soboroff *et al.* hypothesized that statistical sampling from documents in a system pool could generate a set of "pseudo-qrels" [Soboroff et al., 2001]. Subsequent research has advanced the state of the art with respect to judgment-free evaluation by considering the relationship among the documents retrieved by a pool of IR systems [Wu and Crestani, 2003, Spoerri, 2007].

[1] `http://trec.nist.gov`

While evaluating performance without human relevance judgments has obvious appeal, judgment-free performance prediction also has ramifications for fusion-based IR systems, where a final document ranking is obtained by consulting the rankings given from several systems (cf. [Lee, 1997]).

Our work is concerned with constructing pseudo-qrels without recourse to human relevance judgments. Furthermore, we make no assumption that a pool of results is available. Instead of relevance judgments or multiple system comparisons, the following discussion proposes a method for building pseudo-qrels by eliciting multiple representations of each test topic. Our motivation is that a relatively small amount of effort can yield many representations of a single information need and that we can use these various representations to gain a good idea of which documents treat the test topic.

The core idea we propose is the notion of a "query aspect," a textual instantiation of an information need. Building on query aspects, we also define "aspect sets." An aspect set is a collection of query aspects. These aspects are related in various ways (e.g. elaboration, generalization, specification, etc.) to a searcher's information need. This idea is heavily indebted to the principal of polyrepresentation as outlined in [Larsen et al., 2006, Ingwersen and Järvelin, 2005] and [Skov et al., 2008]. In IR systems, objects such as documents and queries are often represented in many parallel forms. A great deal of research concerns how to marry these "polyrepresented" objects to inform IR [Belkin et al., 1993, Belkin et al., 1995, Kelly and Fu, 2007].

Like most of these studies, our work is predicated on the notion that an information need can be represented in many ways and that these divergent representations bespeak differing facets of the topic. Unlike prior work, however, we are concerned with using multiple query representations for IR system ranking, as opposed to document ranking. Additionally, we make no effort to unify our polyrepresented topics; in fact a diversity of vocabulary and emphasis is key in our approach.

3 Approach and Motivation

A given information need can be expressed in many ways. For example, most topics in TREC test collections have three representations: title, description, and narrative. Each of these fields treats the topic at hand, but each uses distinct phrasing. For example, consider TREC topic 402:

```
Title:  behavioral genetics


Description:  What is happening in the field of behavioral genetics,
the study of the relative influence of genetic and environmental
factors on an individual's behavior or personality?


Narrative:  Documents describing genetic or environmental factors
relating to understanding and preventing substance abuse and addictions
are relevant.  Documents pertaining to attention deficit disorders
tied in with genetics are also relevant, as are genetic disorders
affecting hearing or muscles.  The genome project is relevant when
tied in with behavior disorders (i.e., mood disorders, Alzheimer's
disease).
```

A casual query based on this topic might be *behavioral genetics.* But we could focus our search by querying *behavioral genetics research.* And perhaps after a bit of reading, we learn that twin studies are common in the information relevant to this topic. Then we might search for *twin studies.*

We assume that topic 402 speaks to a real but intangible information need $\mathcal{I}$. Each of the searches described in the preceding paragraph articulates something related to that need. But each does so with different emphasis. Each of the preceding searches are what we call query aspects related to $\mathcal{I}$.

3.1 Query Aspects and the Principle of Polyrepresentation

The principle of polyrepresentation is based on Ingwersen's hypothesis that documents are more likely to be relevant to a user if they evince query-document similarity on a variety of cognitively unique features [Ingwersen, 1996]. Particularly important to our discussion is the idea of *functionally* different representations of information. Functionally different information representations are distinct articulations created by a particular person. For example, three functional representations of a document are its title, abstract, and citations, all of which are created by its author [Skov et al., 2008].

At first glance the principle of polyrepresentation does not seem to bear on the matter of IR system ranking. Instead, polyrepresentation has most often been invoked to inform the process of data fusion for document ranking [Belkin et al., 1995, Belkin et al., 1993, Kelly et al., 2005].

