Using Preference Judgments for Novel Document Retrieval

Praveen Chandar and Ben Carterette
{pcr,carteret}@udel.edu
Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA 19716

ABSTRACT

There has been considerable interest in incorporating diversity in search results to account for redundancy and the space of possible user needs. Most work on this problem is based on *subtopics*: diversity rankers score documents against a set of hypothesized subtopics, and diversity rankings are evaluated by assigning a value to each ranked document based on the number of novel (and redundant) subtopics it is relevant to. This can be seen as modeling a user who is always interested in seeing more novel subtopics, with progressively decreasing interest in seeing the same subtopic multiple times. We put this model to test: if it is correct, then users, when given a choice, should prefer to see a document that has more value to the evaluation. We formulate some specific hypotheses from this model and test them with actual users in a novel preference-based design in which users express a preference for document A or document B *given* document C. We argue that while the user study shows the subtopic model is good, there are many other factors apart from novelty and redundancy that may be influencing user preferences. From this, we introduce a new framework to construct an ideal diversity ranking using only preference judgments, with no explicit subtopic judgments whatsoever.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]

Keywords: diversity, user study, preference judgments

1. INTRODUCTION

Research on novelty and diversity aims to improve the effectiveness of search engines by providing results that serve a range of possible user intents for the given query. These problems have been the subject of much interest in IR and web search recently, including the focus of a TREC task[1]. Batch effectiveness evaluation of retrieval systems serves several important purposes: first, giving developers and researchers a measurable objective; second, allowing for failure analysis and troubleshooting; and third, trying to estimate how useful search results will be to users. For the last of these, it is helpful to think of the evaluation measure and relevance judgments as a model of user utility. Measures like precision and recall can be seen as modeling utility in terms of the proportion of retrieved documents that are relevant and the proportion of relevant documents retrieved; measures like discounted cumulative gain or expected reciprocal rank offer a more refined model of user utility that incorporates graded judgments and rank-based discounts.

None of these models capture novelty or redundancy in ranked results. All of them will reward a system for retrieving the same relevant document 10 times in a row, or 10 relevant documents that are only superficially different from each other; while that may be useful for knowing whether retrieval features are working correctly, it is not likely to be very useful to a user. *Diversity evaluation* attempts to model novelty and redundancy in ranked results so as to give a more precise model of user utility.

Current diversity evaluation measures in the literature require judgments of relevance to *subtopics* (also called aspects, facets, or nuggets) of a topic. For example, judgments for the query *Windows* would include binary relevance judgments for each document for the subtopics *window panes*, *windows operating system*, etc. These subtopic judgments are used to determine whether a document is redundant with a previously-ranked document, or whether it contains some new information that a user might find interesting, or whether it is relevant to an alternative intent and perhaps not useful to this user but still useful enough to a different user. Measures like α-nDCG, ERR-IA, subtopic recall, and D-measures [10, 8, 17, 14] all use this same basic model, assigning more value to a document with more novel subtopics and less value to one with more redundant subtopics.

Like any model, the subtopic model surely has shortcomings. Novelty and redundancy are certainly not the only reasons a user might prefer one relevant document over another. Apart from a study by Sanderson et al. that showed that user preferences for rankings correlate with α-nDCG [15], there has not been much work on validating this model against real user preferences. And if the model does *not* track user preferences, then it is hard to justify its continued use: it conflates various aspects of effectiveness in such a way that, if used as an objective function, it can be difficult to understand the precise effect of change in the ranker.

[1] TREC 2009-2011 Web Track Diversity task focuses on novelty and diversity in search results.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

Fortunately these measures produce directly-testable hypotheses about user preferences. In this work we describe a novel user study to test these hypotheses. In Section 2 we start by describing the diversity retrieval problem in more detail and define the model more precisely. In Section 3 we present a user study, including a crowdsourced design; we show that while the model is not perfect, it is certainly not invalid. Section 4 builds on this by presenting a preference-based method for determining a diversity-aware ranking of documents. We conclude in Section 5.

2. NOVELTY RANKING TASK

Consider a user that has an unambiguous but broad information need and goes to a search engine to help satisfy it. This user will input a query and then see a ranked list of results, some of which will be relevant and some of which will not. The user will presumably click the relevant results to view and absorb the information they contain. Ideally, each relevant result would provide some new information that was not provided by previous relevant results; in other words, the relevant results would not be redundant with each other. The idea is that, each time a user clicks on a new relevant document, the amount of knowledge a user gains must be maximized by the novel content in the document.

The goal of ranking documents with novelty is to ensure that each relevant document a user sees as they progress down a ranked list provides new, non-redundant information that will help them satisfy their need. This means that a ranking of documents cannot be based solely on the probability of relevance; the novelty of a document depends to no small degree on the documents that have been ranked above it. Similarly, evaluation of these results cannot be based solely on binary or even graded relevance judgments, since these judgments are made to individual documents independently of all the other documents that might have been ranked. Part of studying the task is defining evaluation measures that can model redundancy and novelty.

2.1 Relationship With Other Tasks

The novelty task has similarities with some existing tasks such as the diversity task studied as part of the TREC Web track. Diversity aims at retrieving a subset of documents that has the maximum coverage of subtopics with the assumption that different users may be interested in different subtopics. In novelty ranking, the goal is to provide a set of documents for a single topic from which the user can get as much information as possible for that particular topic. We assume all users are interested in all of the subtopics, like the standard ad hoc assumption that all users are interested in all of the relevant material.

2.2 Intrinsic vs Extrinsic Diversity

Researchers in the past have identified two types of diversity: extrinsic and intrinsic [12]. Extrinsic diversity addresses the uncertainty in an ambiguous query where the intent is unclear and is best served by a ranking of documents covering several intents. Intrinsic diversity can be described as diversification that focuses on reducing redundancy and providing *novel* information for an unambiguous but still underspecified information need. In our work, we focus on intrinsic diversity which we refer to as *novelty ranking*, as we believe it will be easier for assessors to express preferences when there is no ambiguity of intent.

2.3 Evaluation

Evaluation measures for novelty and diversity must account for both relevance and novelty in the result set. It is important that redundancy caused by documents containing previously retrieved subtopics be penalized and documents containing novel information be rewarded. Most evaluation measures solve this problem by requiring that the subtopics for a query be known and that documents have been judged with respect to subtopics.

