

A Query-Basis Approach to Parametrizing Novelty-Biased Cumulative Gain

Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose

School of Computing Science, University of Glasgow
Glasgow, G12 8RZ, United Kingdom
{kimm,guido,jj}@dcs.gla.ac.uk

Abstract. Novelty-biased cumulative gain (α -NDCG) has become the *de facto* measure within the information retrieval (IR) community for evaluating retrieval systems in the context of sub-topic retrieval. Setting the incorrect value of parameter α in α -NDCG prevents the measure from behaving as desired in particular circumstances. In fact, when α is set according to common practice¹ (i.e. $\alpha = 0.5$), the measure favours systems that promote redundant relevant sub-topics rather than provide novel relevant ones. Recognising this characteristic of the measure is important because it affects the comparison and the ranking of retrieval systems. We propose an approach to overcome this problem by defining a safe threshold for the value of α on a query basis. Moreover, we study its impact on system rankings through a comprehensive simulation.

Keywords: diversity, sub-topic retrieval, effectiveness measure, web search

1 Introduction

The purpose of an IR system is to respond to a given query with relevant documents so as to satisfy a information need. Nevertheless, queries posed by users are often inherently ambiguous and/or under-specified. Presenting redundant information may also be undesirable as users have to endure examining duplicate information repeatedly. Therefore, the IR system should present documents covering a complete combination of possible query-intents, in order to maximise the probability of retrieving relevant information (i.e. “provide complete coverage for a query”²). Such intents address several sub-topics of the information need and so they should be all retrieved; consequently, there is a need to avoid redundantly repeating them in the document ranking (i.e. “avoid excessive redundancy”²).

2 Analysis of α -NDCG

Clarke et. al. [1] proposed a modified version of normalised discounted cumulative gain, called α -NDCG, for evaluating novelty and diversity in search results. Information needs are represented with respect to a query as sets of nuggets, or sub-topics. Consider a query Q with a total of $|S| > 1$ sub-topics. Let $J(d_r, s)$ be a graded relevant judgement indicating whether a document d_r at rank r is

¹ See <http://plg.uwaterloo.ca/~treweb/2010.html> guidelines.

² Quote extracted from the TREC 2009 and 2010 Web Diversity Tracks guidelines.

Table 1. Five documents relevant to the sub-topics of query 26, “lower heart rate”, from the TREC 2009 Web Diversity Track (*Left*), and corresponding evaluations of three imaginary system rankings, when $\alpha=0.5$ (*Right*).

Document ID	Sub-topic				Total	system									
	1	2	3	4		r	doc	g(r)	ng(r)	dng(r)	dcng(r)	α -ndcg(r)	s-r(r)		
a. “en0001-55-27315”	1	-	1	1	3	system	A	1	a	3	<u>3.0</u>	3.0	3.0	1.0	0.75
b. “en0004-47-03622”	-	1	-	-	1		A	2	c	3	<u>1.5</u>	0.9	3.9	1.0	0.75
c. “en0001-69-19695”	1	-	1	1	3		A	3	e	0	<u>0.0</u>	0.0	3.9	0.9	0.75
d. “en0003-94-18489”	-	-	1	1	2		B	1	a	3	<u>3.0</u>	3.0	3.0	1.0	0.75
e. “en0000-31-13205”	-	-	-	-	0		B	2	d	2	<u>1.0</u>	0.6	3.6	0.9	0.75
							B	3	e	0	<u>0.0</u>	0.0	3.6	0.8	0.75
							C	1	a	3	<u>3.0</u>	3.0	3.0	1.0	0.75
							C	2	b	1	<u>1.0</u>	0.6	3.6	0.9	1.00
							C	3	e	0	<u>0.0</u>	0.0	3.6	0.8	1.00

relevant to a sub-topic s or not. A duplication measure³ $D_{s,r-1}$ is defined to monitor the degree of redundancy of documents ranked above r , given a sub-topic s . The measure has the role of quantifying the benefit of a document in a ranking, or what we call *novelty-biased gain*, $NG(Q, r)$:

$$NG(Q, r) = \sum_{s=1}^{|S|} J(d_r, s)(1 - \alpha)^{D_{s,r-1}} \quad (1)$$

where the parameter $0 < \alpha \leq 1$ represents the probability that a user is less likely to be interested in the sub-topic that is redundantly repeated by the document. In practice, this parameter is used to manipulate the reward of a document carrying novel information. To account for the late arrival of documents containing relevant sub-topics, the gain is discounted by a function of the rank position and then progressively cumulated⁴. The discounted cumulative gain at rank r is then normalised by that of the optimal ranking.

