What Qualities Do Users Prefer in Diversity Rankings?

Praveen Chandar & Ben Carterette
Department of Computer & Information Sciences
University of Delaware
[pcr | carteret]@udel.edu

1. INTRODUCTION

Novelty and diversity ranking aims to provide individual users or groups of users with the documents that will cover a *space* of information needs or different aspects of a single information need [7, 11]. Most approaches to diversity evaluation require a list of subtopics that either disambiguate a short query or give further specification of aspects of the underlying information need. Documents are judged for relevance to these subtopics.

Evaluation measures for diversity use these subtopic judgments to evaluate a system's ranking. The basic idea is that stepping down a ranking of documents, each subsequent document that is relevant to a given subtopic should be worth less than the one before. This models decreasing value of seeing the same information. These measures are generally based on a few simple principles:

1. A document that is relevant to more unseen subtopics is worth more than a document that is relevant to fewer.

2. A document that is relevant to both unseen subtopics and already-seen (redundant) subtopics is worth more than a document that is only relevant to the same unseen subtopics.

3. A document that is relevant to unseen subtopics is worth more than a document that is only relevant to redundant subtopics.

To express it in a somewhat formal way,

$$\begin{aligned} \textit{novelty} + \textit{novelty} &\geq \textit{novelty} + \textit{redundancy} \\ &\geq \textit{novelty} \\ &\geq \textit{redundancy} \end{aligned}$$

Measures like α-nDCG [7], subtopic recall [11], and ERR-IA [5] can all be seen as using these inequalities, albeit with different magnitudes in the differences [6, 5, 9]. Our aim in this work is to formulate some simple hypotheses reflecting these assumptions and test whether they hold for actual assessors.

Radlinski et al. [8] make a distinction between *extrinsic diversity*—when a query is ambiguous and best served by documents relevant to several intents—and *intrinsic diversity*—when an information need is unambiguous but underspecified, and best served by documents reflecting different aspects of the information need. We will focus on the intrinsic case, which we also refer to as *novelty ranking*, because it is easier for assessors to understand how to evaluate relevance when there is no ambiguity of intent.

To test our hypotheses, we propose a novel user study based on *preference judgments* of the form "document A is preferred to document B for topic T" [4]. Sanderson et al. [10] have previously used preference judgments to suggest that user preferences of *rankings* of documents correlate well with the α-nDCG scores. We aim to test whether user preferences of individual documents correlate well with the principles listed above that α-nDCG is based on. In Section 2, we formally describe the hypotheses we mean to test. We describe the experimental design in Section 3, and analyze the results in Section 4. We conclude in Section 5.

2. USER PREFERENCE FOR INTRINSIC DIVERSITY

As described above, we model the presence of aspects of information about a topic in documents using subtopic judgments. We have distilled the chain of inequalities above to three simple hypotheses about novel information in a ranking, specifically about the content of a second document conditional on a given first document:

1. Given a document about a subtopic S_1, users will generally prefer to next see a document about a new subtopic S_2 to one about the same subtopic S_1.

2. Given a document about a subtopic S_1, users will generally prefer to next see a document about S_2 with some redundancy about S_1 to one about S_2 alone.

3. Given a document about a subtopic S_1, users will generally prefer to see a document about two new subtopics S_2, S_3 to a document about only one new subtopic S_2.

To investigate these hypotheses, we set up an experiment as follows: assessors would be shown three documents and told that one of them appears as the top document in a ranked list. They would be asked to choose which of the other two they would most like to see as the *second* document in that ranked list. By selecting documents that match the conditions of each hypothesis, we can get a sense of what type of novelty users most desire.

We will denote a document as D_i. We will assume we have relevance judgments to a topic, and for each relevant document, binary judgments of relevance to a set of subtopics.

DDR'12, Feb 12, 2011 Seattle WA, USA
ACM X-XXXXX-XX-X/XX/XX.

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.

