Inf Retrieval (2011) 14:89–106
DOI 10.1007/s10791-010-9157-1

THE SECOND INTERNATIONAL CONFERENCE ON THE THEORY OF INFORMATION RETRIEVAL (ICTIR2009)

An analysis of NP-completeness in novelty and diversity ranking

Ben Carterette

Received: 2 August 2010 / Accepted: 16 August 2010 / Published online: 14 December 2010

Abstract A useful ability for search engines is to be able to rank objects with *novelty* and *diversity*: the top k documents retrieved should cover possible intents of a query with some distribution, or should contain a diverse set of subtopics related to the user's information need, or contain nuggets of information with little redundancy. Evaluation measures have been introduced to measure the effectiveness of systems at this task, but these measures have worst-case NP-hard computation time. The primary consequence of this is that there is no ranking principle akin to the Probability Ranking Principle for document relevance that provides uniform instruction on how to rank documents for novelty and diversity. We use simulation to investigate the practical implications of this for optimization and evaluation of retrieval systems.

Keywords Evaluation · Test collections · Novelty · Diversity · Simulation · Theory

1 Introduction

There has recently been interest in designing retrieval systems to rank documents with novelty and diversity: the retrieved documents should cover some set of subtopics or cover different possible intents of a query (Agrawal et al. 2009; Vee et al. 2008; Clarke et al. 2008; Radlinski et al. 2008; Chen and Karger 2006; Zhai et al. 2008; Carbonell and Goldstein 1998; Carterette and Chandar 2009; Clarke et al. 2009a). Various evaluation measures have been proposed for this task: Zhai et al. (2008) introduced variations of recall and precision that count the number of unique subtopics retrieved, Clarke et al. (2008) introduced a "nugget"-based version of DCG that penalizes systems for retrieving redundant subtopics, and Agrawal et al. proposed "intent-aware" versions of classical measures that average those measures calculated with respect to particular intents. In theory these measures can be used for optimization as well. They are based on a Cranfield-like setting in which assessors have annotated documents not only on their relevance but

B. Carterette (✉)
Department of Computer and Info. Sciences, University of Delaware, Newark, DE, USA
e-mail: carteret@cis.udel.edu

also with respect to subtopics, intents, or nuggets. The system is rewarded for finding documents that contain subtopics or nuggets that have not previously been seen in higher-ranked documents.

These measures have something in common: the computation needed to understand them is NP-hard (Agrawal et al. 2009; Clarke et al. 2008; Zhai et al.2008). Let $\mathcal{S}$ be a set of subtopics, intents, nuggets, or facets related to a given query Q, and let $\mathcal{C}$ be a corpus of documents in which each document $\mathcal{D}$ contains zero or more elements of $\mathcal{S}$. Those that contain zero elements are nonrelevant. Apart from the intent-aware measures, those listed above are based on comparing a value calculated over subtopics retrieved up to some rank j to the maximum value that could have been retrieved at the same rank. Finding this maximumis generally an NP-hard problem. As a result, specific decisions made in the design of a novelty/diversity retrieval system may *appear* to lead to worse results by these measures even when those same decisions would actually *improve* the experience of the user the measures intend to model.

This paper is presented in two parts. The first considers the worst-case implications of optimizing to and evaluating with NP-hard effectiveness measures. The second uses simulations to draw conclusions about the implications in the average case.

2 Worst-case analysis

Let us first define our evaluation measures using the notation above, then show how each is NP-hard. For simplicity we will refer to elements of $\mathcal{S}$ as *subtopics*, though they need not literally be subtopics.

2.1 Evaluation measures

We consider four measures from the literature: S-recall and S-precision, α-nDCG, and intent-aware precision (prec-IA). Before defining them, let us follow Zhai et al. (2008) in defining $\textsc{minRank}(\mathcal{S}, k)$ as the size of the smallest subset of documents in $\mathcal{C}$ that could contain ("cover") at least k subtopics in $\mathcal{S}$.[1] We will use unadorned $\textsc{minRank}$ for the case where $k = |\mathcal{S}|$. We prove that computing $\textsc{minRank}$ is NP-complete in the "Appendix" (Theorem 1).

2.1.1 S-recall

S-recall at rank m is defined as the number of subtopics retrieved up to a given rank m divided by the total number of subtopics (size of $\mathcal{S}$) (Zhai et al. 2008):

$$\textit{S-recall@}m = \frac{|\cup_{i=1}^{m} \mathcal{D}_i|}{|\mathcal{S}|}.$$

Computing S-recall at an arbitrary m is polynomial time; we only need count the unique subtopics retrieved. But because $|\mathcal{S}|$ could vary greatly from topic to topic, it is useful to look at S-recall at rank $m = \textsc{minRank}(\mathcal{S}, |\mathcal{S}|)$. Analogously to R-precision, S-recall at $\textsc{minRank}$ has a minimum value of 0 and a maximum of 1 for every topic. It is, however, NP-complete as a consequence of $\textsc{minRank}$ being so.

[1] Note that while Zhai et al. defined this quantity in terms of a recall value, we define it in terms of the number of subtopics. The definitions are functionally equivalent.

2.1.2 S-precision

Zhai et al. (2008) defined S-precision at rank m as the ratio of the minimum rank at which a given recall value could optimally be achieved to the first rank at which the same recall value actually has been achieved. Let $k = |\cup_{i=1}^{m} \mathcal{D}_i|$. S-precision is then equivalent to $\text{MINRANK}(\mathcal{S}, k)$ divided by the first rank by which at least k unique subtopics have appeared.

$$S\text{-}precision@m = \frac{\text{MINRANK}(\mathcal{S}, k)}{m^*}, \quad \text{where } m^* = \arg\min_j |\cup_{i=1}^{j} \mathcal{D}_i| \geq k.$$

2.1.3 α-nDCG

Standard DCG calculates a gain for each document based on its relevance and a logarithmic discount for the rank it appears at (Jarvelin and Kekalainen 2002). The nugget version for diversity evaluation defines the gain of a document in terms of the subtopics (or nuggets) it contains and the frequency with which those subtopics appear in documents ranked above it (Clarke et al. 2008). The gain is incremented by 1 for each new subtopic, and α^k ($0 \leq \alpha \leq 1$) for a subtopic that has been seen k times in previously-ranked documents.

Since DCG is unbounded, it is standard to normalize it by the maximum possible value it could have given a perfect ranking of documents; this is called nDCG. In the case of α-DCG, determining that maximum appears to be NP-hard. Though the argument is not straightforward, we present a sketch in the "Appendix" (Conjecture 1).

Clarke et al. have also introduced a variant of α-nDCG called *novelty- and rank-biased precision* (NRBP) that is based on Moffat and Zobel's rank-biased precision (Moffat and Zobel 2008; Clarke et al. 2009b). Rather than use an exact normalization factor, it normalizes using an upper bound on the maximum possible NRBP calculated by assuming there is an "ideal" ranking in which every document contains every subtopic. Because of this, NRBP does not have well-defined range. Note that this is not necessarily detrimental (DCG does not have a well-defined range either); the practical question is whether it affects conclusions drawn from evaluation or whether it has any effect on the way we optimize system performance.