However, the principle of polyrepresentation is based on the idea that information can be expressed in multiple, diverse means. Thus a document's topic may be represented by the features listed above. Likewise an information need can be represented by, say, a keyword query, a Boolean query, and a relevant document.

Our work begins with the idea that by collecting retrievals from multiple representations of a topic we can obtain a set of putatively relevant documents to be used as "pseudo-qrels." Each of these representations highlights a facet of the user's underlying information need and thus we refer to them as *query aspects.*

3.2 Query Aspects for Ranking IR Systems

We offer this definition of a query aspect:

> **Definition 1: Query aspect.** A linguistic articulation of an information need.

In the context of Definition 1, standard *ad hoc* IR, which uses a single query to retrieve documents, makes use of a single query aspect. However, we assume that the underlying information need that generated such an aspect could be pursued by diverse query aspects. These aspects might be synonymous with the original query. Alternatively, they might emphasize specific facets of the information need, or the broader context of the user's interests.

In addition to query aspects we introduce the notion of an aspect set:

> **Definition 2: Aspect set.** A collection of terms found in one or more query aspects.

Collectively, an aspect set consists of terms that speak broadly to the user's information need. In the example of query aspects for TREC topic 402 given above, our aspect set would be {*behavioral genetics behavioral genetics research twin studies*}. Note that an aspect set is not a proper set in the sense that its members (terms) may appear multiple times.

To enable system ranking we propose using the relationship between information needs and query aspects to build a set of pseudo-qrels for the queries of a test collection. The intuition behind our approach is simple. Given a test query q we can generate a list of pseudo-qrels for q on the collection D by skimming the top k documents returned for searches on several related query aspects. Our hope is that each query aspect's top-ranked documents will have a high likelihood of being relevant to $\mathcal{I}$. But at the same time, by dint of using multiple query representations, we try to cast a wide net over the topic's

Table 1: Notation used throughout this paper

D	a collection of documents
d	a particular document
Q	a collection of p statements of information need
q	a particular statement of information need
S	an IR system
$S(q, D)$	the ranked list of documents retrieved from D by S for q
$S^k(q, D)$	The top k documents retrieved by S for q

domain, building pseudo-qrels based on varied aspects of the underlying need.

Table 1 gives the notation that informs the following explanation of our method.

We assume that we have access to a corpus of documents D and a set of test queries Q. Our goal is to create useful proxies of relevance judgments for each query q in Q. To do this, we analyze q and create n distinct query aspects. We collect these aspects into a single aspect set A_1. We then repeat this process m times, resulting in a collection of m aspect sets. Exactly how to solicit these aspects is an open question. But in this paper, we have asked research subjects to analyze each test topic and create suitable query aspects (this is described in detail below).

Having obtained m aspect sets for q, $A_1 \ldots A_m$, we submit each set as a query to a "seed" IR system" (again, we discuss this in detail below). This results in m document rankings $S_1(A_1, D) \ldots S_m(A_m, D)$ each of which contains the top c, say 1000, documents obtained from its respective run.

Our hope is that the most highly ranked documents in each result set are likely to be relevant to the information need that created q. Thus we create our pseudo-qrels by skimming off the top k documents from each ranking $S_1^k(A_1, D) \ldots S_m^k(A_m, D)$, calling the union of these documents the putatively relevant documents for q.

Having built our pseudo-qrels, system ranking now proceeds as usual. Using software such as NIST's `trec_eval` we may calculate effectiveness measures such as mean average precision, P@10, bpref, etc. If our approach of skimming results obtained from multiple query representations has worked, effectiveness measures computed from the pseudo-qrels should correlate highly with measures computed from qrels obtained by an established method such as pooling.

In the experiments described below we rank systems submitted to TREC using mean average precision calculated from both NIST's official relevance judgments (we abbreviate this as MAP) and from our pseudo-relevance judgments (abbreviated aMAP, for "aspect MAP").