Table 1 (Left) shows five documents relevant to (some of) four sub-topics of query 26 belonging to the TREC 2009 Web Diversity Track. For the purpose of showing how an incorrect setting of α affects α -NDCG, we illustrate three imaginary system rankings (A, B, C), where the top three documents are ranked differently. In Table 1 (Right), the first column shows the rank position, (r), followed by document id, (doc), and the gain, $g(r)$, wrt. sub-topic relevance. The next columns are the novelty-biased gain, $ng(r)$, discounted novelty-biased gain, $dng(r)$, discounted cumulative novelty-biased gain, $dcng(r)$, its normalised gain, α -ndcg(r) when $\alpha=0.5$, and finally sub-topic recall [3], $s-r(r)$. Note that, while $a-b-c-d-e$ is an ideal ordering of the documents, setting α to 0.5 produces a maximal gain, resulting in the *false* ideal document ranking $a-c-b-d-e$, which in turn is used when normalising, as shown in the table. If systems are evaluated according to α -NDCG with $\alpha=0.5$, the following system rankings are obtained: $\{A, B, C\}$ or $\{A, C, B\}$. Note that system C obtains a lower α -NDCG than system A at positions 2 and 3 although at rank 2 it covers the only missing sub-topic (26.2), thus achieving complete sub-topic coverage (i.e. $s-r(2)=1.0$) earlier than A . In these circumstances α -NDCG with $\alpha=0.5$ rewards documents containing *novel* relevant sub-topics *less* than *redundant* ones.

³ $D_{s,r-1} = \begin{cases} \sum_{i=1}^{r-1} J(d_i, s) & \text{if } r > 1 \\ 0 & \text{if } r = 1 \end{cases}$

⁴ $DCNG(r) = \sum_{i=1}^r NG(Q, i) / \log_2(1 + i)$

3 Deriving a Threshold for α

We consider the case where the gain obtained by a system retrieving novel relevant sub-topics, say system X , is expected to be higher than the gain of a system retrieving only redundant sub-topics, say system Y .

Let s^* be a novel relevant sub-topic⁵, and s a redundant relevant sub-topic. At rank position r , in the worst case scenario (i.e. when system X retrieves only a single *novel* relevant sub-topic whereas system Y retrieves the remainder $|S|-1$ relevant but *redundant* sub-topics) system X should have higher α -NDCG than system Y . Thus, since we expect $NG_X(r) > NG_Y(r)$, we can rewrite this as:

$$J(d_r, s^*) \cdot (1 - \alpha)^{D_{s^*, r-1}} > \sum_s^{|S|-1} J(d_r, s) \cdot (1 - \alpha)^{D_{s, r-1}}$$

This inequality can be used to define boundaries on α so that the inequality is true, i.e. a system retrieving novel relevant sub-topics is awarded with an higher α -NDCG than a system retrieving redundant sub-topics. At this stage we make a simplifying assumption, following the relevance judgements that have been collected in the TREC Web Diversity track: we assume a binary decision schema regarding the relevance of documents to each sub-topic. Therefore:

$$(1 - \alpha)^{D_{s^*, r-1}} > \sum_s^{|S|-1} (1 - \alpha)^{D_{s, r-1}}$$

and with a further assumption that the $D_{s, r-1}$ of all redundant relevant sub-topics are identical, we obtain

$$(1 - \alpha)^{D_{s^*, r-1}} > (|S| - 1) \cdot (1 - \alpha)^{D_{s, r-1}}$$

Let $\beta = D_{s, r-1} - D_{s^*, r-1}$ be the difference in redundancy⁶. Note that β is always an integer when relevance judgements are binary. Thus, we can resolve wrt. α , ignoring the case $\alpha < 1 + \left(\frac{1}{|S|-1}\right)^{1/\beta}$, as $\alpha < 1$ by definition:

$$(st) : \quad \alpha > 1 - \left(\frac{1}{|S| - 1}\right)^{1/\beta} \quad (2)$$

Eq (2) is the necessary and sufficient condition that has to be satisfied if we expect α -NDCG to reward systems retrieving novel relevant sub-topics more than systems retrieving redundant sub-topics. Figure 1 shows the safe threshold (*st*) on α according to Eq (2) for varying circumstances, suggesting that considering values of α *below* or *equal to* the threshold (inside highlighted areas) can lead to an unexpected behaviour of the measure. That is, if $\alpha=0.5$ for all the information needs, α -NDCG may misjudge documents conveying novel information. This is crucial, in particular, when analysing *high quality*⁷ ranking results @2, @3, etc., or when the redundancy difference of the rankings (β) at lower positions is small. For queries containing 2 or less sub-topics this problem does not occur, as $\alpha=0.5$ is greater than the safe threshold.

⁵ Or the sub-topic with smaller degree of redundancy.

⁶ Measuring a relative amount of novel information in documents, where redundant sub-topics have higher degree of redundancy than novel sub-topics, i.e. $D_{s, r-1} > D_{s^*, r-1}$, and thus $\beta > 0$.

⁷ i.e. when a high number of relevant documents are ranked within the early ranking positions.

Fig. 1. Values of the safe threshold for α .

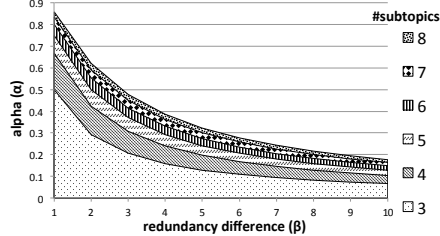
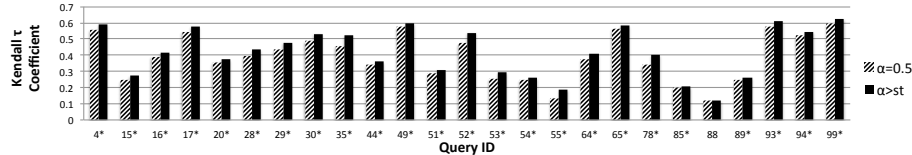


Fig. 2. Kendall’s τ coefficient for 25 queries wrt. sub-topic recall and α -NDCG for $\alpha=0.5$ and $\alpha > st$.



4 Remarks and Conclusion

To verify the impact of setting α according to the proposed threshold, we conducted a comprehensive study by simulating system rankings using relevance judgements from the TREC 2009-2010 Web Diversity Tracks. We used the Fisher-Yates shuffle algorithm (with 100 re-starts) to generate $6!$ (factorial) random samples of all possible permutations of relevant document rankings. Rankings were then evaluated according to sub-topic recall and α -NDCG @10 with $\alpha=0.5$, and $\alpha=st+0.01$ where $\beta=1$ to avoid possible undesired scenarios. Figure 2 presents the Kendall’s τ correlation of system rankings between sub-topic recall and the two different settings of α for 25 example queries (out of 98 total queries). Setting $\alpha > st$ produces rankings of systems based on α -NDCG that are significantly⁸ more correlated to the ones obtained using sub-topic recall⁹ than those obtained with $\alpha=0.5$. These results are consistent over all the query set. Although the use of correlation with sub-topic recall as a mean to assess whether a measurement is better than another might be criticised, we believe that this can provide an indication of the measure behaviour, in particular because the intent of the Diversity Tracks is to provide complete coverage of all sub-topics.

In summary, by setting α on a query basis according to the safe threshold of Eq (2), the diversity of document rankings can be correctly measured without recurring to further modify α -NDCG, as suggested in [2].

References

1. C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Bütcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666, 2008.
2. T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Lin. Simple Evaluation Metrics for Diversified Search Results. In *EVIA '10*, pages 42–50, 2010.
3. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.

⁸ Measured by a 1 tail t-test ($p < 0.01$) and indicated by * in Figure 2.

⁹ Correlation pairs are relatively low. This is because once complete sub-topic coverage is achieved at position r , the value of sub-topic recall for ranks $> r$ is always 1. Therefore, sub-topic recall is unable to measure the utility of ranking in such circumstances.