2.1.4 Intent-aware precision

Intent-aware precision (prec-IA) is calculated by first calculating precision for each distinct subtopic separately, then averaging these precisions according to some distribution indicating the proportion of users that are interested in that subtopic. Using the notation we defined above, this may be expressed as:

$$\begin{aligned} prec\text{-}IA@m &= \sum_{S \in \mathcal{S}} P(S|Q) prec_S@m \\ &= \sum_{S \in \mathcal{S}} P(S|Q) \frac{1}{m} \sum_{i=1}^{m} I(S \in \mathcal{D}_i) \end{aligned}$$

where $I(S \in \mathcal{D}_i)$ is 1 if and only if subtopic S appears in document $\mathcal{D}_i$, and $P(S|Q)$ is the probability that a user issuing query Q would be interested in subtopic S. Intent-aware measures do not penalize redundancy, but using a weighted average ensures that more desirable subtopics will influence the final value to a greater degree than less desirable subtopics.

Prec-IA is efficiently computable. Like α-DCG and NRBP, the maximum achievable value for a query is not necessarily 1.0, nor is it necessarily even clear what the maximum value is—it depends on the distribution of subtopics in documents (see Theorem 2 in the "Appendix"). However, a normalizing constant for prec-IA can be computed using a simple greedy algorithm (see Theorem 3 in the "Appendix"). Thus prec-IA is efficiently computable whether it is normalized or not.

We note that it is possible to define intent-aware versions of S-recall, S-precision, and α-nDCG. This might be valuable in cases where the query is ambiguous, so there are multiple possible intents, and each intent has its own set of subtopics or nuggets, creating a sort of hierarchy of subtopics. Clarke et al. (2009b) consider this case in their definition of intent-aware NRBP. For simplicity, we will focus on a single level of that hierarchy.

All of these measures have strengths; each contributes something unique to an overall understanding of performance. Our concern is not with the measures themselves, but with the cases at their boundaries: those topics for which we cannot properly evaluate or optimize systems because of the computational requirements. These cases cannot be averaged out; they will be a source of systemic error in our evaluations. Our goal is to begin to estimate how frequent such cases may be and what the implications of their existence are.

2.2 Approximability

An approximation algorithm is an efficiently-computable algorithm that gives an approximate solution to a hard problem. Approximation algorithms are typically evaluated by an *approximation ratio* expressed as the rate of growth of the ratio of the approximate solution to the optimal solution.

2.2.1 Evaluation

There is a simple greedy algorithm for calculating $\textsc{minRank}(\mathcal{S}, k)$ and the normalizing factor in α-nDCG: first take the document that contains the most subtopics, then the document that contains the most subtopics that have not already been taken, and so on until k subtopics have been covered. This greedy approach is in fact roughly the best approximation that can be achieved. As we show in the "Appendix", minRank is equivalent to Set Cover, and Feige showed that set cover is inapproximable within $(1-\epsilon)\ln|\mathcal{S}|$ for $\epsilon > 0$ unless NP has quasi-polynomial algorithms (Feige 1998). The greedy algorithm has approximation ratio H_s, where $s = \max_{S\in\mathcal{S}}|S|$ and $H_n = \sum_{i=1}^{n} 1/i$; the fact that $H_s \leq 1 + \ln s$ gives the result.

While the approximated minRank or normalizing factor can therefore be quite bad, the situation is somewhat better for the measures themselves. The measures exhibit *submodularity*, which means they can be approximated within a constant factor of $1 - 1/e$ (Agrawal et al. 2009). Intuitively, even if we are overestimating the denominator by a large factor, the fact that there is a limited number of subtopics means that the marginal error in the approximate value of S-recall or S-precision decreases as that factor increases.

2.2.2 Optimization

The optimization problem is to rank documents such that S-recall, S-precision, α-nDCG, or prec-IA are maximized. The standard principle for optimization in IR is the *Probability Ranking Principle*, which says that ranking documents in decreasing order of probability of

relevance gives the optimal expected precision and recall (and therefore R-precision and average precision and other such measures) (Robertson 1977). It can be extended to graded relevance to provide a ranking principle for DCG (Li et al. 2008). Either way, the PRP assumes that documents are relevant independently of one another, so it is not suitable for optimization of novelty or diversity rankings (Goffman 1964). Robertson illustrates this with an example of a query with two possible intents, showing that there is no PRP-based ranking that can uniformly satisfy both intents (Robertson 1977).

An optimization analog to the greedy algorithm for approximating evaluation measures is a greedy algorithm for ranking documents: given k ranked documents, the $(k+1)$st should be the one that is most likely to satisfy the greatest number of previously-unsatisfied subtopics (Agrawal et al. 2009; Clarke et al. 2008; Zhai et al.2008). However, unlike the PRP, which maximizes precision and recall at *every* rank, a greedy document-by-document ranking principle cannot necessarily provide maximum S-recall, S-precision, or α-nDCG at every rank. This follows from the NP-completeness of the evaluation problem; if it *were* possible to optimize at every rank, evaluation measures would be computable with the greedy algorithm. The worst case for optimization, then, is that the system is optimized at rank $1 + \log |\mathcal{S}|$ but not at any higher rank.

Intent-aware precision is an important exception. An expanded PRP that estimates the probability of relevance of a document to each subtopic would optimize prec-IA at every rank. This is because prec-IA, in contrast to the other measures, does not explicitly penalize redundancy. We explore the consequences of this below.

2.3 Example

Suppose there are 14 subtopics and 5 relevant documents (that is, five documents that contain at least one subtopic).[2] Documents contain subtopics as follows:

$$\mathcal{D}_1 = \{S_1, S_2\}$$
$$\mathcal{D}_2 = \{S_3, S_4, S_5, S_6\}$$
$$\mathcal{D}_3 = \{S_7, S_8, S_9, S_{10}, S_{11}, S_{12}, S_{13}, S_{14}\}$$
$$\mathcal{D}_4 = \{S_1, S_3, S_4, S_7, S_8, S_9, S_{10}\}$$
$$\mathcal{D}_5 = \{S_2, S_5, S_6, S_{11}, S_{12}, S_{13}, S_{14}\}$$

Let us consider each of our evaluation measures:

1. To calculate MINRANK, the greedy algorithm will take $\mathcal{D}_3$ followed by $\mathcal{D}_2$ followed by $\mathcal{D}_1$, resulting in S-recall being evaluated at rank 3. The optimal is at rank 2; $\mathcal{D}_4$ and $\mathcal{D}_5$ cover all 14 subtopics. The approximation ratio of MINRANK is therefore 3/2.
2. S-precision at any rank depends on being able to calculate $\text{MINRANK}(\mathcal{S}, k)$, where k is the number of unique subtopics observed to that rank. For $k = 7$ and $k = 8$, the greedy and optimal algorithms agree that $\text{MINRANK}(\mathcal{S}, 7) = \text{MINRANK}(\mathcal{S}, 8) = 1$. They also agree for $k = 12$ (the first two documents selected by the greedy algorithm): $\text{MINRANK}(\mathcal{S}, 12) = 2$. But for $k = 14$ (in the two documents selected by the optimal algorithm) there is disagreement. The greedy approach says $\text{MINRANK}(\mathcal{S}, 14) = 3$, while the optimal says $\text{MINRANK}(\mathcal{S}, 14) = 2$. This means that calculating $\text{MINRANK}(\mathcal{S}, 14)$ greedily for a system that place $\mathcal{D}_4, \mathcal{D}_5$ at ranks 1 and 2 will result in an S-precision of 3/2, which is greater than 1.