4 Experimental Evaluation

We tested our approach on results obtained from three of the TREC conferences, TREC-3 [Harmon, 1996], TREC-7 [Harmon and Voorhees, 1996b], and TREC-8 [Harmon and Voorhees, 1996a]. To make our results comparable to those reported in [Wu and Crestani, 2003] and [Spoerri, 2007] we limited our analysis to a subset of the systems that participated in each year. Specifically, we use only so-called *automatic* runs, as opposed to *manual* runs. Our analysis does, however, retain runs based on any of the topic fields: *title*, *description*, and *narrative*. Table 2 lists basic statistics about the test collections used for our experiments[2].

4.1 Creating Query Aspects

For each TREC topic listed in Table 2 (a total of 150 topics) we obtained four query aspect sets (i.e. 600 aspect sets). Query aspects were generated

[2] In the table *-CR* means that the collection did not use *The Congressional Record* documents.

Table 2: Datasets used for experimentation

Corpus	Disks	# Docs	Topics	#Systems
TREC-3	1-3	1,078,169	151-200	29
TREC-7	4-5 (-CR)	527,094	351-400	86
TREC-8	4-5 (-CR)	527,094	401-450	115

by a sample of participants recruited from our department. Six paid subjects were each given a set of 50 TREC topics. An additional two subjects were given all 150 topics. Based on each assigned topic, each participant created a corresponding aspect set.

Study participants were shown only each topic's description field. The description field was chosen because it is more detailed than the title field, but we felt that the narrative fields were too prescriptive and likely to constrain aspect construction.

To create each aspect the participant consulted the corresponding TREC topic description. Then, using a variety of online resources (e.g. Google, Yahoo, Wikipedia–any resource the participant chose), he or she conducted a series of searches related to the topic. During the course of the search, any queries entered into a search service were added to the list of aspects for that user's search on that topic. Finally, each participant's aspects for a particular topic were merged into a single, long query–the aspect set for that user on that topic. In future work we plan to test more sophisticated methods of combining evidence from multiple aspects created by a particular user.

As an example, consider the following aspects that one subject submitted for topic 402:

- behavioral genetics
- news in behavioral genetics

- criminal behavior in behavioral genetics
- genetics and environmental influence on criminality
- biological theories behavioral genetics.

This led to the participant's aspect set for topic 402 to be {*behavioral genetics news in behavioral genetics criminal behavior behavioral genetics genetics and environmental influence on criminality biological theories behavioral genetics*}. All told, study participants created a total of four aspect sets for each of the 150 topics listed in Table 2.

During aspect creation, participants' research process was pursued quickly and casually. The aspect creation time was clocked for the two volunteers who covered all 150 topics. For these subjects, aspect creation for 150 topics took approximately seven hours per person, about three minutes per topic. Thus each TREC data set took about 2.5 hours per person. The median length of a manually generated query aspect was 13 words (mean 14.75), and the standard deviation was 7.26 words.

To create a set of pseudo-qrels for a given topic, we ran each query aspect set against its corresponding document collection (see Table 2). These runs used the KL Divergence approach (cf. [Zhai and Lafferty, 2006]) as their retrieval model. In the vocabulary we introduced above, KL constitutes our "seed system." Our seed system used no stemming and no stoplist. The system smoothed its language models by Bayesian updating using a Dirichlet hyperparameter of $\mu = 2000$. The system (and all systems referred to below) were implemented using the lemur toolkit[3].

Each run generated a ranking of at most 1000 documents against a particular aspect. The pseudo-qrels were obtained by taking the union of the top k

[3] `http://lemurproject.org`

documents retrieved by the seed system for all four aspect sets. The variable k is a tunable parameter, and unless otherwise stated we set $k = 50$.

4.2 Baseline Correlation of aMAP with MAP

The simplest question that bears on our analysis is, to what extent does performance evaluation based on judgment-free aspect qrels correlate with evaluation by traditional, human-judged qrels? We measure the correlation between rankings using two rank correlation metrics: Kendall's τ (abbreviated τ) and the Spearman rank correlation (abbreviated src) [Kendall, 1990].