2.3.1 Existing Evaluation Measures

Subtopic recall. Subtopic recall measures the number of unique subtopics retrieved at a given rank [17]. Given that a query q has m subtopics, the subtopic recall at rank k is given by the ratio of number of unique subtopics contained by the subset of document up to rank k to the total number of subtopics m.

$$S\text{-}recall@k = \frac{\left|\bigcup_{i=1}^{k} subtopics(d_i)\right|}{m} \tag{1}$$

α-nDCG scores a result set by rewarding newly found subtopics and penalizing redundant subtopics. In order to calculate α-nDCG we must first compute the gain vector [10]. The gain vector is computed by summing over subtopics appearing in the document at rank k:

$$G[i] = \sum_{j=1}^{m} (1-\alpha)^{c_{j,i}-1} \tag{2}$$

where $c_{j,i}$ is the number of times subtopic j has appeared in documents up to (and including) rank i. Once the gain vector is computed, a discount is applied at each rank to penalize documents as the rank decreases. The most commonly used discount function is the $log_2(1+i)$, although other discount functions are possible. The *discounted cumulative gain* is given by

$$\alpha DCG@k = \sum_{i=1}^{k} \frac{G[i]}{log_2(1+i)} \tag{3}$$

α-DCG must be normalized to compare the scores against various topics. This is done by finding an "ideal" ranking that maximizes α-DCG, which can be done using a greedy algorithm. The ratio of α-DCG to that ideal gives α-nDCG.

Intent-aware family. Agrawal et al. studied the problem of answering ambiguous web queries, which is similar to the subtopic retrieval problem [2]. The focus of their evaluation measure is to measure the coverage of each intent separately for each query and combine them with a probability distribution of the user intents. They call this the *intent-aware* family of measures. It can be used with most of the traditional measures for evaluations such as precision@k, MAP, nDCG, and so on.

ERR-IA. Expected Reciprocal Rank (ERR) is a measure based on "diminishing returns" for relevant documents [9]. According to this measure, the contribution of each document is based on the relevance of documents ranked above it. The discount function is therefore not just dependent on the rank but also on relevance of previously ranked documents. A weighted average of the ERR measures for each interpretation would give the intent-aware version of ERR [8].

D-Measure. The D and the D# measures described by Sakai et al. [14] aims to combine two properties into a single evaluation measure. The first property is to retrieval documents covering as many intents as possible and the second is to rank documents relevant to more popular intents higher than documents relevant to less popular intents.

2.3.2 *Principles of Existing Evaluation Measures*

All of these measures estimate effectiveness of a system's ranking by iterating over the ranking, rewarding relevant documents containing a unseen subtopic(s) and penalizing relevant documents containing subtopic(s) seen before in the ranking. They are all based on a few principles in general:

1. A document with more unseen subtopics is worth more than a document with fewer unseen subtopics;
2. A document with both unseen and already-seen subtopics is worth more than a document with only the same unseen subtopic;
3. A document with unseen subtopics is worth more than a document with only redundant subtopics.

One of our goals with this work is to test whether these principles hold for real users.

2.4 Data

Our analysis was conducted primarily on the *Newswire* data created by Allan et al. [3] to investigate the relationship between system performance and human performance on a subtopic retrieval task. The data consists of 61 topics, each with a short (3-6 word) query, and judgments of relevance to documents in a subset of the TDT5 corpus. The *Newswire* data includes relevance judgments for the top 130 documents retrieved by a query-likelihood language model for the short query for each query. The judgments consists of binary relevance judgments for each document, and for each relevant document, a list of subtopics contained in that document. This data reflects an intrinsic diversity task and is therefore most appropriate to this work.

3. FACTORS INFLUENCING USER PREFERENCES

In Section 2.3.2, we identified some principles on which the evaluations for diversity are based on. In this section we tests if these principles hold for real users and further study in detail the role of subtopics in influencing user preference. Although in practice the same evaluation measures are used for both intrinsic and extrinsic diversity, our focus is on intrinsic diversity as it is easier for assessors to understand the concept of relevance when there is no ambiguity of intent. We explore the factors that influence user preference for novelty ranking using a preference based framework.

3.1 Triplet Framework

The idea of pairwise preference judgments is relatively new in the IR literature, having been introduced by Rorvig in 1990 [13] but not subject to empirical study until the past several years [6, 5]. Comparison studies between absolute and preference judgments show that preference judgments can often be made faster than graded judgments, with better agreement between assessors (and more consistency with individual assessors) [6]. Also with preferences tassessors can make much finer distinctions between documents.

We propose a preference-based framework to study novelty consisting of a set up in which three relevant documents that we refer to as a *triplet* are displayed such that one of them appears at the top and the other two are displayed as a pair below the top document. We will use D_T, D_L, and D_R to denote the top, left, and right documents respectively, and a triplet as $\langle D_L, D_R | D_T \rangle$. An assessor shown such a triplet would be asked to choose which of D_L or D_R they would prefer to see as the *second* document in a ranking given that D_T is first, or in other words, they would express a preference for D_L or D_R conditional on D_T. For the purpose of this study we will assume we have relevance judgments to a topic, and for each relevant document, binary judgments of relevance to a set of subtopics. Thus we can represent a document as the set of subtopics it has been judged relevant to, e.g. $D_i = \{S_j, S_k\}$ means document i is relevant to subtopics j and k. Varying the number of subtopics in top, left and right documents yields specific hypotheses about preferences for novelty over redundancy.

3.2 Hypotheses

The triplet framework allows us to collect judgments for novelty based on preferences and also enables us to test various hypotheses. As discussed above, varying the number of subtopics in D_T, D_L and D_R it is possible to enumerate various hypotheses concerning the effect of subtopics in a document. We define two types of hypotheses; one very specific with respect to subtopic counts and redundancy, and the other more general.

Hypothesis Set 1 : First we propose the simplest possible hypotheses that capture the three principles above. We will denote a preference between two documents using $\succ$, e.g. $D_L \succ D_R$ means document D_L is preferred to document D_R. Then the three hypotheses stated formally are:

H_1: if $\langle D_L, D_R | D_T \rangle = \langle \{S_2\}, \{S_1\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (novelty is better than redundancy)

H_2: if $\langle D_L, D_R | D_T \rangle = \langle \{S_1, S_2\}, \{S_2\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (novelty+redundancy is better than novelty alone)

H_3: if $\langle D_L, D_R | D_T \rangle = \langle \{S_2, S_3\}, \{S_2\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (novelty+novelty is better than novelty alone)

Hypothesis Set 2 : Here we define a class of hypotheses in which the number of subtopics contained in each document in a triplet is categorized by relative quantity. We identify six variables based on number of subtopics that almost completely describe the novelty and redundancy present in the triplet. The six variable are as follows:

1. Tn - Number of subtopics in D_T;
2. NLn - Number of subtopic in D_L not present in D_T;
3. NRn - Number of subtopic in D_R not present in D_T;
4. Sn - Number of subtopics shared between D_L and D_R;
5. RLn - Number of subtopics in D_L and present in D_T;
6. RRn - Number of subtopics in D_R and present in D_T.