[2] This example is derived from Wikipedia's page on SET COVER (http://en.wikipedia.org/wiki/Set_cover_problem).

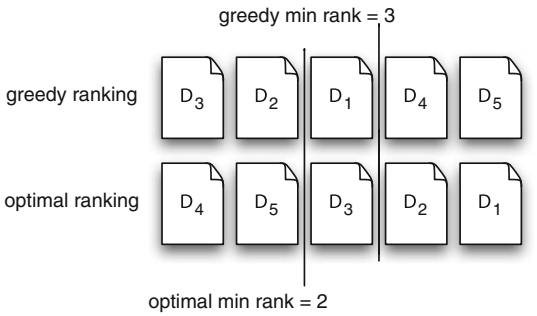

Fig. 1 A system that ranks documents greedily to optimize S-recall would place $\mathcal{D}_3$ above $\mathcal{D}_2$ above $\mathcal{D}_1$. A system that ranks documents greedily to optimize α-nDCG would place $\mathcal{D}_3$ above $\mathcal{D}_4$ and $\mathcal{D}_5$ (not shown). A system that optimizes S-recall at $\text{minRank}(\mathcal{S})$ would place $\mathcal{D}_4, \mathcal{D}_5$ at the first two positions. Using a greedy algorithm to determine $\text{minRank}(\mathcal{S})$ places it at rank 3; the true value is at rank 2

3. The normalizing factor for α-nDCG presents a problem in that the optimal set of documents over which it is computed can depend on the rank. At rank 1, the best possible α-DCG is achieved with $\mathcal{D}_3$ ($\alpha\text{-}DCG = 8/\log_2(2)$). But at rank 2, the best possible α-DCG is achieved with $\mathcal{D}_4, \mathcal{D}_5$ ($\alpha\text{-}DCG = 7/\log_2(2) + 7/\log_2(3)$). The optimal set at rank 1 is not a subset of the optimal set at rank 2, and therefore optimal α-nDCG at every rank is unachievable by any ranking algorithm.
4. Assuming $P(S|Q)$ is uniform, prec-IA is maximized by taking $\mathcal{D}_3$ first, then $\mathcal{D}_4$ and $\mathcal{D}_5$. Note that no matter what rank we look at, despite the fact that we can find the maximum value, prec-IA is rather far from 1: $prec\text{-}IA@1 = 0.57$, $prec\text{-}IA@2 = 0.54$, $prec\text{-}IA@3 = 0.52$.

Now let us consider how the two types of evaluation interact with greedy optimization versus optimizing for S-recall at minRank. We will assume the system has perfect knowledge of subtopics, and consider two cases:

1. a system optimizing S-recall/S-precision, greedily taking $\mathcal{D}_3, \mathcal{D}_2, \mathcal{D}_1$ followed by $\mathcal{D}_4, \mathcal{D}_5$ in any order to maximize the number of unique subtopics retrieved;
2. a system optimizing α-nDCG/prec-IA, greedily taking $\mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_2, \mathcal{D}_1$ to provide some redundancy along with new subtopics.

The first of these greedy approaches is illustrated in Fig. 1, along with the optimal ranking for S-recall at minRank and the minRanks calculated by greedy and optimal approaches.

Table 1 shows the complete set of evaluations for three systems: greedy systems with greedy evaluation; greedy systems with optimal evaluation; optimal system with greedy evaluation; and optimal system with optimal evaluation. Note that some of the values are greater than one for the optimal system evaluated greedily; this is because it is simply able to outperform any greedy algorithm.[3] Also note that the optimal system is uniformly outperformed at rank 1 by the greedy systems regardless of evaluation measure computation; this is because, as mentioned above, the document that is optimal at rank one ($\mathcal{D}_3$) is not a subset of the documents that are optimal at rank two ($\mathcal{D}_4, \mathcal{D}_5$). Since the system is restricted to choosing a document at rank 1 that *is* a subset of the documents at ranks 1 and 2, it cannot optimize at *both* ranks and therefore must suffer at one of them.

[3] A simple "hack" for this case might be to redefine S-precision and α-nDCG to have maximum values of 1, but this seems unfair to a system that uses alternatives to the greedy ranking.

Table 1 Greedy and optimal evaluations for two systems that rank documents greedily and a system that optimizes for S-recall at the minimum rank

		Greedy eval			Optimal eval		
		Rank 1	Rank 2	Rank 3	Rank 1	Rank 2	Rank 3
Greedy S-rec/S-prec $(\mathcal{D}_3, \mathcal{D}_2, \mathcal{D}_1, \mathcal{D}_4, \mathcal{D}_5)$	S-prec	1.000	1.000	1.000	1.000	1.000	0.667
	S-rec	0.571	0.857	1.000	0.571	0.857	1.000
	α-nDCG	1.000	0.943	0.844	1.000	0.922	0.844
	norm prec-IA	1.000	0.800	0.636	1.000	0.800	0.636
Greedy α-nDCG/prec-IA $(\mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_2, \mathcal{D}_1)$	S-prec	1.000	1.000	1.000	1.000	1.000	0.667
	S-rec	0.571	0.786	1.000	0.571	0.786	1.000
	α-nDCG	1.000	1.000	1.000	1.000	0.977	1.000
	Norm prec-IA	1.000	1.000	1.000	1.000	1.000	1.000
Optimal S-rec/S-prec $(\mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_3, \mathcal{D}_2, \mathcal{D}_1)$	S-prec	1.000	1.333	1.333	1.000	1.000	1.000
	S-rec	0.500	1.000	1.000	0.500	1.000	1.000
	α-nDCG	0.875	1.023	0.983	0.875	1.000	0.983
	Norm prec-IA	0.875	0.933	1.000	0.875	0.933	1.000

Prec-IA is normalized by its maximum achievable value

The α-nDCG case is particularly interesting. We calculated α-nDCG with $\alpha = 1/2$, i.e. the second time a subtopic appears it contributes 1/2 to the document's gain, the third time it contributes 1/4, and so on. The system that optimizes S-recall therefore has incentive to go on to find the second-best set of documents and rank them second, thereby achieving an α-nDCG greater than 1 at rank 2 with the greedy evaluation. The greedy system evaluated optimally, on the other hand, sees a decrease in nDCG despite continuing to find novel subtopics; this is because it could have retrieved all 14 unique subtopics at rank 2, and 14 unique subtopics plus 8 redundant subtopics at rank 3.