Triplet $\langle D_2, D_3|D_1\rangle$ means the assessor is given a choice of D_2 or D_3 given D_1. Since we can represent documents by the set of subtopics they have been annotated with, we can write a triplet as the set of subtopics in the three documents. We will denote a preference between two documents using $\succ$, e.g. $D_1 \succ D_2$ means document D_1 is preferred to document D_2. Then the three hypotheses stated formally are:

H_1: if $\langle D_2, D_3|D_1\rangle = \langle\{S_2\}, \{S_1\}|\{S_1\}\rangle$, then $D_2 \succ D_3$ (novelty is better than redundancy)

H_2: if $\langle D_2, D_3|D_1\rangle = \langle\{S_1, S_2\}, \{S_2\}|\{S_1\}\rangle$, then $D_2 \succ D_3$ (novelty+redundancy is better than novelty alone)

H_3: if $\langle D_2, D_3|D_1\rangle = \langle\{S_2, S_3\}, \{S_2\}|\{S_1\}\rangle$, then $D_2 \succ D_3$ (novelty+novelty is better than novelty alone)

We don't expect that any of these would hold for every case (i.e. every matching triplet of documents for every query). For instance, the first may not hold if S_1 is deemed to be far more important to the information need than S_2; the second may not hold if D_1 is deemed to be "complete" for S_1. And there are of course many reasons a user might prefer one document to another; the presence of a novel subtopic is only one reason and may not be the most important.

Given this formal setup, in the next section we describe the actual experimental design we implemented to test it.

3. EXPERIMENTAL DESIGN

In order to test the above mentioned hypothesis we need two kinds of judgments; subtopic level judgments and conditional preference judgments. Subtopic level judgments were obtained from the data described in Section 3.1. Conditional judgments are necessary as the hypotheses are such that conclusions are made based on a condition i.e given a document D_1 which of the two document D_2 or D_3 is better. We decided to use crowdsourcing to collect user data for the conditional judgments as it is a fast and easy way to collect user judgments. Also prior work has shown that crowdsourcing platforms like Amazon Mechanical Turk (AMT) have several advantages such as low cost, fast turnaround, flexibility, etc [3].

Amazon Mechanical Turk (AMT) [1] is an online labor marketplace were workers complete a small task for a certain amount of money. The AMT system works as follows: A requester creates a group of *Human Intelligence Tasks* (HITs) with various constrains and workers from the marketplace who satisfy these constraints may work on these tasks to complete the task. In this section, we discuss the HITs in detail, which is basic unit of work performed a worker and the constrains associated with it. Additional details about AMT are available in the developer documentation.

3.1 Data

As discussed above, we do not believe these hypotheses would hold for queries with extrinsic diversity (such as those used by the TREC Web tracks), since for those types of queries a user usually has one intent in mind and the rest are not relevant. Thus we need data that models intrinsic diversity, i.e. a user has an unambiguous information need that can be represented by different aspects that appear in relevant documents. An example is "earthquakes" for a user that wants to find locations of recent earthquakes. If there had been earthquakes in Iran, Algeria, India, and Pakistan, and information about them appears in relevant documents, those would be the subtopics.

We have data reflecting this intrinsic diversity need. It is described by Allan et al. [2], from whom we obtained it. It consists of 60 topics, each of which has a keyword query, a description of an information need, and a list of subtopics identified by an assessor. For each topic, 130 documents were judged for topical relevance as well as for relevance to each of the subtopics. The corpus is a set of about 300,000 newswire articles originally part of the AQUAINT corpus.

3.2 HIT Design

Designing a HIT was by far the trickiest part of this user study. In this section, we discuss the variables associated with a HIT and the experimental settings used.

3.2.1 HIT Properties

A detailed description is necessary for the HIT in order to be identified by the workers. In general, workers use the AMT's web interface to search for a task to work on. Requesters set variables such as HIT Title, Keyword that aid workers to search for tasks that are more suitable to their skill set.

Title: A short description of the task to the workers. The title text is indexed, thus HITs could be searched by title. In our study, *"Document Preference"* was used as the title.

Description: A detailed explanation about the task. This gives workers a bit more information before they actually decide to preview a HIT. The workers can not search based on the description text. We used *"Read the document at the top and pick the document from the two documents shown below that gives most new information"* as the description.

Keyword: A set of keywords that will help workers search for HITs. The keywords used in our study includes *search, news articles, prefer, preference and opinion.*

Time allotted: AMT allows the requester to set a time limit within which a worker has to complete an accepted HIT. It is important not to rush workers into finishing their task. We set *three hours* as the limit to complete a HIT.

Pay: Workers are paid for each HIT they complete. Pay rate has obvious implications for attracting workers and incentivizing them to do quality work. Higher pay rates are more attractive to genuine workers but they also attract more spammers, therefore care must be taken while determining the pay rates. On the other hand lower pay rates could result in workers abandoning the task, therefore an appropriate amount needs to be picked. We paid *$0.80* for every HIT used in our study.