The number of subtopics for each of the six variables are categorized as *low* or *high*. The six variables enable us to test the effect of novelty and redundancy w.r.t the number of subtopics in a triplet. The variables NLn and NRn focus on novelty whereas RLn and RRn focuses on redundancy. For instance, by varying NLn and NRn and holding the other variables constant, it is possible to test the effect of the relative quantity of novel subtopics in a document.

3.3 Experimental Design

In this section we describe the experimental design used to test the hypotheses defined above. Notice that the defined hypotheses are based on the number of subtopics contained in the documents and they fit into the triplet framework which requires conditional preference judgments. Therefore, to test our hypotheses, two kinds of judgments were needed: subtopic level judgments and conditional preference judgments. The subtopic level judgments were obtained from the data described in Section 2.4. Conditional preference judgments were collected using crowd sourcing as it is a fast, easy and a low cost way of collection user judgments [4].

We used Amazon Mechanical Turk (AMT) [1]; an online labor marketplace to collect user judgments. AMT works as follow: requestor create a group of Human Intelligence Task (HITs) with various constraints and worker from the marketplace works on these task to complete the task. In this work we, use a design similar to the one used by Chandar and Cartertte [7]. Designing a user study using AMT involves deciding on the HIT layout and HIT properties.

3.3.1 HIT Layout

In order to collect user judgments for our hypotheses using AMT, we had to organize the triplets satisfying a given hypothesis into HITs. Each HIT consisted of the following (in order of display): a set of instructions about the task, original keyword query, topic description, five preference triplets, and a comment field allowing worker to provide feedback. A brief description about each element is given below:

Guidelines: Workers were provided with a set of instructions and guidelines prior to judging. Guidelines specified that workers should assume that everything they know about the topic is in the top document and are trying to find a document that would be most useful for learning more about the topic. Guidelines did not mention anything about subtopics, or even novelty/redundancy except as examples of properties assessors might take into account in their preferences (along with recency, ease of reading, and relevance).

Query text and topic description: The *query text* described the topic in a few words (we used the topic "titles" in the traditional TREC jargon) and *topic description* provided a more verbose and informative description about the topic. Again, there was no mention of explicit subtopics.

Preference triplet: Figure 1 shows an example preference triplet with the query text and topic description. Each preference triplet consists of *three* documents, all of which were relevant to the topic and the document were picked randomly from the data described in Section 2.4 to meet the constraints of a given hypothesis. One document appeared at the top followed by two documents below it, the triplets were chosen randomly such that the hypothesis constraints were satisfied. A HIT consisted of five preference triplets belonging to the same query shown one below the other.

The triplets were followed by a preference option for the workers to indicate which of the two documents they preferred. The workers were asked to pick the document from the lower two that provided the most new information, assuming that all the information they know about the topic is in the top document. They could express a preference based on whatever criteria they liked; we listed some examples in the guidelines. Note that we do not show them any subtopics, nor do we ask them to try to determine subtopics and make a preference based on that.

Comments Field was provided at the end, so that the workers could to provide a common feedback for all the five triplets, if they chose to do so.

3.3.2 HIT Properties

Workers are paid for each HIT they complete and picking an appropriate amount for each task is always tricky. In our study, workers were paid $0.80 for each completed HIT. Also each HIT had a time limit of three hours before which the HIT had to be completed. While the actual task might not take three hours to complete; the extra time allows them to take breaks if needed, since the workers had to read fifteen documents per HIT. We had five separate workers judge each HIT for the our *first set of hypotheses* and three separate workers judge each HIT for the our *second set of hypotheses*.

3.3.3 Triplets

Triplets were generated by randomly picking the three relevant documents for a given query and representing them as subtopic(s). Triplets for the first set of hypotheses in Section 3.2 were considered such that the constraints are satisfied for each hypothesis. For example, for hypothesis H_1 given a query x the triplet would consist of D_T containing only the subtopic S_1 and D_L containing the subtopics S_1 and S_2. Six queries were used to test the first set of hypotheses with four triplets for each query.

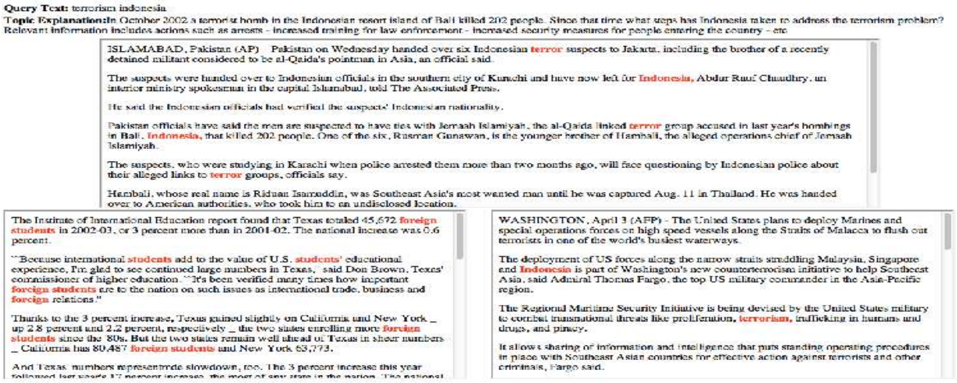
The triplets were generated in a similar way for the second set of hypotheses but the constraints for each hypothesis were based on the six variables described in 3.2. For example, a triplet with a variable setting of Tn = High, Sn = High, NLn = High and NRn = High would contain 5 or more subtopics in the top document D_T and 3 or more subtopics in the left and right documents (D_L and D_R) such that there are 1 or more subtopics shared between D_L and D_R. The details of the number of subtopics for each categories of high and low levels for each variables are provided in the Table 1. Eight queries were used to test the second set of hypotheses with four different triplets for each query.

3.3.4 Quality Control

There are two major concerns in collecting judgments through crowdsourcing platform such as AMT. One is "Do the workers really understand the task?" and the other is "Are they making faithful effort to do the work or clicking randomly?". We address these concerns using three techniques: majority vote, trap questions, and qualifications.

Majority vote: Since novelty judgments to be made by the workers are subjective and it is possible some workers are clicking randomly, having more than one person judge a triplet is common practice to improve the quality of judgments. In our study, each HIT was judged by 5 or 3 different workers (depending on hypothesis set). We look at the individual preferences as well as the majority preference.

Trap questions: Triplets for which answers are obvious were included to assess the validity of the results. We included two kinds of trap questions: "non-relevant document trap" and "identical document trap". For the former, one of the bottom two documents was not relevant to the topic and should never be preferred. For the latter, the top document and one of the bottom two documents were the same. The workers were expected to pick the non-identical document as it provides novel information. One of the five triplets in a HIT was a trap and the type was chosen randomly.