Normalized prec-IA exhibits the most extreme behavior. For the system that greedily optimizes prec-IA, it is 1 throughout. For the system that greedily optimizes S-recall, prec-IA decreases with rank. For the system that optimizes true S-recall, prec-IA increases with rank. There is no other measure that produces such wide differences in behavior, and for that reason we question the applicability of prec-IA to this task.

NRBP is not shown in the table, since it is not calculated at individual ranks but rather calculated for an entire ranking. It also has no NP-complete component, because it uses an efficiently-computable upper bound to normalize. It provides an additional interesting case, however: the system that optimizes for S-recall at minRank has an NRBP of 0.711, while the greedy S-rec/S-prec system has an NRBP of 0.673 and the greedy α-nDCG/prec-IA system has an NRBP of 0.713. All three values are fairly far from 1.0, though there is no ranking that provides a higher NRBP. Like α-nDCG, NRBP will prefer a greedy system, though it comes closer to recognizing that a greedy ranking is not the only possible approach.

The table shows that for optimization there is a firmly imposed tradeoff. When optimizing for S-recall at minRank, it is impossible to achieve perfect S-recall, S-precision, α-nDCG, and prec-IA at rank 1. When optimizing greedily for S-recall/S-precision or α-nDCG/prec-IA at each rank, it is impossible to achieve perfect S-recall at minRank. In standard retrieval problems founded on the PRP, there is an empirical tradeoff between precision and recall, but it is theoretically possible to optimize for both. For these measures there may be topics for which that is theoretically impossible; the developer is forced to choose.

This example can be generalized. If $|\mathcal{S}| = 2^{k+1} - 2$ and there are k relevant documents that are pairwise disjoint and $\mathcal{D}_i$ contains $2i$ subtopics, and there are two additional relevant documents that are disjoint and that each contain one half of each $\mathcal{D}_i$, the approximation ratio for MINRANK is $O(k/2)$. As k increases, the greedily-computed S-recall for a greedy system is 1, but the true S-recall is $(2^k + 2^{k-1})/(2^{k+1} - 2)$, which goes to 3/4. Note that this is a constant approximation ratio for S-recall despite the logarithmic approximation ratio for MINRANK. This is due to the submodularity of S-recall (Agrawal et al. 2009).

3 Simulation and analysis

While worst-case analysis shows that it is possible to construct cases in which the evaluation and optimization fail, the practical question is whether such cases occur in real data, and if so, how often and to what extent they affect evaluation and optimization. Having only a small sample of subtopic queries to analyze and no theory regarding the distribution of subtopics in documents, we cannot make definitive statements. But we can run simulations of the type done in average-case complexity studies (Bogdanov and Trevisan 2006).

We report results exclusively for S-recall at MINRANK. S-recall is slightly simpler than S-precision and α-nDCG because it involves no parameters and is always between 0 and 1. The general conclusions hold regardless of measure.

3.1 Real data

There is little annotated data available for studying these problems. Currently two large sets exist. The first was constructed by Allan et al. (2005) for a report-writing task with a newswire corpus.[4] It comprises a set of 60 topics with about 13,000 document-level relevance judgments as well as labeled "aspects" for each relevant document. "Aspects" are defined as individually distinct pieces of relevant information. For instance, the first query is "oil producing nations" and its relevant aspects are *Algeria, Angola, Azerbaijan, Bahrain, Brazil, Cameroon, Chad, China, ...* Each document is labeled as to whether it is relevant to each of the topic's aspects. The aspects were defined by the assessors themselves during the course of judging. If while judging their 10th document they discovered it contained an aspect that had not been in any of the first nine, they added it to the list of aspects for that topic. There was no limit on the number of aspects they could define; the average for a topic is 22, but two topics have over 100.

The second was assembled by NIST for the diversity task for the TREC 2009 Web track. It comprises 50 topics with about 28,000 document-level relevance judgments to web pages, with each page judged for relevance with respect to predefined subtopics (of which there were at most eight) (Clarke et al. 2009a). Subtopics largely reflect different information needs or intents of the query. For example, the query "kcs" has two subtopics relating to the Kansas City Southern railroad, two relating to two separate school districts, and one relating to an energy company.

We obtained these datasets to use as starting points. For the Allan et al. data, we treat aspects as subtopics. We consider each subtopic to be equally valuable to the user, so this problem is somewhat different from the diversity problems of Agrawal et al. and others that model a users' interest in particular subtopics. The Web track data is closer to that

[4] The link to the data provided in this reference no longer works, but the data can be obtained by contacting the author of this work.

Table 2 Examples of topics from the Allan et al. data

Topic no.	Query	# Subtopics	# Relevant docs
5	Ohio highway shootings	33	52
	Near I-270, near Columbus, a house, a freeway interchange, ...		
7	Greenspan testimony congress	8	75
	Wed. Feb 11 2004, Thu. Feb 12 2004, Tue. Feb 24 2004, Apr 2004, ...		
18	Haiti protest	7	48
	Port-au-Prince, Montreal, St. Marc, Raboteau, Gonaives, ...		
48	Reduce dependence oil	17	12
	Nuclear energy, shift to biodiesel, invest in hydrogen, ...		

Though we have not shown the topic descriptions, the reader can probably infer it from the query and the listed subtopics

Table 3 Examples of topics from the TREC 2009 Web track data

Topic no.	Query	# Subtopics	# Relevant docs
1	Obama family tree	3	93
	TIME photo essay; heritage of Obama's forebears; bio of Obama's mother		
10	Cheap internet	8	124
	Low-cost providers; dial-up providers; Vonage homepage; ...		
24	Diversity	4	117
	Workplace diversity; diversity training programs; cultural diversity; ...		
33	Elliptical trainer	4	142
	Reviews; sources of used trainers; relative benefits; ...		

Subtopics are paraphrased from the full TREC subtopic questions

diversity problem in that the subtopics are much more clearly delineated between documents. This emerges clearly when looking at the average number of subtopics documents are relevant to: for the former set, each document contains 2.7 subtopics on average; for the latter, each document contains an average of only 1.2.

Tables 2 and 3 show some example topics from the two sets along with their subtopics. Table 2 illustrates a task in which there is a clearly-defined information need, and to answer that need a system must retrieve as many unique aspects as possible. Table 3 illustrates a task in which there are multiple possible needs, and the system must be useful to users who have any of them (proportionately).

Among these 110 topics, there are seven that are trivial (two from Allan et al.; five from Web): they have only one relevant document, only one subtopic, or one relevant document that covers all the subtopics. We have excluded these. Additionally, there are 34 that are quasi-trivial (27 (46.5%) from Allan et al.; 7 (15%) from Web); in these, some subtopics only appear in one relevant document each, and taking those documents (and in some cases one additional document) covers the set trivially. There are seven topics for which the greedy algorithm overestimates the true minRank, with four from Allan et al. and three from Web. Therefore, 7 out of 103 non-trivial topics (6.8%) and 10% of non-quasi-trivial topics can have performance overestimated by the greedy algorithm.