3.2.2 HIT Layout

The content of the HIT Design Layout is what a worker sees for a HIT. A common template consisting of various elements was used for all the HITs in the experiment and is shown in Figure 1. The various elements used in the template includes: a set of instructions about the task, the original keyword query, topic description, article texts (with query keywords highlighted), preference options for indicating which of the two documents the assessor prefers, and a comment field allowing them to provide feedback for that HIT. A brief description about each element is given below:

Guidelines: The worker was provided with a set of instructions and guidelines prior to judging. The guidelines specified that the worker 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. Some suggestions included in the guidelines were: one has more new information about the topic than the other; one has more focused new information about the topic than the other; one has more detailed new information than the other; one is easier to read than the other. The actual guidelines used are shown in Figure 2.

Query text and topic description: Each HIT consists of a *query text* field that describes the topic in a few words (we used the topic "titles" in the traditional TREC jargon) and a *topic description* field that provides more verbose and informative description about the topic, which are typically expressed in one or two sentences. Below is an example query text and topic description used.

> *Query Text:* John Kerry endorsement
> *Topic Explanation:* Documents containing information about individuals/groups that has endorsed or have announced their plan to endorse John Kerry's presidential primary bid are relevant.

Preference triplet: Figure 1 shows an example preference triplet with the query text and topic description. A HIT consisted of five preference triplets belonging to the same query shown one below the other. Each preference triplet consists of *three* documents, all of which were relevant to the topic. One document appeared at the top; this was a document chosen from the data described in Section 3.1 relevant to exactly one subtopic. The bottom two documents in the triplets were chosen randomly such that the hypothesis constraints were satisfied. For example, the documents in a triplet for hypothesis *H1* would contain the following subtopics in them: Top Document - S_1, Left Document - S_1, Right Document - S_2.

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. We did not show them any subtopics, nor did we ask them to try to determine subtopics and make a preference based on that. A comment field was provided at the end to provide a common feedback for all the five triplets, if they chose to do so.

3.2.3 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. A variety of methods such as *majority votes* can be used to determine the preferred document in each triplet. In our study, each HIT was judged by 5 different workers.

Trap questions: Triplets for which the answers were already known 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 document was not relevant to the topic. For the latter, the top document and one of the bottom two documents were the same. The assessors were expected to pick the non-identical document as it provides novel information. One of the five triplets in a HIT was a trap question and the type was chosen randomly.

Qualifications: It is possible to qualify workers before they are allowed to work on your HITs in Amazon Mechanical Turk. Qualifications can be determined based on historical performance of the worker such as percentage of approved HITs. Also, worker's qualification can be based on a short questionnaire or a test. A HIT could have multiple qualifications that a worker must satisfy in order to preview the HIT. A brief description of the two qualifications used in are study are explained below:

1. **Approval rate:** HITs can be restricted to workers with a minimum percentage of approval for their task. This method is a commonly used to improving accuracy and reducing spammer from working on your task. A worker required an overall approval rate of *95%* 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. By requiring workers to take a test, requester can illustrate the kind of response expected for a 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. Additionally, we had instructions to the workers for each triplet aiding them in making a preference, e.g. "prefer the document containing information not in the top document" for the identical traps and "prefer the document that is topically relevant" for the the non-relevant traps.