Query Text: terrorism indonesia

Topic Explanation: In October 2002 a terrorist bomb in the Indonesian resort island of Bali killed 202 people. Since that time what steps has Indonesia taken to address the terrorism problem? Relevant information includes actions such as arrests - increased training for law enforcement - increased security measures for people entering the country - etc

ISLAMABAD, Pakistan (AP) Pakistan on Wednesday handed over six Indonesian **terror** suspects to Jakarta, including the brother of a recently detained militant considered to be al-Qaida's pointman in Asia, an official said.

The suspects were handed over to Indonesian officials in the southern city of Karachi and have now left for **Indonesia,** Abdur Rauf Chaudhry, an interior ministry spokesman in the capital Islamabad, told The Associated Press.

He said the Indonesian officials had verified the suspects' Indonesian nationality.

Pakistan officials have said the men are suspected to have ties with Jemaah Islamiyah, the al-Qaida linked **terror** group accused in last year's bombings in Bali, **Indonesia,** that killed 202 people. One of the six, Rusman Gunawan, is the younger brother of Hambali, the alleged operations chief of Jemaah Islamiyah.

The suspects, who were studying in Karachi when police arrested them more than two months ago, will face questioning by Indonesian police about their alleged links to **terror** groups, officials say.

Hambali, whose real name is Riduan Isamuddin, was Southeast Asia's most wanted man until he was captured Aug. 11 in Thailand. He was handed over to American authorities, who took him to an undisclosed location.

The Institute of International Education report found that Texas totaled 45,672 **foreign students** in 2002-03, or 3 percent more than in 2001-02. The national increase was 0.6 percent.

``Because international **students** add to the value of U.S. **students'** educational experience, I'm glad to see continued large numbers in Texas,'' said Don Brown, Texas' commissioner of higher education. ``It's been verified many times how important **foreign students** are to the nation on such issues as international trade, business and **foreign** relations.''

Thanks to the 3 percent increase, Texas gained slightly on California and New York _ up 2.8 percent and 2.2 percent, respectively _ the two states enrolling more **foreign students** since the '80s. But the two states remain well ahead of Texas in sheer numbers _ California has 80,487 **foreign students** and New York 63,773.

And Texas' numbers representrode slowdown, too. The 3 percent increase this year followed last year's 17 percent increase, the most of any state in the nation. The national

WASHINGTON, April 3 (AFP) - The United States plans to deploy Marines and special operations forces on high speed vessels along the Straits of Malacca to flush out terrorists in one of the world's busiest waterways.

The deployment of US forces along the narrow straits straddling Malaysia, Singapore and **Indonesia** is part of Washington's new counterterrorism initiative to help Southeast Asia, said Admiral Thomas Fargo, the top US military commander in the Asia-Pacific region.

The Regional Maritime Security Initiative is being devised by the United States military to combat transnational threats like proliferation, **terrorism,** trafficking in humans and drugs, and piracy.

It allows sharing of information and intelligence that puts standing operating procedures in place with Southeast Asian countries for effective action against terrorists and other criminals, Fargo said.

Figure 1: Screenshot of the preference triple along with the query text and description.

variable	number of subtopics	
	low	high
Tn	1-4	5-9
Sn	0	1-2
NLn	0-2	3-6
NRn	0-2	3-6
RLn	0	1-2
RRn	0	1-2

Table 1: Number of subtopics corresponding to the high and low categories for each variable in our second set of hypotheses.

Qualifications: It is possible to qualify workers before they are allowed to work on your HITs in AMT. Worker qualifications can be determined based on historical performance such as percentage of approved HITs. Also, worker's qualification can be based on a short questionnaire or a test. The two qualifications used in are study are explained below:

1. **Approval rate:** HITs can be restricted to workers with an overall minimum percentage of approval. It is commonly used for improving accuracy and reducing spammer from working on your task. An overall approval rate of *95%* was required to work on our HITs.
2. **Qualification test:** Qualification tests can be used to ensure that workers have the required skill and knowledge to perform the task. In our case, workers had to be trained to look for documents that provide novel information given the top document. We created a qualification test having the same design layout as the actual task but had only three triplets. Two of the three triplets were identical document traps and the other was a non-relevant trap with instructions for each triplet aiding in making a preference.

3.4 Results and Analysis

Judgments for a total of 60 triplets (out of which 12 triplets were traps) were obtained for the hypothesis set 1. Since we had each triplet assessed by five separate assessors, a total of 300 judgments were collected out of which 60 were traps. We had 39 unique workers (identified by worker ID) on AMT judge these triplets across six topics.

Table 2 shows results for H_1. It turns out that there is no clear preference for either redundant or novel documents for the four queries. For two of our queries assessors tended to prefer the novel choice; for the other two they tended to prefer the redundant choice. When we use majority vote to determine a consensus for each triplet, we find that the outcomes are exactly equal. Thus while we cannot reject H_1, we have to admit that if it holds it is much less strong than we expected.

H_1	all prefs		consensus	
topic	same	new	same	new
childhood obesity	6	14	1	3
terrorism indonesia	8	12	1	3
earthquakes	15	5	3	1
weapons for urban fighting	15	5	3	1
total	**44**	**36**	**8**	**8**

Table 2: Results for H_1: that novelty is preferred to redundancy. The "all prefs" columns give the number of preferences for the redundant and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

H_2	all prefs		consensus	
topic	new	same+new	new	same+new
kerry endorsement	9	11	2	2
childhood obesity	4	16	0	4
terrorism indonesia	13	7	4	0
libya sanctions	4	16	0	4
total	**30**	**50**	**6**	**10**

Table 3: Results for H_2: that novelty and redundancy together are preferred to novelty alone. The "all prefs" columns give the number of preferences for the redundant+novel document and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

Table 3 shows a clearer (but still not transparent) preference for H_2, novelty and redundancy together over novelty alone. Over all assessors and all triplets, the preference is significant by a binomial test (50 successes out of 80 trials; $p < 0.05$). Still, there is one query ("john kerry endorsement") for which the difference is insubstantial, and one that has the opposite result ("terrorism indonesia"). The latter

H_3	all prefs		consensus	
topic	new	new+new	new	new+new
kerry endorsement	9	11	1	3
childhood obesity	3	17	0	4
terrorism indonesia	2	18	0	4
libya sanctions	8	12	1	3
total	**22**	**58**	**2**	**14**