3.2 Simulated topics

Starting from real topics, we simulate new topics by sampling from a space defined by the marginal distributions of subtopics within documents. Specifically, each topic can be written as a matrix T with documents on the rows, subtopics on the columns, and $T_{ij} = 1$ if document i is relevant to subtopic j or $T_{ij} = 0$ otherwise. An example is shown in Table 4. We will simulate topics by sampling uniformly at random from the space of 0–1 matrices that have the same row sums and column sums as the initial topic matrix. This ensures that even if we cannot precisely model the distribution of subtopics in documents, we can at least model the numbers of subtopics contained in each document and the number of documents each subtopic appears in.

The sampling algorithm is based on a random walk procedure described by Zaman and Simberloff (2002). It is used in ecological studies for statistical testing of hypotheses about distributions of species in regions. It is based on the observation that within a larger matrix T, a 2×2 diagonal matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ can be changed to an anti-diagonal matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ (and vice versa) without altering the row or column sums. The algorithm works by sampling two rows and two columns uniformly at random, and if the 2×2 matrix formed from the cells at their intersections is diagonal or anti-diagonal, changing it to an anti-diagonal or diagonal matrix (respectively). Over many iterations this randomizes the distribution of subtopics in documents while keeping the marginal sums constant.

The algorithm requires a "burn-in" period to sufficiently randomize the original matrix. After that, a large enough number of sampling iterations ensures a uniform distribution over all possible matrices with the same row and column sums as the original. We used a burn-in period of 10,000 iterations, with 1,000 additional samples from the burned-in matrix to generate random topics. Thus for any given topic, we could generate a new random topic by iterating 1,000 times starting from the burned-in matrix for that topic.

We have limited our simulations to start from the Allan et al. data. Because there are substantially more subtopics, and the variance in the number of subtopics is higher, this data provides somewhat more interesting results.

3.2.1 Results

Results on simulated topics are based on evaluating a greedy system with perfect knowledge of subtopic containment. This is because the worst case for a system without

Table 4 Part of the document-subtopic matrix for topic 18 "haiti protest"

	Port-au-Prince	Montreal	St. Marc	Cap-Haïtien	Gonaives	Raboteau	Petionville	Sum
$\mathcal{D}_1$	1	0	0	0	0	0	0	1
$\mathcal{D}_2$	1	0	0	0	0	0	0	1
$\mathcal{D}_3$	0	1	0	0	0	0	0	1
$\mathcal{D}_4$	0	0	0	1	0	0	0	1
$\mathcal{D}_5$	0	0	1	0	1	1	0	3
$\mathcal{D}_6$	1	0	1	0	1	0	0	3
…				…				…
Sum	34	1	5	1	16	3	1	61

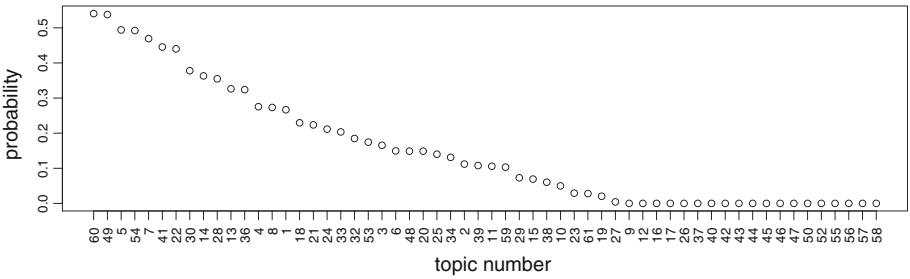

Fig. 2 Proportion of matrices sampled from the space defined by each of the baseline topics with MINRANK approximation ratio greater than 1

perfect knowledge is arbitrarily bad: if such a system did not retrieve any relevant documents in the top $j =$ OPTIMAL-MINRANK, but it retrieved relevant documents at the following ranks up to $j =$ GREEDY-MINRANK, its S-recall approximation ratio goes to infinity. We consider simulated imperfect systems in the next section.

First we investigated the probability that the greedy algorithm for MINRANK would overestimate the minimum rank. Figure 2 shows the proportion of sampled matrices starting from each actual topic for which the true minimum rank (found by exhaustive search[5]) was less than the greedy minimum rank. Note for some topics the probability is very high: for topic 60, over half the randomly sampled matrices were suboptimal.

There were 19 topics (roughly one third) for which the greedy and true minimum rank matched in every sample. Overall, the greedy algorithm overestimated MINRANK for about 15% of sampled topics, which is a little higher than would be likely if the rate of 4 every 60 that was observed in the data is true.

Next we investigated the average MINRANK approximation ratio for the cases for which the greedy algorithm was suboptimal. Figure 3 shows the results for the 39 topics that were not always greedy-optimal. Topic 7 is the worst, with an average approximation ratio nearly 1.5 (minimum 1; maximum 1.667; median 1.333). Over all sampled topics, the mean approximation ratio is 1.16. The greedy is never more than 4 greater than the optimal, suggesting cases like our example above (worst case $\log |\mathcal{S}|$) are not occurring.

Finally we looked at the factor by which S-recall was overestimated when the rank was overestimated. Again, S-recall can only be overestimated by a constant $1 - 1/e$. Figure 4 shows that the average worst case is about 1.16 times the true value. The maximum factor by which any S-recall is overestimated is 1.33, which happens to be the reciprocal of the 3/4 approximation ratio derived in our example above.

3.3 Simulated systems

As discussed above, the worst case for a system with perfect knowledge of subtopics is that S-recall is overestimated by a constant factor. The worst case for a system with no knowledge of subtopics (i.e. one that makes use of heuristics such as similarities between documents) is arbitrarily bad. Between these two extremes, we are interested in the cases of systems that use heuristics but that "look like" real systems might.

[5] Though this is a relatively small data set, exhaustive search still took a very long time in the most extreme cases, even when parallelized across 64 cores.

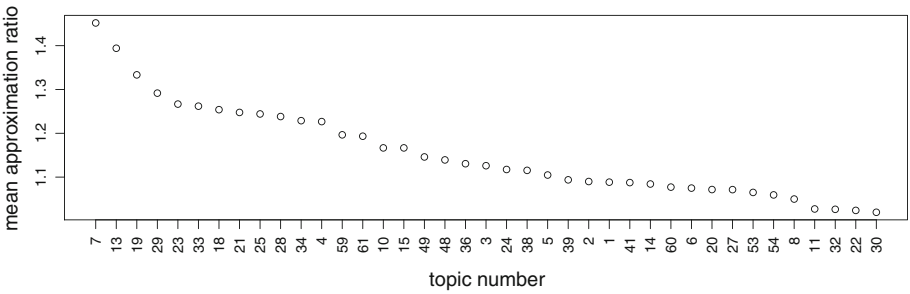

Fig. 3 Average MINRANK approximation ratio when greedy algorithm is suboptimal. Queries for which the greedy algorithm is always optimal not shown