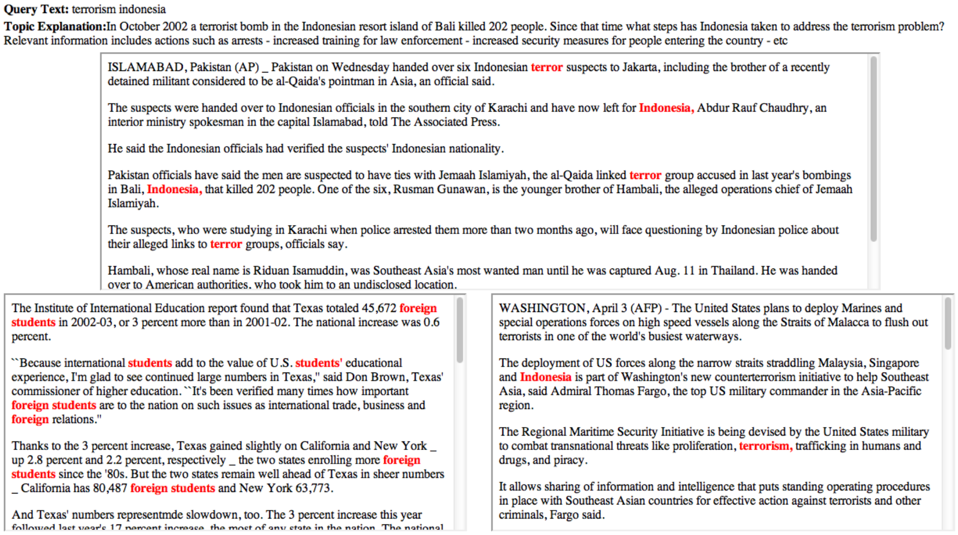
3.3 Topics and Triplets

We found six topics that had triplets of documents matching our hypotheses; two of these only matched H_1, two matched H_2 and H_3, and two matched all three hypotheses[1]. For each of these six topics we identified four different triplets for each matching hypothesis. Each of these triplets were assessed by five different workers. Thus we have a total of 20 preference judgments per topic per hypothesis, and after applying majority vote to each triplet, we have 4 preference judgments per topic per hypothesis. We chose six topics in our study to explore these hypotheses, and wish to extend the study to more topics in the future.

4. RESULTS AND ANALYSIS

Table 1 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

[1] Finding triplets that match the hypotheses exactly turned out to be more challenging than we expected



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 representmde 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

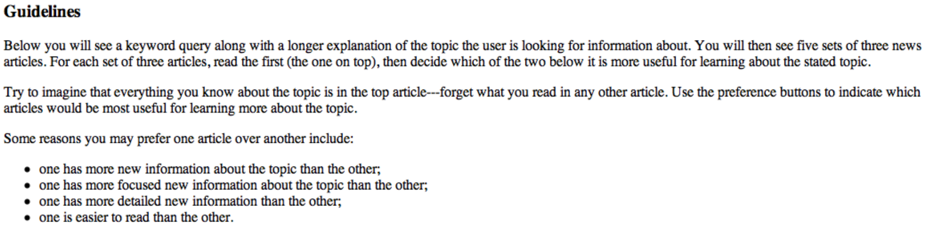
WASHINGTON, April 3 (AFP) - The United States plans to deploy Marines and special operations forces on high speed vessels along the Straits of Malacca to flush out terrorists in one of the world's busiest waterways.

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.



Guidelines

Below you will see a keyword query along with a longer explanation of the topic the user is looking for information about. You will then see five sets of three news articles. For each set of three articles, read the first (the one on top), then decide which of the two below it is more useful for learning about the stated topic.

Try to imagine that everything you know about the topic is in the top article---forget what you read in any other article. Use the preference buttons to indicate which articles would be most useful for learning more about the topic.

Some reasons you may prefer one article over another include:

- one has more new information about the topic than the other;
- one has more focused new information about the topic than the other;
- one has more detailed new information than the other;
- one is easier to read than the other.

Figure 2: Screenshot of the guidelines used in a HIT.

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 1: 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 2: 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.

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.

Table 2 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 case is particularly interesting because it is the opposite of what we would expect after seeing the results in Table 1: given that assessors preferred redundant documents to novel documents for that query, why would they now prefer novel documents to documents with both novelty and redundancy?

Table 3, 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

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 3: 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	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 4: Interassessor agreement scores for each topic.

queries for which the preference is weak.

Based on this, it seems we can conclude *novelty+novelty* $>$ *novelty*, *novelty* + *redundancy* $\geq$ *novelty*, but not *novelty* $\geq$ *redundancy*.

4.1 Interassessor Agreement

As described above, each triplet was judged by five different workers. We calculated an interassessor agreement score for each triplet 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). Table 4 shows the mean agreement for every triplet judged for each query. We had 106 unique workers judge 60 triplets across six unique queries. Five out of six queries featured in all the hypotheses but the query *weapons for urban fighting* featured only in H_1. The overall a high mean agreement of 0.7 was found across all triplets and the scores are close to the agreement observed previously [4].

4.2 Trap Questions

Interestingly, while assessors passed the "nonrelevant document" trap question almost perfectly, most of them did *not* pass the "identical document" trap: they actually indicated a preference for the identical document more often than not. While a single random guesser could easily get a nonrelevant trap right and an identical document trap wrong, it is highly unlikely (statistically) that every assessor would do this! Thus we conclude that there's some reason other than cheating for assessors preferring the identical document.

First, it may be another indication that users sometimes like redundant information. In our corpus, there are many examples of two articles on the same subject that are structured almost identically yet have slightly different information; an example is two articles about the same earthquake, with the more recent article containing more accurate information about the magnitude, source, and damage. Given this, it may be that assessors like an identical document because it confirms the information in the first document rather than contradicting or updating it.

Second, there are some possible confounding effects. Because the bottom two documents appear in narrower windows than the top document, it may not be immediately clear that one of them is identical. Furthermore, given up to 15 articles to read on a page of hits, assessors are likely skimming; since many documents are structured similarly but not identical, they may assume that two identical documents are actually just similar based on their skimming.

