Balancing Aspects in Retrieved Search Results

ABSTRACT

Many queries contain explicit aspects which must be balanced in any retrieved result in order to meet a user's information need: if aspects of the query are missing or disproportionately represented in documents, the results will be of lower quality than desired. This balancing thus needs to occur both within the retrieved documents individually and across the entire set. We introduce the concept of query-aspect balance and describe a new evaluation measure, β-NDCG, that allows the evaluation of query-aspect balance on multivalued query-aspect judgments. We apply β-NDCG to a small test collection and explore its utility. We show that β-NDCG captures problems of query aspect balance within and across documents in the ranked list.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

Keywords

effectiveness, models

1. INTRODUCTION

In many retrieval tasks, it is important to ensure that the entirety of a query is covered by the search results. The different information needs, or aspects, presented in a query are not independent of each other and documents in which they co-occur are more likely to be valuable to the user. If a user requests documents relating to "sports on the beach", it stands to reason that the system should not only return a set of documents that are as a whole evenly balanced between information about "sports" and "beaches" but also each individual document should be well-balanced between these two aspects of the query. The user did not likely intend to get a set which is half documents about sports and half documents about beaches, rather they were seeking documents in which the two information needs occurred together. We address that issue, proposing a new way of evaluating retrieved results based on this notion of query-aspect balance. The measure that we propose, based on the family of NDCG measures, is called β-NDCG.

2. RELATED WORK

There is an extensive line of research into identifying and quantifying query components under the banner of subtopics, also referred to as nuggets, categories or instances. Early work on the problem by Carbonell and Goldstein [2] proposed the maximal marginal relevance method for selecting novel documents while also maintaining relevance to the original query. Clarke et al.[5] codified a framework which distinguished between novelty and diversity and proposed α-NDCG as a measure for comparing ranked lists on the basis of those two factors in the case of subtopic retrieval. Chapelle et al.[3] proposed expected reciprocal rank (ERR)as a means of addressing short comings in traditional relevance measures. Although ERR does not measure diversity, it does capture the notion that the shape of the density of the information retrieved matters. Craswell et al. [6] proposed a cascade model for measuring ranked lists. Argawal et al. [1] propose a series of intent-aware versions of classical retrieval measures. Clarke et al. [4] compared intent-aware and cascade models.

All of those methods and measures are intended to ensure that the ranked list somewhat evenly represents the subtopics within the retrieved set. The aspects of the query itself are not considered except to the extent that documents covering all aspects seem more likely to be relevant and these methods operate on the start of a ranked list. In contrast, this study explicitly incorporates the query's aspects into an evaluation measure and "rewards" ranking methods that correctly balance them.

3. ASPECT COVERAGE

An aspect is an explicit idea expressed within a query. For example, in the query "sports on the beach" there is a "sports" aspect and a "beach" aspect. Intuitively, a relevant document will include both of those aspects. More generally, a document will have a higher degree of relevance the more query aspects it includes. Moreover, a document that balances those aspects, not unduly covering one sub-topic at the expense of others, should be preferred because it evenly "covers" the query. For that reason, we assume that document relevance judgments for aspects are non-binary, supporting a range of possible degrees of relevance. A perfectly balanced ranked list of documents $d_1 ... d_n$ for a query q with

aspects $a_1 ... a_i$ is one in which the relevance of the documents to the aspects is evenly distributed both within each document and within the entire list.

Based on that idea, a measure for aspect coverage should have the following properties:

- *Relevance:* All other properties being identical, it should reward documents that maximize total relevance to the query aspects.
- *Internal Balance:* It should reward documents that have equal coverage of aspects, minimizing the variance across aspects mentioned. For example, a document that covers all aspects equally has zero balance, whereas one that covers one aspect well and others poorly will have high variance. We note that documents retrieved in response to a query will usually have *some* coverage of each query aspect, but they may not be well balanced.
- *List Balance:* It should reward documents that cover aspects under-represented in the previously seen documents. Note that internal balance is still important when selecting documents to improve list balance.
- *Multi-valued:* It should function on aspect judgments with multiple levels of relevance so that it is possible to recognize disproportionate over- or under-representation of aspects.

We assume the classic ad-hoc retrieval setting, where the user is issuing a query for which there is not an exact answer, but rather is looking for a range of material on the subject. For example, a user might be looking for a survey of information related to the use of support vector machines in classifying the authorship of scanned documents and enter the query "SVM for scanned document authorship classification." Although some documents are more relevant to this query than others, it is not posed or intended as a question that can be completely answered by any single document.

4. POSSIBLE MEASURES

There are several commonly-used measures that initially appear to be appropriate for this evaluation task. We contend that although they consider similar issues, they have significant weaknesses that make them inappropriate for the task of measuring aspect coverage.