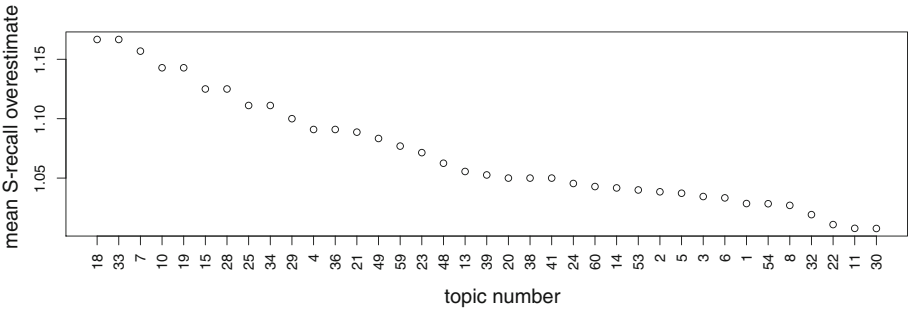

Fig. 4 Average factor by which S-recall is overestimated when greedy algorithm is suboptimal. Queries for which the greedy algorithm is always optimal not shown

We simulated a "real" system that uses a greedy optimization approach as follows: starting with a document-subtopic matrix, we degraded it by changing each 1 indicating the presence of a subtopic i in a document j to a probability p_{ij} drawn from a Beta prior with parameters α_p, β_p. We changed each 0 indicating the absence of subtopic i in document j to a probability q_{ij} drawn from a Beta prior with parameters α_q, β_q. We then applied a greedy algorithm similar to Agrawal et al.'s (2009) IA-SELECT, which attempts to rank the documents that are most likely to satisfy previously-unsatisfied subtopics. The resulting ranked list is evaluated using S-recall.

The Beta distribution parameters $\alpha_p, \beta_p, \alpha_q, \beta_q$ offer some control over the expected quality of the simulated system:

- As $\alpha_p/(\alpha_p + \beta_p) \to 1$ and $\alpha_q/(\alpha_q + \beta_q) \to 0$, the system approaches the best possible performance.
- As $\alpha_p/(\alpha_p + \beta_p) \to 0$ and $\alpha_q/(\alpha_q + \beta_q) \to 1$, the system approaches the worst possible performance.
- When $\alpha_p/(\alpha_p + \beta_p) = \alpha_q/(\alpha_q + \beta_q)$, the system is ranking documents randomly.

To keep the parameter space manageable, we used $\alpha_p = \beta_q$ and $\alpha_q = \beta_p$, increasing α_p and α_q exponentially from 2^0 to 2^7. For large α_p and small α_q, the system is better; for small α_p and large α_q, the system is worse. At $\alpha_p = \alpha_q$ the performance is random.

3.3.1 Results

We selected topics for which the greedy algorithms were suboptimal on either the burned-in matrix or the original matrix. We then degraded the matrix randomly and greedily re-ranked the documents according to the procedure above.[6] We then calculated S-recall both greedily and optimally.

Figure 5 compares the mean performance measured by the greedy evaluation to the S-recall approximation ratio for topics 5 and 7, starting from their burned-in matrices. Each point is the result of averaging over 100 trials with a particular α_p, α_q. Note that as simulated system performance degrades, we actually overestimate its performance more! This is quite disturbing, as it means that when the greedy evaluation is suboptimal, it will overestimate a bad system's performance more than a good system's performance. Bad systems will always appear better than they really are by a greater factor than good systems will.

The degree of overestimation is worse for topic 7 than for topic 5. This is because the optimal minimum rank for topic 7 is 3 (greedy is 4), while the optimal minimum rank for topic 5 is 16 (greedy is 18). With a deeper rank required for evaluation, the system has less opportunity to "catch up" after passing the optimal rank. However, topic 5 has five outlying points with very high approximation ratios. These are all points where α_q is substantially higher than α_p, meaning the system is a priori poor.

Figure 6 shows similar results starting from the original matrices for topics 18 and 30. Like topic 7, topic 18 has low optimal ranks (optimal 4 vs. greedy 5). Like topic 5, topic 30 has high optimal ranks (optimal 53 vs. greedy 52).

4 Discussion and conclusion

We have argued that NP-complete evaluation and optimization can be a serious problem for retrieval systems. Even if the approximation ratio is constant, we can significantly overestimate the performance of a system. In particular, the worse a system is, the more likely its performance is to be overestimated. These errors are not random errors that can be averaged out by sampling more topics; they are systemic problems with evaluation and optimization in this setting.

Furthermore, there will always some topics for which it is theoretically impossible to optimize measures for every rank. As Fig. 1 and Table 1 illustrate, and as the NP-completeness of the computation implies, for some topics the optimal set of documents of size k is not a proper subset of the optimal set of size $k + 1$. This poses a problem for a system that is expected to rank things; it must choose just one of those ranks at which to try to optimize, because there is no consistent way to optimize for both. It is also impossible to optimize all measures simultaneously; there are firm theoretical tradeoffs in choosing to optimize for S-recall versus α-nDCG. The implication is that novelty and diversity systems must have a very clear idea of the user's task in order to provide the best possible experience.

There are other concerns about these measures as well. For one, it is possible to "game" them in a way that is not possible with traditional document-level relevance-based measures: a dishonest researcher or developer can simply introduce a new document that is a concatenation of the entire corpus. This new document will contain every subtopic and

[6] We did not do an optimal ranking, since there are too many documents to be able to do exhaustive search over all subsets.

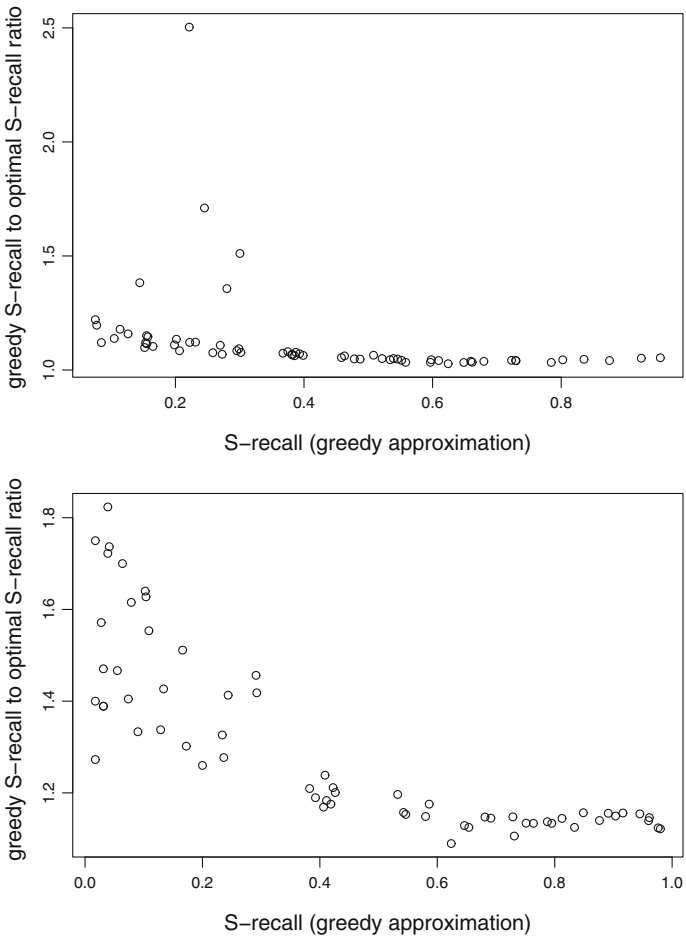

Fig. 5 Comparison of greedy S-recall to S-recall approximation ratio for topic 5 (left) and topic 7 (right) starting from burned-in matrices. Each point represents a different pair of prior parameters (α_p, α_q) and is averaged over 100 random trials

therefore will provide the maximum value for any of these measures, though it is clearly not useful to a user. For another, the cognitive load on assessors is much higher, as they must judge each document with respect to each subtopic. This introduces much more variance in the judgments than is present in standard relevance judgments.