Table 4: Results for H_3: that two novel subtopics are preferred to one. The "all prefs" columns give the number of preferences for the novel+novel document and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

topic	high≻low	left≻ right
earthquakes	76 - 20 (79%)	96 - 96 (50%)
terry nichols guilt evidence	75 - 21 (78%)	100 - 92 (52%)
medicare drug coverage	73 - 23 (76%)	86 - 106 (45%)
oil producing countries	65 - 31 (68%)	89 - 103 (46%)
no child left behind	62 - 34 (65%)	81 - 111 (42%)
european union member	61 - 35 (64%)	103 - 89 (54%)
german headscarf court	59 - 37 (61%)	84 - 108 (44%)
ohio highway shooting	51 - 45 (53%)	104 - 88 (54%)
total	**522 - 246 (68%)**	**743 - 793 (48%)**

Table 5: Results of preference judgments by the number of new subtopics in D_L, D_R over D_T (variables NLn, NRn). Counts are aggregated over all values of Tn, Sn per query. The first column gives preference counts for the document with more new subtopics over the document with fewer when $NLn \succ NRn$. The second column is the baseline, giving counts for preferences for left over right.

case is particularly interesting because it is the opposite of what we would expect after seeing the results in Table 2: given that assessors preferred redundant documents to novel documents for that query, why would they prefer novel documents to documents with both novelty and redundancy?

Table 4, with results for H_3, is the strongest positive result: a clear preference for documents with two new subtopics over documents with just one. In this case both results are significant (58 successes out of 80 trials and $p < 0.0001$ over all triplets and all assessors; 14 successes out of 16 trials and $p < 0.01$ for majority voting). Nevertheless, there are still queries for which the preference is weak.

Based on this, it seems like *novelty* + *novelty* > *novelty*, *novelty*+*redundancy* ≥ *novelty*, but not *novelty* ≥ *redundancy*.

There were a total of 640 triplets (out of which 128 triplets were traps) for the second part of our study. Each of these triplets were judged by three separate assessors, thus a total of 1920 judgments were made out of which 384 were traps. And for this study we had 38 unique workers (identified by worker ID) on AMT working on our triplets. Some of these workers had worked on the first study as well. Almost 70% of the judgments were completed by 15% of the workers and about 93% of the irrelevant traps were passed by the workers. This power law distribution for our task has been observed earlier for other tasks as well [11], we hope to investigate on this issue in the future.

Triplets were generated by controlling four variable: Tn, Sn, NLn and NRn, we obtained sixteen unique settings for the four variable combination as each of the four variables were categorized into *low* and *high* with equal number of triplet in each setting. This allowed us to perform ANOVA such that the number of new subtopics in the left or right document was the primary predictor of preference, with the number of subtopics in the four variables as the secondary predictors. ANOVA indicated that there is a lot of residual variance, suggesting there are various factors influencing preferences that we have not included in the model.

Table 5 analyzes preferences for more new subtopics in D_L or D_R over fewer new subtopics (variables NLn and NRn) by topic. We looked at four cases: the first two (NLn high, NRn low; NLn low, NRn high) can tell us whether users prefer to see more new subtopics over fewer, while the second (NLn high, NRn high; NLn low, NLn low) along with the first two give us a baseline preference for left over right. While we would expect the baseline preference to be 50% (since which document appears on the left versus right is randomized), there may be other unmodeled factors that cause it to be more or less than 50%, so it is useful to compare to this baseline.

It is clear from this table that users as a group prefer to see more new subtopics, just as we saw in the results for H_3 above. Still, there are individual queries for which that preference is not strong, especially when compared to the baseline (e.g. the "Ohio highway shooting" topic), and even when the preference is strong in aggregate there are cases where they do not hold.

There is some effect due to the number of subtopics in D_T, with preferences for more new subtopics stronger when Tn is low. When it is low, the preference for high versus low is 271 to 113 (70%) against a baseline preference for left over right of 347 to 421 (45%)[2]. When Tn is high, the preference for high versus low is 251 to 133 (65%) against a baseline of 396 to 372 (52%). We conjecture that when the top document already has a lot of information about the topic, there is a little less reason to prefer either left or right regardless of how many subtopics they contain.

There is not much effect due to the number of shared subtopics between D_L and D_R. When Sn is low, the preference for more new subtopics over fewer is 268 to 116 (70%) against a baseline of 370 to 398 (48%); when it is high, the preference for more new is 254 to 130 (66%) against a baseline of 373 to 395 (49%). This may be because fewer shared subtopics makes it easier to express a preference. However, the effect is too small to draw any firm conclusion.

There is interesting interaction between the number of new subtopics and the number of redundant subtopics in D_L and D_R. When one has a high number of new subtopics and the other has a low number of new subtopics, number of *redundant* subtopics seems to influence the strength of preference for the one with more new subtopics: if there are more redundant subtopics along with the new subtopics, the preference is 118 to 44 (73%), but when there are fewer redundant subtopics with more new subtopics and more redundant subtopics with fewer new subtopics, the preference is even at 51 to 51 (50%). This suggests again that users *like* redundancy, sometimes enough to overcome a lack of novelty. However we must note that data here is sparse, also the two variables RRn and RLn were not the ones that we controlled for in our experiment.

[2] We presume that the greater-than-expected preference for the right document is just due to random chance.

Topic	Agreement	No. triplets
childhood obesity	0.71	15
weapons for urban fighting	0.92	5
kerry endorsement	0.58	10
libya sanctions	0.62	10
earthquake	0.72	5
terrorism indonesia	0.71	15
Mean	**0.69**	**60**

Table 6: Interassessor agreement scores for each topic for the first study.

Topic	Agreement	No. triplets
oil producing countries	0.63	80
terry nichols guilt evidence	0.72	80
no child left behind	0.61	80
german headscarf court	0.57	80
medicare drug coverage	0.66	80
earthquakes	0.65	80
european union member	0.59	80
ohio highway shooting	0.59	80
Mean	**0.63**	**640**

Table 7: Interassessor agreement scores for each topic for the second study.

3.5 Interassessor Agreement

As described above, each triplet was judged by five different workers for the first study (hypotheses set 1) and by three workers for the second study (hypotheses set 2). We calculated an inter-assessor agreement score for each triplet for the first study as follows. The judgments were considered as 10 pairs of answers given for a single triplet, adding 1 points to the score if the two workers agreed (complete agreement); and adding nothing if they judged different documents (no agreement). The perfect agreement would sum up 10 points, so we divided the score obtained by 10 and normalized from 0 (no agreement at all) to 1 (perfect agreement). Mean agreement for the first study is given in Table 6 for each query. Overall a high mean agreement of 0.7 was found across all triplets and the scores are close to the agreement observed previously [6]. Since the mean agreement was quite high for the first study, it encouraged us to reduce the number of workers for each triplet and increase the number of queries for the second study. Similar mean agreement can be seen for the second study in Table 7.