To ground our analysis we compare our rank correlations with those obtained by the method of "reference counts" detailed by Wu and Crestani in [Wu and Crestani, 2003] (we abbreviate this approach RC). The RC method shows marked improvement over earlier judgment-free evaluation (e.g. [Soboroff et al., 2001]). Ranking systems by RC is straightforward. Given N systems and a system S whose performance we wish to estimate, RC derives a statistic for S obtained by summing 1000 minus the rank of each document that appears for each query among the remaining $N - 1$ systems. The intuition here is that systems that return documents that are highly ranked among other systems are likely to be high performers.

Comparing RC and aMAP is perhaps complicated because RC is completely automatic, requiring no human intervention. While our approach requires no relevance judgments, it does depend on people creating multiple query aspects. On the other hand, RC assumes that we have multiple systems on hand for evaluation, while our approach requires only a single seed system to generate aspect qrels.

Table 3 shows τ and Spearman rank correlations between the official TREC MAP rankings and system rankings obtained by the RC method and by aMAP.

Table 3: Rank correlations with MAP: comparison of the reference counting technique and aMAP as described in Section 3

	RC		aMAP		% Improved	
Data	tau	src.	tau	src	tau	src.
TREC-3	0.374	0.556	0.875	0.972	133.96	74.20
TREC-7	0.639	0.843	0.804	0.943	25.82	11.86
TREC-8	0.603	0.784	0.775	0.918	28.52	17.09

All correlations are significant above the 99% level. Additionally aMAP appears more strongly correlated with MAP than the ranking given by RC. The table's two rightmost columns show the percent improvement in rank correlation using aMAP, our method based on judgment-free aspect qrels, over RC; for all three data sets, the improvement is high.

To contextualize these results, we refer to Voorhees' work in [Voorhees, 2001]. Voorhees suggests that human agreement regarding relevance typically results in τ correlations of approximately 0.9. As a heuristic, she posits that we consider systems scoring at least τ=0.8 with a benchmark measure not "noticeably" different than that benchmark. In the cases of TREC-3 and 8 aMAP surpasses τ correlation of 0.8 with respect to MAP, while the reference counting method falls below this threshold in all cases.

Figure 1 goes here.

Figure 1: System rankings using aMap. aMap scores plotted in the order generated by ranking systems with respect to MAP calculated from NIST's qrels

Figure 1 plots aMap for the systems participating in each TREC listed in Table 2. The systems are listed in decreasing order of their rank using MAP computed from NIST's qrels. The panels of Figure 1 suggest that aMap is very successful at identifying poorly performing systems. This is very common in the literature of judgment-free evaluation. The correlation between aMAP and

MAP remains strong even for high-performing systems in TREC-3 and TREC-7. But towards the left side of the rightmost panel of the figure the correlation on the TREC-8 data weakens. Nonetheless, throughout the TREC-8 ranking, a solid line of points is visible in the center of the data cloud, suggesting that many high-ranked systems are also ranked highly by aMAP.

4.3 Sufficient Query Aspect Representations

This section concerns the question: how many query aspect sets are needed to rank IR systems well? In the experiments described above, for each topic we had four aspect sets–that is, we have four unique retrievals per topic. To examine how the number of aspect sets effects system rankings we generated an additional four aspect sets for the TREC-8 data. These additional aspect sets were created by a second set of four volunteers.

Figure 2 goes here

Figure 2: Effect on MAP-aMAP correlation of the number of aspect sets used to create pseudo-qrels.

Figure 4.3 plots the rank correlations (τ and Spearman) obtained by generating pseudo-qrels based on one aspect set, two groups, etc. The data point for the ith number of aspect sets was obtained by sampling $N = 10$ aspect sets of size i (obviously, for runs using seven or eight aspect sets, $N < 10$. For example there are 70 ways to choose four aspect sets from our eight. The point at 4 on Figure 2's x-axis is the average correlation among a random sample of 10 among these 70.

The figure suggests that adding aspect sets to the process of generating pseudo-qrels is beneficial. While the strongest gains appear as we increase the number of aspect sets from one to approximately three, Figure 2 shows non-trivial gains in correlation all the way to eight groups.