Assuming the measure is well-chosen for the task, for most topics there is no problem. The greedy algorithm is optimal in 93% of the cases in "real" data, and in about 85% of cases in simulated data. The problem cases are those for which the greedy algorithm is not optimal, in particular those for which a bad system is significantly overrated by the greedy algorithm. Future work may investigate characterizing the problematic topics so that results may be adjusted appropriately, though when considering additional problems described above, more fruitful work may lie in investigation of alternative representations of interdependent document relevance.

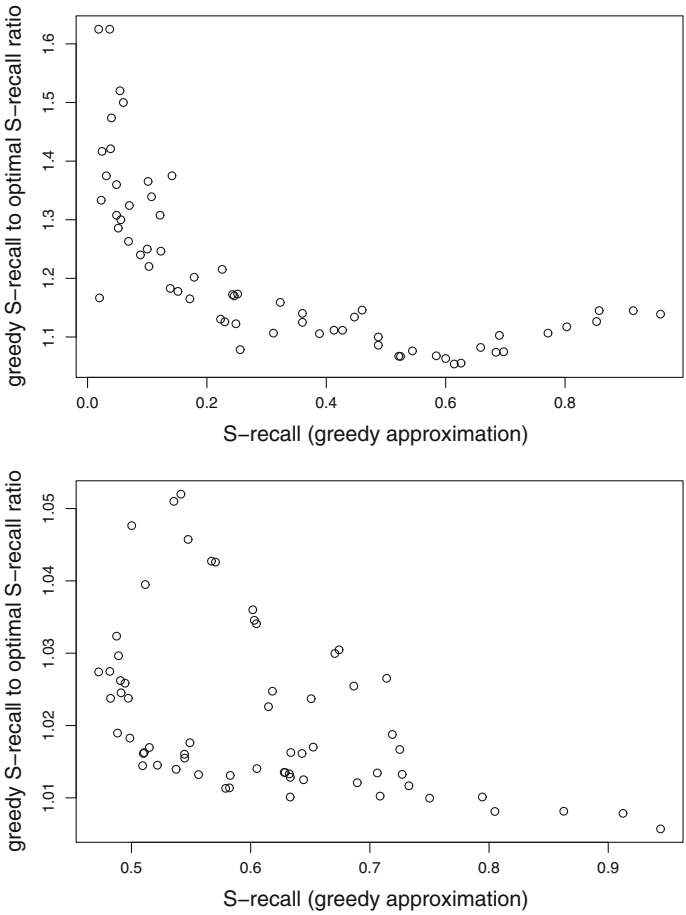

Fig. 6 Comparison of greedy S-recall to S-recall approximation ratio for topic 18 (*left*) and topic 30 (*right*) starting from original matrices. Each point represents a different pair of prior parameters (α_p, α_q) and is averaged over 100 random trials

Appendix

Here we present proofs or sketches of proofs for some of our claims regarding worst-case complexity. We use $\mathcal{S}$ to represent a set of subtopics for a query Q, $\mathcal{C}$ to represent a set of documents in which each document $\mathcal{D} \in \mathcal{C}$ is a subset of $\mathcal{S}$, and $\mathcal{R}$ to represent a subset of documents in a ranking.

Our first result establishes the hardness of S-recall at minRank.

Lemma 1 *S-recall at rank* $k = \text{minRank}(\mathcal{S}, |\mathcal{S}|)$ *is NP-hard.*

We prove this by reducing $\text{minRank}(\mathcal{S}, |\mathcal{S}|)$ from Set Cover.

Proof Let $I = (\mathcal{U}, \mathcal{V})$ be an instance of Set Cover in which $\mathcal{U}$ is a universe of items and $\mathcal{V}$ is a family of subsets of $\mathcal{U}$. The optimization problem is to find a subfamily $\mathcal{W} \subseteq \mathcal{V}$ such that the union of subsets in $\mathcal{W}$ is equal to $\mathcal{U}$ and $|\mathcal{W}|$ has minimum size among all such subfamilies.

To transform this into an instance of minRank, define the set of subtopics $\mathcal{S}$ to be $\mathcal{U}$ and the set of documents $\mathcal{C}$ to be $\mathcal{V}$. A solution to minRank is the size of a subset of documents $\mathcal{R} \subseteq \mathcal{C}$ such that all subtopics in $\mathcal{S}$ are contained in the union of documents in $\mathcal{R}$. Note that this is exactly equivalent to the size of the minimum covering subfamily $\mathcal{W}$ in the definition of Set Cover. Since Set Cover is NP-complete, it follows that minRank is NP-hard. □

Two assumptions are useful in establishing a lower bound on computational complexity:

1. $\mathcal{C}$ is a cover of $\mathcal{S}$, i.e. each subtopic appears in at least one document in the corpus;
2. the size of $\mathcal{C}$ is in $O(|\mathcal{S}|^c)$ for some constant c, i.e. the number of documents in a corpus grows at worst polynomially with the number of subtopics.

We claim the first assumption is reasonable for practical reasons: if there were some subtopic that did not appear in any document, it would not be useful for evaluating systems since there is no possibility for any system to retrieve it. Therefore it would be removed from consideration. For the second, we believe that support is lent by well-attested conjectures that the relationship between vocabulary size and corpus size is sublinear (such as Zipf's Law (1949).

Theorem 1 *S-recall at rank* $k = \text{minRank}(\mathcal{S}, |\mathcal{S}|)$ *is NP-complete under assumptions 1 and 2.*

Proof We know minRank is NP-hard from Lemma 1. It remains to be shown that it is in NP.

Let us define an algorithm for finding minRank iterate from $k = 1$ to $|\mathcal{C}|$, and for each k determine whether there is a set cover of $\mathcal{U} = \mathcal{S}$ in $\mathcal{V} = \mathcal{C}$ of size k. The first assumption ensures that the algorithm will halt: no matter how large $\mathcal{C}$ is, eventually a cover will be found. The second ensures that it iterates a polynomial number of times and therefore is not outside of NP.

Thus if the two assumptions above hold, this algorithm is in NP, and therefore minRank is NP-complete. □

Next we argue that computing the normalization constant for α-nDCG@k is NP-hard. The normalization constant is the maximum possible value of α-DCG@k.