3.6 Possible Confounding Effects in Display

The way the hits were displayed may introduce some confounding effects, possibly causing assessors to choose documents for reasons other than novelty or redundancy. We investigated two such effects:

Document length A preference towards shorter documents was observed in general, though the preference gets weaker over the three hypotheses. For H_1, assessors preferred the shorter document in 79% of triplets. For H_2, that decreased to 71% of triplets, and for H_3 it dropped steeply to only 44%. However, it is also true that the mean difference in length for the pair of documents they were choosing between was greatest for H_1 triplets and least for H_3 triplets (H_1:158 terms, H_2:126 terms, H_3:47 terms). Therefore its safe to conclude there seems to be a preference towards shorter documents.

Highlighted terms It turns out that assessors tended to prefer the document with *fewer* highlighted query terms. For H_1, assessors preferred the document with more query terms only 35% of the time. For H_2 that drops to 13%, and for H_3 it comes back up to 29%. The mean difference in number of query term occurrences is quite low, only on the order of one additional occurrence on average for H_1 and H_3 documents, and only 0.2 additional occurrences for H_2. While the effect is significant, it seems unlikely that assessors can pick up on such small differences. We think the effect is more likely due to the distribution of subtopics in documents.

3.7 Additional Investigation

While the results suggest that the number of subtopics influences user preferences, it is also clear that from the analysis that other factors are affecting preferences. The results from H_1 and the weaker preference in H_2 were not what we expected. We investigated this more by looking at a number of triplets ourselves and identifying some new hypotheses about why assessors were making the preferences they were. From looking at triplets for the "earthquakes" topic, we identified three possible reasons for preferring a document with a redundant subtopic:

- it updates or corrects information in the top document;
- it significantly expands on the information in the top document;
- despite having a novel subtopic, the other choice provides little information of value.

This suggests to us that there are other factors that affect user preferences, in particular recency, completeness, and value. It may also suggest that there are implicit subtopics (at finer levels of granularity) that the original assessors did not identify, but that make a difference in preferences. None of this is surprising, but there is currently no evaluation paradigm of note that take all of these factors into account in a holistic way. Preference judgments can, and this analysis suggests additional hypotheses for testing with preferences.

4. PREFERENCE JUDGMENTS FOR AN IDEAL NOVELTY RANKING

The user study shows that although users tend to prefer documents containing more novel subtopics, it is also evident that factors other than subtopics play a vital role. The study also shows that the presence of subtopic in a document is taken into account implicitly and preferences are based not only on the number of subtopics but also on several other factors that include subtopic importance, relevance of the subtopic, readability of the document, etc. In this section, we propose an approach that attempts to capture these factors implicitly using a preference based framework to form a full ranking of documents with novelty as an implicit quality.

Our approach involves a series of sets of preference comparisons. Each set is essentially a comparison sort algorithm, with the comparison function a simple preference conditional on information contained in top-ranked documents from prior sets of comparisons, generalizing the triplet framework we introduced above.

The first set of preferences is meant to produce a relevance ranking: given a choice between two documents, assessors select the one they prefer, with topical relevance being the primary consideration in the judgment. Once these comparisons are done for all pairs, it is possible to obtain the best or "most relevant" document, i.e. the most preferred document based on the number of times a document was selected.

For the second set of preferences, the assessor needs to consider the *novelty* of information in the document along with relevance. This leads to exactly the triplet framework we used previously. For this second set, the assessor will see the top-ranked document from the previous set as D_T, then pick from two documents D_L, D_R conditional on that.

The sequence continues by adding more documents to the top. For the third set, the comparison involves information in *two* previously ranked documents along with a pair of documents; for the fourth, it involves information in three previously ranked documents along with a pair. This continues to the final set, in which there are only two documents to compare conditional on $n-2$ previous top documents.

When complete, the most preferred document in the first set takes rank 1, the most preferred document in the second set takes rank 2, and so on. Observe that the first set of judgments correspond to relevance judgments and sets 2 through $n-1$ correspond to novelty.

This method asks for a very large number of preferences: if fully judged, there would be $O(n^2)$ preferences in the first set, $O((n-1)^2)$ in the second, and so on, for a total of $O(n^3)$ judgments, which is almost certainly infeasible. We hypothesize that the first two sets of preferences (one for relevance and one for novelty) will provide a near-optimal approximation to the full set and if judgments are transitive (that is, if document A is preferred to B and B is preferred to C, then A should be preferred to C as well), the number of judgements needed can be reduced drastically. We will test both of these hypotheses below.

4.1 Experimental Design

As described above, we asked assessors to make the first two sets of judgments for each topic. The first set of judgments attempts to rank documents by relevance to the topic; intuitively, these judgments could be used to find the most relevant document in the ranked list: that which is preferred to everything else (assuming judgments are transitive) is most relevant. The second set of judgments attempts to rank the remaining documents by the degree of novelty they provide given that we know the document that is ranked at position one from the first set.

For this experiment we elected not to use MTurk. We wanted a single assessor to do all the preferences for a single topic, first so they would be able to build a familiarity with the topic as they judge, and second so we could assess their self-consistency. Thus we asked students at our institution to participate in the study. These students are mostly in computer science, mostly studying NLP and language technologies. Like the workers in the previous section, they were not given explicit instruction regarding subtopics; they were only asked to express a preference. We had 6 assessors complete preferences for at least one topic.

We designed two new web interfaces running on a local server to be used by assessors to collect preferences for both relevance and novelty, the first two sets of preferences described above. Common elements in both interfaces are the original keyword query, topic description, article texts (with query keywords highlighted), preference buttons for indicating which of the two documents the assessor prefers, a progress bar with a rough estimate of the percentage of preferences completed, and a comment field allowing them to say why they made their choice (if they wish). Elements specific to each experiment are described in more detail in the respective sections below.

For this study, we asked assessors to judge all pairs of documents in the first two sets. Topics were chosen from the data described in Section 2.4. We wanted to include all known relevant documents for the topic in the preference experiment. Since we were asking assessors for all pairs, we limited our selection to topics with a relatively small number of relevant documents. We then added a randomly-selected set of nonrelevant documents from among the top-ranked documents for the topic. We kept the total number of preferences in an experiment to less than 200.

The first two documents shown to an assessor were chosen randomly from the set of all documents to be ranked. After that, whichever document the assessor preferred remained fixed in the interface; only the other document changed. This way the assessor only had to read one new document after each judgment, just as they would in normal single-document assessing. Furthermore after the first $O(n)$ judgments we know the top-ranked document for the current set, and thus if transitivity holds it follows that we only need a linear number of preferences at each set.