In practice, the number of available aspect sets is likely to be limited by resources (e.g. the amount of time available to create them). However, the data reported here suggest that aMAP sees stronger correlation with MAP as the number of aspect sets grows. Without additional testing, it is not clear if a still higher number of aspects would become a liability.

4.4 Effect of Aspect Pool Depth on Correlation

Figure 3 goes here

Figure 3: Effect on MAP-aMAP Correlation of the Number of Documents Skimmed from each aspect set Run (Aspect Pool Depth)

As described in Section 3, when creating pseudo-qrels, we must choose the value for a parameter k: the number of documents "skimmed" from each aspect set's retrieval. This is very similar to the question of how many documents should be treated as putatively relevant during blind relevance feedback. If we choose a value for k that is too small, we suspect that the resulting pseudo-qrels will include only a few relevant documents, and will therefore yield a weak statistic for system ranking. On the other hand, setting k too high will introduce documents to the analysis that have very low likelihood of relevance.

Figure 3 shows the effect of changing k, the number of documents considered "relevant" from each aspect set's retrieval. The Figure's x-axis is k. The y-axis gives rank correlation between MAP and aMAP, with τ correlation in black and Spearman in gray.

The figure suggests that for all three data sets it is possible to set k too high. In all cases, $k > 300$ decreases the correlation. Likewise, in most cases k can be set too low. All of the corpora show a region of optimality near $k = 100$. This result jibes with intuition. If the first documents retrieved by our seed system are the most likely to be relevant, setting k to a large number risks labeling low-scoring documents as putatively relevant (in our pseudo-qrels). If we fail to include a sufficiently high number of documents in our pseudo-qrels, however, we will fail to include genuinely relevant documents. This will impact our system ranking by penalizing systems that rank "non-obvious" relevant documents highly.

4.5 Sensitivity of aMAP to Input Quality

The results discussed in Section 4.2 raise an important question. Ranking systems by aMAP requires use of a "seed system." In this case our seed was a KL divergence retrieval system with no stemming or stopwords used. To what extent does our particular seed system bear on the success of aMAP? In other words, we might ask whether the high correlations we observed were due merely to a very good retrieval system. Would our results be different given a poor seed system?

Table 4: aMAP-MAP correlations derived from two seed systems.

	Cor. KL		Cor. Okapi		MAP (topic desc.)		% Decline	
Data	tau	src.	tau	src.	KL	Okapi	MAP	tau
TREC-3	0.875	0.972	0.856	0.965	0.164	0.133	23.3	2.2
TREC-7	0.804	0.943	0.743	0.913	0.208	0.172	20.9	8.0
TREC-8	0.775	0.918	0.656	0.825	0.216	0.155	38.7	18.1

Table 4 compares results obtained using two seed systems. The columns labeled *KL* refer to our seed system described above. In contrast, we have computed τ and Spearman rank correlations with MAP obtained using an Okapi system (using the lemur default parameters). In addition to the correlations, Table 4 gives the mean average precision obtained by applying TREC Description topic fields as queries using each query. These MAP scores are intended to demonstrate the difference in retrieval effectiveness between the KL and Okapi systems.

The columns of Table 4 under the MAP heading show that the KL system outperformed the Okapi system for all three corpora (paired Wilcoxon tests yielded $p < 0.01$ in all cases). We can see that on TREC 3, 7, and 8, Okapi sees %23.3, %20.9, and %38.7 decreases in MAP versus KL. On the other hand, aMAP correlations only decline %2.2, %8.0, and %12.0, respectively. Thus it seems that the quality of the seed system does effect aMAP's correlation with MAP. However, the degree of this relationship does not appear very strong; in fact, in the case of TREC-3 aMAP calculated on the Okapi system is almost as highly correlated as the KL-derived aMAPs.

A second important question hinges on the matter of the specific query aspects that our research subjects generated for this study. Perhaps these are very good query aspects. Maybe we were simply lucky in the sets our participants created, and if we were to sample aspect sets again our results would be less compelling.