Conjecture 1 *Maximizing α-DCG@k with respect to orderings of documents is NP-hard.*

Proving this is actually quite difficult; since the order of the documents affects the calculation, it cannot be easily reduced from graph-based problems like Set Cover. We make an informal argument by showing that the problem can be viewed as an instance of a resource-constrained scheduling problem with no precedence constraints. In such problems, there are sets of m resources, n jobs, and k discrete time periods. Jobs request resources for consumption. Resources are limited (when a jobs is using a resource, it cannot be used by other jobs), and in some cases they are non-renewable (that is, there is a fixed quantity that decreases each time a job uses it). The problem is to schedule jobs to maximize some utility function.

We transform that problem to the problem of maximizing α-DCG by mapping resources to subtopics, jobs to documents, and time periods to ranks. Documents "consume" the subtopics they contain to create utility; the amount of utility created is a function of the reciprocal of the log rank and the "amount" of the subtopic remaining after having been consumed by other documents. Specifically, the document scheduled at rank j consumes $100(1 - \alpha)$ percent of whatever remains of the subtopic after the documents at ranks 1 through $j - 1$ have consumed

it. The problem is to "schedule" documents to ranks such that utility is maximized. This scheduling problem is NP-hard for general utility functions (Lenstra et al. 1977), and therefore unless there is a clever algorithm that takes advantage of the particular form of α-DCG's utility function, we claim it is NP-hard to maximize α-DCG. Note, however, that if $\alpha = 1$, the greedy algorithm is optimal—the measure would be equivalent to DCG with relevance grade defined as the number of unique subtopics in the document. It is the non-renewable/redundancy-penalizing property of the measure that makes the problem hard.

The next result establishes that the upper bound of *prec-IA@k* is associated with the distribution of subtopics in documents.

Theorem 2 *prec-IA@k has a maximum value of one if and only if there are at least k documents that contain every subtopic.*

Proof If there are k documents that contain every subtopic, the maximum value of *prec-IA@k* is achieved by ranking those k documents. The precision of each subtopic $prec_S@k$ is one, and therefore the weighted average of precisions is one.

To show the converse, we will use contradiction. Suppose there are at most $k-1$ documents that contain every subtopic. Then there must be at least one document in the top k that does not contain at least one subtopic. Let us denote that subtopic S_m. Since it is not contained in every document in the top k, $prec_{S_m}@k \leq \frac{k-1}{k} < 1$. Assuming $P(S_m|Q) > 0$, it follows that the weighted average *prec-IA@k* < 1. Therefore there must be at least k documents that contain every subtopic if *prec-IA@k* is to have a maximum value of 1. □

This is of course analogous to a similar result for binary-relevance precision, which says that precision at rank k has a maximum value of one if and only if $k \leq |R|$ (the number of relevant documents). But this is a more important result for two reasons: first, it is unusual that there would be many documents that contain every subtopic. Second, if such documents exist, precision-IA *rewards* a system for finding more of them. This is contrary to what every other measure of novelty or diversity does.

Finally we show that the greedy algorithm is optimal for calculating the maximum value of *prec-IA@k*.

Theorem 3 *Ranking documents in decreasing order of* $d_i = \sum_{S \in \mathcal{D}_i} P(S|Q)$ *is optimal for calculating the maximum value of prec-IA@k.*

Proof Let $p_k = \frac{1}{k}\sum_{i=1}^{k} d_i = \frac{1}{k}\sum_{i=1}^{k}\sum_{S\in\mathcal{D}_i} P(S|Q)$, where $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have been numbered according to the greedy algorithm. Suppose *prec-IA@k* $> p_k$. Then there is some alternative ordering $\pi(n)$ such that:

$$\sum_{S\in\mathcal{S}} P(S|Q) prec_S@k > \frac{1}{k}\sum_{i=1}^{k}\sum_{S\in\mathcal{D}_i} P(S|Q)$$

$$\frac{1}{k}\sum_{S\in\mathcal{S}} P(S|Q)\sum_{i=1}^{k} I(S \in \mathcal{D}_{\pi(i)}) > \frac{1}{k}\sum_{i=1}^{k}\sum_{S\in\mathcal{D}_i} P(S|Q)$$

$$\sum_{i=1}^{k}\sum_{S\in\mathcal{D}_{\pi(i)}} P(S|Q) > \sum_{i=1}^{k}\sum_{S\in\mathcal{D}_i} P(S|Q)$$

$$\sum_{i=1}^{k} d_{\pi(i)} > \sum_{i=1}^{k} d_i$$

This shows that *prec-IA@k* can be formulated as a sum of values $d_{\pi(i)}$. Thus if *prec-IA@k* $> p_k$, then there must be at least one document $\mathcal{D}_n$ such that $d_{\pi(n)} > d_n$. But if such a document exists, it would have been selected by the greedy algorithm. Therefore, *prec-IA@k* cannot be greater than p_k, and therefore the greedy algorithm is optimal. □

References

Agrawal, R., Gollapudi, S., Halverson, H., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM international conference on web search and data mining* (pp. 5–14).

Allan, J., Carterette, B., & Lewis, J. (2005). When will information retrieval be 'good enough?'. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 433–440).

Bogdanov, A., & Trevisan, L. (2006). *Average-case complexity*. Hanover, MA: Now Publishers Inc.

Carbonell, J. G., & Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 335–336).

Carterette, B., & Chandar, P. (2009). Probabilistic models of novel document rankings for faceted topic retrieval. In *Proceedings of the 18th ACM international conference on information and knowledge management*.

Chen, H., & Karger, D. R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 429–436).

Clarke, C. L., Craswell, N., & Soboroff, I. (2009a). Overview of the TREC 2009 web track. In *Proceedings of the 18th text retrieval conference (TREC)*.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., et al. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 659–666).

Clarke, C. L., Kolla, M., & Vechtomova, O. (2009b). An effectiveness measure for ambiguous and underspecified queries. In *Advances in information retrieval theory: Proceedings of the 2nd international conference on the theory of information retrieval* (pp. 188–199).

Feige, U. (1998). A threshold of ln *n* for approximating set cover. *Journal of the ACM, 45*(4),634–652.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval, 2*(3), 201–203.

Jarvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information and Systems, 20*(4), 422–446.

Lenstra, J. K., Kan, A. H. G. R., & Brucker, P. (1977). Complexity of machine scheduling problems. In P. L. Hammer (Ed.), *Studies in integer programming* (Vol. 1). North Holland: Addison-Wesley.

Li, P., Burges, C. J., & Wu, Q. (2008). McRank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems* (Vol. 20, pp. 897–904).

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information and Systems, 27*, 1–27.

Radlinski, F., Kleinberg, R., & Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on machine learning* (pp. 784–791).

Robertson, S. E. (1977). The probability ranking principle in information retrieval. *Journal of Documentation, 33*, 294–304.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., & Amer-Yahia, S. (2008). Efficient computation of diverse query results. In *Proceedings of the 24th international conference on data engineering* (pp. 228–236).

Zaman, A., & Simberloff, D. (2002). Random binary matrices in biogeographical ecology—instituting a good neighbor policy. *Environmental and Ecological Statistics, 9*, 405–421.

Zhai, C., Cohen, W. W., & Lafferty, J. D. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 10–17).

Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.