4.1.1 First Level Judgments: Relevance Preferences

In the first set of judgments, the assessor was shown two documents (news articles) and a statement of an information need (a topic); the task was to pick the most preferred document using the "prefer left" or "prefer right" buttons. A screenshot of the first level judgments is shown in Figure 2

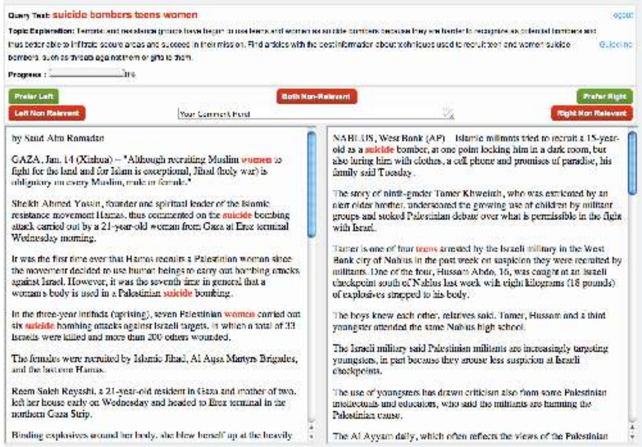
The assessor was provided with a set of instructions and guidelines prior to judging. The guidelines specified that the assessor should assume they know nothing about the topic and are trying to find documents that are topically relevant, that is, that provide some information about it. If a document contains no topical information, the assessor could judge it "not relevant"; if they do so, the system will assume they prefer every other document to that one and remove it from this set as well as all subsequent sets so it will not be seen in future comparisons. Assessors could also judge "both not relevant" to remove both from the set and see a new pair. These buttons can make the task easier by reducing the total number of preference judgments the assessors need to make.

If both documents were topically relevant, the assessor could express a preference based on whatever criteria they liked. Some suggestions included in the guidelines were: one document is more focused on the topic than the other; one document has more information about the topic than the other; one document has more detailed information than the other; one document is easier to read than the other. Assessors could exit for a break as long as they liked and return at the point where they stopped. A progress indicator let them know roughly how close they were to the end

4.1.2 Second Level Judgments: Novelty Preferences

For the second set of preferences, the assessor was shown *three* documents and a statement of an information need (a topic); the task was to pick the most useful document from two of the three to learn more about the topic given what is presented in the third.

The interface for the second level judgment was very similar to the triplet layout shown is Figure 1. One document



Query Text: suicide bombers teens women

Topic Explanation: Terrorist and resistance groups have begun to use teens and women as suicide bombers because they are harder to recognize as potential terrorists and thus better able to infiltrate secure areas and succeed in their mission. Find articles with the best information about techniques used to recruit teen and women suicide bombers, such as threats against them or gifts to them.

Progress : 11%

by Saud Abu Ramadan

GAZA, Jan. 14 (Xinhua) -- "Although recruiting Muslim women to fight for the land and for Islam is exceptional, Jihad (holy war) is obligatory on every Muslim, male or female."

Sheikh Ahmed Yassin, founder and spiritual leader of the Islamic resistance movement Hamas, thus commented on the suicide bombing attack carried out by a 21-year-old woman from Gaza at Erez terminal Wednesday morning.

It was the first time ever that Hamas recruits a Palestinian woman since the movement decided to use human beings to carry out bombing attacks against Israel. However, it was the seventh time in general that a woman's body is used in a Palestinian suicide bombing.

In the three-year intifada (uprising), seven Palestinian women carried out six suicide bombing attacks against Israeli targets, in which a total of 33 Israelis were killed and more than 200 others wounded.

The females were recruited by Islamic Jihad, Al Aqsa Martyrs Brigades, and the last one Hamas.

Reem Saleh Reyashi, a 21-year-old resident in Gaza and mother of two, left her house early on Wednesday and headed to Erez terminal in the northern Gaza Strip.

Binding explosives around her body, she blew herself up at the heavily

NABLUS, West Bank (AP) -- Islamic militants tried to recruit a 15-year-old as a suicide bomber, at one point locking him in a dark room, but also luring him with clothes, a cell phone and promises of paradise, his family said Tuesday.

The story of ninth-grader Tamer Khweireh, who was extricated by an alert older brother, underscored the growing use of children by militant groups and stoked Palestinian debate over what is permissible in the fight with Israel.

Tamer is one of four teens arrested by the Israeli military in the West Bank city of Nablus in the past week on suspicion they were recruited by militants. One of the four, Hussam Abdo, 16, was caught at an Israeli checkpoint south of Nablus last week with eight kilograms (18 pounds) of explosives strapped to his body.

The boys knew each other, relatives said. Tamer, Hussam and a third youngster attended the same Nablus high school.

The Israeli military said Palestinian militants are increasingly targeting youngsters, in part because they arouse less suspicion at Israeli checkpoints.

The use of youngsters has drawn criticism also from some Palestinian intellectuals and educators, who said the militants are harming the Palestinian cause.

The Al Ayyam daily, which often reflects the views of the Palestinian

Figure 2: Screenshot of the preference collection interface for relevance preferences.

appeared at the top of the screen; this was the most preferred document as identified by the assessor after the first set of preferences. The assessors were asked to pick a document from a pair of documents (appearing below the top document) that provided the most novel information given that they know all the information in the top document.

Guidelines specified that the assessor should pretend that the top document is the entirety of what they know about the topic, and their goal is now to find the best document for learning *more* about the topic. Beyond that, they could express a preference based on whatever criteria they liked, including those listed above.

There are no nonrelevant judgment buttons in this interface. Any document that was judged nonrelevant in the first set of preferences will not be seen in this set. Anything that was relevant in the first set is assumed to still be relevant; if a relevant documents provides no new relevant information, we assume the assessor's preferences will result in that document being ranked near the bottom of this set.

4.2 Experimental Analysis

As described above, we conducted novelty preference judging with five topics from the data described in Section 2.4. On average, 16.8 documents were judged for each topic. A total of 605 pairs were judged for 4 topics by 6 assessors for experiment levels 1 and 2. We compared these judgments to the original subtopic-based judgments in the data.

Agreement on Relevance. To assess agreement on the relevance of a document, we assume that any document not explicitly judged not relevant must be relevant. We consider the original relevance judgments derived from the subtopic judgments as the ground truth and assess the performance of our assessors relative to that. Table 9 shows the confusion matrix between assessors making preference judgments and the original assessors making subtopic judgments; broad agreement on the two classes is 71%. Preference assessors identified 76% of the relevant documents that the original assessors found, and 60% of the documents judged relevant by at least one assessor were judged relevant by both. This is a high level of agreement for IR tasks; compare to the 40% agreement on relevance reported by Voorhees [16].

topic	Rank Correlation	
	Level 1	Level 2
OPEC actions	0.563	0.534
OPEC actions - Alternate	0.568	0.377
childhood obesity	0.467	0.264
childhood obesity - Alternate	0.403	0.394
suicide bombers teens women	0.320	0.200
foreign students visa restrictions	0.532	0.030

Table 8: Kendall's τ correlations between rankings from real preference judgments and rankings from simulated preference judgments (for the relevance ranking (level 1) and the novelty ranking (level 2)).