Figure 4 goes here

Figure 4: Sensitivity of aMAP to the quality of the Query Aspects (TREC-7)

Figure 4 speaks to this question. The plot shows data based on the TREC-7 collection. To interpret the graph, recall that for TREC-7 we have four aspect sets, a_1, a_2, a_3, a_4 for each of the collection's 50 topics. The figure was generated by running each of these aspects as a query. At each stage (reading from left to right) we added Gaussian noise to the resulting query-document scores and reordered the results. The noise was distributed $\mathcal{N}(0, x\sigma)$, where σ is the sample standard deviation of the query-document similarity scores for a given topic. The x-axis shows the factor x. Thus reading left to right, the scores become noisier and noisier.

The four lines at the bottom of the figure shows P@10 (calculated using NIST's qrels) using each aspect set as a query. As the noise increases we see that the aspects become "worse" queries; i.e. they score lower P@10[4].

The two lines at the top of the figure we see the τ and Spearman correlations between aMAP and MAP calculated from increasingly noisy aspects.

Figure 4 shows that as the aspects decline in quality the strength of correlation between aMAP and MAP gets weaker. However, this dynamic is relatively mild. Moving from no noise to very noisy data degrades P@10 performance for a_1 by %155.94. However, the corresponding shift for aMAP's correlation with MAP is %8.2 (τ) and %3.9 (src.). Thus, aspect quality appears to be only a mild factor in aMAP-MAP correlation.

5 Discussion

In Section 4 we found that aMAP (mean average precision calculated from pseudo-qrels derived from query aspects) correlates highly with mean average precision generated by human-created relevance judgments. However, we do not mean to suggest that using query aspects for system ranking is superior or preferable to traditional evaluation. Instead, we offer the method of generating aspect-based pseudo-qrels for situations where complete test collections are unavailable. Creating query aspects is easy and fast, making them appealing for situations where "in-house" retrieval systems are to be compared, for instance.

Additionally, we propose that documents retrieved using multiple query aspects could entail useful starting points for manual relevance assessments (much as pooled results are used).

An interesting question that arises from our discussion is the status of relevance in the method of system ranking that we propose in this paper. Inter-rater

[4] We report P@10 here instead of MAP for scaling purposes on the plot.

reliability with respect to relevance has sparked debate in the IR community [Harman, 1998, Harter, 1996]. The proposed methodology obviates the need for inherently subjective judgments of relevance. Instead, however, our approach raises the question, is a given set of query aspects sufficient to gauge system quality?

Likewise, researchers have argued that binary relevance judgments are unrealistic. How our approach relates to this criticism is not immediately obvious. The pseudo-qrels derived by query aspects, when used to calculate aMAP play a role analogous to traditional relevance judgments. However, to obtain these pseudo-qrels, no subjective judging takes place. Our pseudo-qrels, then, seem qualitatively different than standard qrels. However, their specific status in the system ranking problem poses a question we hope to pursue in further research.

6 Conclusion

Ranking IR systems with respect to their effectiveness is a crucial problem in retrieval. System ranking forms the core of experimental IR research, and it has implications for metasearch problems, as well. In this paper we presented a novel method of ranking IR systems without recourse to human-generated relevance judgments. Instead of judging document relevance, we propose enlisting human effort to create query aspects, alternative articulations of the information needs responsible for each test query. The method we describe requires only a small amount of human effort, and the resulting pseudo-qrels lead to system rankings that correlate strongly with those obtained using official TREC rankings.

The core of our approach is the idea of *query aspects*. Query aspects are diverse statements of a single information need. Following the idea of polyrepresentation advanced in cognitive theories of IR, we argue that documents relevant to an information need are likely to manifest evidence of that need in a variety of forms. By collecting query aspects into *aspect sets* and submitting these sets to a seed IR system, we can obtain a catholic group of documents relevant to the user's underlying information need. Treating these documents as "pseudo-qrels"allows us to undertake analysis similar to traditional Cranfield IR experimentation without recourse to explicit human relevance judgments.

More specifically our principal argument is that gathering the top k documents retrieved from a small number m (in this paper $m = 4$ in most cases) aspect sets casts a wide net, pulling together diverse relevant documents without including a harmful number of non-relevant documents in the set of pseudo-qrels. Our chief contribution lies in transferring the idea of polyrepresentation from its traditional place in ranking documents to the matter of ranking IR systems.