Finally, it may be that assessors simply did not understand the task. But as we discuss in Section 4.4 below, that does not seem to be the case; rather, it seems more likely that sometimes they just preferred documents for reasons other than the subtopics they contained.

In any case it seems that we can (at least tentatively) conclude that some redundancy is desirable, whether it be redundancy of content or redundancy of form and structure.

4.3 Possible Confounding Effects in Display

The way the hits were displayed may introduce some confounding effects, causing assessors to choose documents for reasons other than novelty or redundancy. In particular:

1. Sometimes the two documents have a large difference in lengths. Assessors may prefer the shorter just to avoid having to read more.

2. Assessors may prefer the document in which more query terms have been highlighted.

3. Assessors may even subconsciously normalize highlighted terms for document length and weight by document frequency, which we could check by looking at preferences due to some retrieval scoring function like language modeling.

We investigated each of these.

4.3.1 *Document length*

It seems that assessors did prefer shorter documents 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% of triplets. 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 (158 terms for H_1, 126 terms for H_2, and 47 terms for H_3). It therefore seems safe to conclude that assessors really do prefer shorter documents.

4.3.2 *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 documents. 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.

4.3.3 *Language model score*

There was only a slight preference by language model score (using linear smoothing), and it was a preference for documents with a lower score. For H_1, 51% of preferences were for the document with the higher score, but for H_2 and H_3 the preference was 44% and 41% respectively. Since these are not significant, it is unlikely that any interaction between length and query term occurrence had an effect on preferences.

4.4 Additional Investigation

The results from H_1 and the identical document trap were not what we expected; we thought there would be a much stronger preference for a novel document over a redundant document. 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 preferences, in particular recency, completeness, and value. It may also suggest that there are implicit subtopics (perhaps at finer levels of granularity) that the original assessors did not identify, but that make a difference in preferences.

None of this is surprising, of course, but there is currently no evaluation paradigm of note that can take all of these factors into account in a holistic way. Preference judgments can, and this analysis suggests additional hypotheses for testing with preferences.

5. CONCLUSIONS

We have described a user study to test some basic hypotheses about preferences for documents with novel and/or redundant information. The hypotheses we tested are implicitly assumed to be true by the use of subtopic judgments in diversity evaluation measures like α-nDCG. Based on the results of the study, we have reason to believe that users would generally prefer documents with *more* information to less (H_2 and H_3), but when asked to choose between two documents with the same amount of information (in terms of subtopic relevance), they seem to have no strong opinion.

From this we draw a few conclusions:

- The subtopic judgments we have may not completely accurately reflect all the aspects of the topics that users identify; in particular, there may be deeper levels of granularity that users use to distinguish between documents.
- Different users are interested in different aspects; there may not be one set of subtopics that can model the needs of all users.
- There are reasons for preferences other than novelty and redundancy; these reasons include recency, completeness, value, and perhaps ease of reading (as modeled by document length).
- Nevertheless, subtopic judgments seem to provide a pretty good model of user preferences.
- But the preferences themselves may be better.

Our analysis suggests more hypotheses we may want to investigate in the future. In addition, there are ways to extend the three hypotheses we present here. For example, we may want to investigate the effect of a subtopic that occurs in all three documents, e.g. a new $H_3' : \langle D_2, D_3 | D_1 \rangle = \langle \{S_2, S_4\}, \{S_2, S_3, S_4\} | \{S_1, S_4\} \rangle \Rightarrow D_3 \succ D_2$. With our formal way of stating the hypotheses, we can easily enumerate and investigate many different scenarios.

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 author's and do not necessarily reflect those of the sponsor.

6. REFERENCES

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

[2] James Allan, Ben Carterette, and Joshua Lewis. When will information retrieval be "good enough"? In *Proceedings of SIGIR*, pages 433–440, 2005.

[3] Omar Alonso, D.E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, number 2, pages 9–15. ACM, November 2008.

[4] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of ECIR*, pages 16–27, 2008.

[5] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, pages 1–21, May 2011.

[6] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of WSDM*, pages 75–84, 2011.

[7] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR '08*, pages 659–666, 2008.

[8] Filip Radlinski, Paul N. Bennett, Ben Carterette, and Thorsten Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43:46–52, December 2009.

[9] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of SIGIR*, pages 1043–1052, 2011.

[10] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of SIGIR*, pages 555–562, 2010.

[11] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR '03*, pages 10–17, 2003.