Preference Judgments	Subtopic Judgments	
	Relevant	Non-Relevant
Relevant	58	20
Non-Relevant	18	37

Table 9: Confusion matrix for relevance judgments derived from the subtopic judgments in the original Newswire collection and derived from our preference judgments (all queries aggregated).

Rank Correlation. Another way to compare preference judgments to the original subtopic judgments is by using both to construct a ranking of documents, then computing a rank correlation statistic between the two rankings. The subtopic judgments included in the *Newswire* data were obtained by assessors explicitly labeling subtopics for each relevant document. We use the subtopic information to simulate preference judgments that might have been obtained via our experiment. For the first set, we always prefer the document with the greatest number of subtopics. (Except in the case of a tie, when we prefer a random document.) For the second set, the top-ranked document from the first set becomes the "top document", and then for each pair we prefer the document that contains the greatest number of subtopics that are not in that top-ranked document. The final ranking has the most-preferred document from the first set of preferences at rank 1 followed by the ranking obtained from the second set of preferences.

Kendall's τ rank correlation for each topic for both level 1 and level 2 preference judgments is shown in Table 8. Kendall's τ ranges from -1 (lists are reversed) to 1 (lists are exactly the same), with 0 indicating a random reordering. The values we observe are positive and statistically significant (except for level 2 judgments for topic *foreign students visa restrictions*). Kendall's τ is based on pairwise swaps, and thus can be converted into agreement on pairwise preferences by adding 1 and dividing by 2. When doing this we see that agreement is again high for the relevance ranking, and also high for the novelty ranking, well over the 40% observed by Voorhees (except for topic *foreign students visa restrictions*). We believe this validates our second set of preferences, though certainly the question is not closed.

4.2.1 Transitivity in Preference Judgments

One issue in using our preference judgments for novelty is that the number of pairwise judgments increases quickly with number of documents. Increase in number of judgments means increase in assessor time, but if the assessors are consistent i.e. if their judgments are transitive, then we can

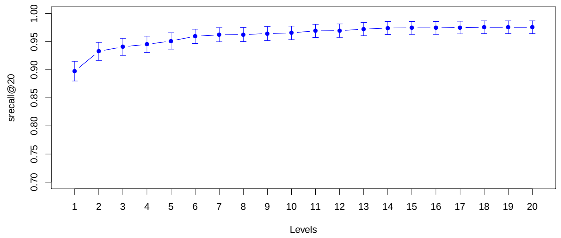


Figure 3: S-recall increases as we simulate deeper levels of preference judgments, but the first set of novelty preferences (level 2) gives an increase that nearly exceeds all subsequent levels combined.

reduce the number of preferences from $O(n^2)$ to $O(n \log n)$ at each level; furthermore, since we really only need the "best" document at each level, transitivity would allow us to reduce the number of preferences to $O(n)$ at each level.

We performed experiments to check for transitivity in the novelty task by looking at triplets of documents. A triplet of documents $\langle i, j, k\rangle$ is transitive if and only if i is preferred to j, j is preferred to k, and i is preferred to k. The ratio of number of triplets found to be transitive to the total number of triplets give a measure of transitivity in the preference judgments. On average transitivity holds for 98% across all queries with each query being transitive 96% of the time. This suggests that the assessors are highly consistent in their judgments; thus using a sorting algorithm with minimum information loss could further reduce the number of judgments required. It also suggests that whatever other features of documents (apart from topical relevance and novelty) the assessors are using in their decision process, they are consistent in their use of those features.

4.2.2 *How many levels of judgments are needed?*

In this section we show that the first two sets of preferences, i.e. experiments 1 and 2, are approximately sufficient to produce an optimal ranking. We again use preferences simulated from subtopic judgments: a relevance ranking is found by always preferring the document with more subtopics ("level 1"); a first approximation to a novelty ranking is found by always preferring the document with the most subtopics that are not in the top document ("level 2"); a second approximation by always preferring the document with the most subtopics that are not in the first two documents ("level 3"); and so on up to level 20.

Figure 3 shows the S-recall scores increasing as the number of preference sets increases. Clearly the increase in S-recall from level 1 to level 2 is the largest, nearly exceeding the total increase obtained from all subsequent levels put together. This suggests that the first approximation novelty ranking is likely to be sufficient; this has the benefit of reducing the amount of assessor effort needed to produce the data.

5. CONCLUSION AND FUTURE WORK

We have taken initial steps into investigating the use of preference judgments for novelty ranking tasks. We have proposed a novel framework for obtaining preference judgments for the novelty task and explicated the pros and cons of using preference judgments. Preliminary results for comparing explicit subtopic labels with preference judgments suggest that preference judgments can give similar information about both relevance and novelty as the subtopic judgments that are typically used.

Based on this, we proposed a preference-based approach to obtaining a full ranking for relevance, novelty, and all other factors that contribute to user preferences. We showed that rankings obtained in this way correlate well to rankings based on subtopic judgments, and since assessors are highly self-consistent, probably capturing a great deal of other information as well. Of course, if subtopic judgments were replaced with preferences, we would need a new set of evaluation measures. This clearly is a direction for future work.

Acknowledgments: This work was supported in part by the National Science Foundation (NSF) under grant number IIS-1017026. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] Amazon mechanical turk. `http://www.mturk.com`.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proc. of WSDM*, 2009.

[3] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. of SIGIR*, pages 433–440, 2005.

[4] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, number 2, pages 9–15, Nov. 2008.

[5] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *Proceedings of ECIR*, 2011.

[6] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of ECIR*, pages 16–27, 2008.

[7] P. Chandar and B. Carterette. What qualities do users prefer in diversity ranking. In *Proceedings of the 2nd Workshop on Diversity in Document Retrieval*, 2012.

[8] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results. *Information Retrieval*, pages 1–21, 2011.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceeding of CIKM*, pages 621–630, 2009.

[10] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659–666, 2008.

[11] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In *In 12th International Financial Cryptography and Data Security Conference*, pages 16–30. Springer-Verlag, 2008.

[12] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43:46–52, Dec 2009.

[13] M. E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590–598, 1990.

[14] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Y. Lin. Simple evaluation metrics for diversified search results. In *Proc. EVIA*, 2010.

[15] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of SIGIR*, pages 555–562, 2010.

[16] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR*, pages 315–323, 1998.

[17] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR*, 2003.