References

[Belkin et al., 1993] Belkin, N. J., Cool, C., Croft, W. B., and Callan, J. P. (1993). The effect multiple query representations on information retrieval system performance. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 339–346, New York, NY, USA. ACM.

[Belkin et al., 1995] Belkin, N. J., Kantor, P. B., Fox, E. A., and Shaw, E. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448.

[Buckley and Voorhees, 2000] Buckley, C. and Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd Annual iInternational ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, New York, NY, USA. ACM.

[Buckley and Voorhees, 2004] Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, New York, NY. ACM Press.

[Carterette and Allan, 2005] Carterette, B. and Allan, J. (2005). Incremental test collections. In *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 680–687, New York, NY, USA. ACM.

[Carterette et al., 2006] Carterette, B., Allan, J., and Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, Seattle, WA. ACM Press.

[Cleverdon and Mills, 1963] Cleverdon, C. W. and Mills, J. (1963). The testing of index language devices. *ASLIB Proceedings*, 15(4):106–130.

[Harman, 1998] Harman, D. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323.

[Harmon, 1996] Harmon, D. K. (1996). *Overview of the Third Text Retrieval Conference (TREC-3)*. DIANE Publishing Company.

[Harmon and Voorhees, 1996a] Harmon, D. K. and Voorhees, E. M. (1996a). *Overview of the Eighth Text Retrieval Conference (TREC-8)*. DIANE Publishing Company.

[Harmon and Voorhees, 1996b] Harmon, D. K. and Voorhees, E. M. (1996b). *Overview of the Seventh Text Retrieval Conference (TREC-7).* DIANE Publishing Company.

[Harter, 1996] Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49.

[Ingwersen, 1996] Ingwersen, P. (1996). Cognitive perspectives on information retrieval interaction – elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50.

[Ingwersen and Järvelin, 2005] Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series).* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Kelly et al., 2005] Kelly, D., Dollu, V. D., and Fu, X. (2005). The loquacious user: a document-independent source of terms for query expansion. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 457–464, New York, NY, USA. ACM.

[Kelly and Fu, 2007] Kelly, D. and Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. *Information Processing and Management*, 43(1):30–46.

[Kendall, 1990] Kendall, M. (1990). *Rank Correlation Methods, Third Edition.* Griffin.

[Larsen et al., 2006] Larsen, B., Ingwersen, P., and Kekäläinen, J. (2006). The polyrepresentation continuum in IR. In *IIiX: Proceedings of the 1st Interna-*

tional Conference on Information Interaction in Context, pages 88–96, New York, NY, USA. ACM Press.

[Lee, 1997] Lee, J. H. (1997). Analyses of multiple evidence combination. *SIGIR Forum*, 31(SI):267–276.

[Sakai, 2007] Sakai, T. (2007). Alternatives to bpref. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 71–78, New York, NY, USA. ACM.

[Sanderson and Joho, 2004] Sanderson, M. and Joho, H. (2004). Forming test collections with no system pooling. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, New York, NY, USA. ACM.

[Skov et al., 2008] Skov, M., Larsen, B., and Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing and Management*, 44(5):1673–1683.

[Soboroff et al., 2001] Soboroff, I., Nicholas, C., and Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 66–73, New Orleans, Louisiana, United States. ACM. 383961.

[Spoerri, 2007] Spoerri, A. (2007). Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing and Management*, 43(4):1059–1070.

[Voorhees, 1998] Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings* *of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, New York, NY, USA. ACM.

[Voorhees, 2000] Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716.

[Voorhees, 2001] Voorhees, E. M. (2001). Evaluation by highly relevant documents. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, New York, NY, USA. ACM.

[Wu and Crestani, 2003] Wu, S. and Crestani, F. (2003). Methods for ranking information retrieval systems without relevance judgments. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 811–816, New York, NY, USA. ACM.

[Zhai and Lafferty, 2006] Zhai, C. and Lafferty, J. (2006). A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55.