The first measure we consider is α-NDCG [5]. In its original form, we have the following gain G at rank k:

$$G_k = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_i, k-1}$$

where $J(d_k, i)$ represents the judgment of the document at rank k for topic i and $r_i, k-1$ is the number of documents from ranks 1 to $k-1$ in which topic i occurs. When judgments are binary as they are assumed to be in α-NDCG, then the only way to increase the gain in a document is to cover more topics. Thus a document can only be high scoring if it covers many topics. However, because we desire the *multi-valued* property, we must incorporate non-binary judgments. In that case, by α-NDCG, a document can be high scoring because it heavily represents a single topic which violates the *internal balance* property: we want documents that have multiple aspects, and would rather see a balanced document that talks a little about an under-discussed topic then an unbalanced document that talks only about an undiscussed topic. To adapt α-NDCG to handle multi-valued judgments, we change $r_i, k-1$ to $s_i, k-1$, the sum of the judgments for aspect i on all documents up to rank $k-1$. Unfortunately, the result is that $(1-\alpha)^{r_i, k-1}$ rapidly becomes so small that the gain at any rank after the topmost is very small as well. We have been unable to find a way to handle this issue well. We further note that α-NDCG does not explicitly honor the *internal balance* property because it considers topics independently.

These observations are not surprising: α-NDCG was designed to evaluate the diversity and novelty of a ranked list relative to a set of binary-valued topic judgments [5], where novelty is "the need to avoid redundancy" and diversity is "the need to resolve ambiguity." There are two key differences between our task and the one envisioned for α-NDCG [5], and these inform the differences between the two measures. The first is that aspects, unlike topics, are explicit in the query. Thus we are concerned with the aforementioned issue of balance as opposed to diversity, as we are not trying to resolve ambiguity in the query. The second difference is that we do not assume that our aspects are given binary judgments. For aspects, unlike topics, binary judgments would be fairly uninformative, as most results will likely be at least relevant on some minor level to most aspects since they are explicitly in the query and most retrieval systems will retrieve results that reference all of the terms in the query. The issue is that there is a big difference between a document about beaches that mentions volleyball once and a document about beach volleyball even though it would be perfectly reasonable to argue that both documents would be relevant on a binary scale to both beaches and sports.

The cascade model [6] assumes that the quality of prior documents is important to the value of the current document, however it is predicated on the notion that the user's information need is finite and thus attempts to calculate the rank at which the user is satisfied. It also does not support any notion of query aspects. Thus measures based on the cascade model, such as ERR do could not be used.

Intent-Aware NDCG (NDCG-IA) [1] calculates a separate NDCG score for each category and then aggregates them with the score for each category weighted in proportion to the distribution of the categories for that query. Since each aspect's score is calculated independently of the others it neither rewards documents that cover aspects under-represented in the previously seen documents (i.e., that satisfy the *list balance* goal) nor does it reward documents that minimize aspect variance (i.e., satisfy the *internal balance* property). Since all intent-aware measures likewise calculate the score for each category independently, they cannot serve as the basis for our measure.

5. β-NDCG

As our score should be calculated on multi-valued judgments, the widely-used NDCG family of measures is a logical starting point for designing a measure capable of integrating query-aspect balance. This measures have a cumulative gain function that is discounted as the ranked list is traversed and provide a score normalization against an ideal ranking.

The document gain functions in the original NDCG and its α-NDCG variant do not support our desired objectives, so we will replace them with a new function. Based on our previously specified requirements, the gain for a document should include three components. The first is a positive function of the combined value of the judgment scores for the aspects. The second is a penalty for having unequal coverage of aspects. The third is a penalty for covering aspects that are over-represented in previously seen documents. All of these components should function on multi-valued relevance judgments, not just of documents (as does NDCG) but on aspects (unlike α-NDCG). We start with a skeleton in which the gain for a document is the sum of its aspect judgments, down-weighted by two functions: f_1 is a function of the aspect's relative magnitude and alpha, which is a parameter indicating our preference for internal aspect balance; and f_2 is a function of the variance of the aspects in the document and β, which is an indicator of our preference for list aspect balance.

$$G_k = \sum_{i=1}^{m} J(d_k, i) f_1(i, \alpha) f_2(\sigma_{d_k}, \beta) \quad (1)$$

Since we want f_1 to be negatively correlated with the *proportion* of the previously seen aspects accounted for by aspect i, we set it to be equal to:

$$f_1 = 1 - \alpha \frac{\sum_{h=1}^{k-1} J(d_h, i)}{\sum_{h=1}^{k-1} J(d_h)} \quad (2)$$

where $J(d_h)$ is the sum of the judgment scores for the aspects of the document at rank h. The penalty will thus increase as the proportion increases with the penalty being zero if we have never seen this aspect before and one if this aspect accounts for the entirety of the previously seen aspects.

Since we want f_2 to be negatively correlated with the variance of the document aspects, we set it to:

$$f_2 = \frac{1}{1 + \beta \sigma_k} \quad (3)$$

A document with perfectly even balance will have a variance of zero and thus no penalty, while there is a growing penalty as variance increases. Inserting these two equations into equation (1) gives us

$$G_k = \sum_{i=1}^{m} J(d_k, i) \left(1 - \alpha \frac{\sum_{h=1}^{k-1} J(d_h, i)}{\sum_{h=1}^{k-1} J(d_h)}\right) \frac{1}{1 + \beta \sigma_k} \quad (4)$$

For both α and β, setting the parameter to zero will completely remove any penalty from their respective functions. Setting both to zero gives us NDCG with the gain from each document being simply the sum of its aspect judgments.

We call this measure β-NDCG to reflect its purpose of evaluating the "balance" of aspects within and across documents in a ranked list.

6. EXPERIMENTS

We applied β-NDCG to a set of 21 queries from the TREC Web Track years 2009, 2010 and 2011. The queries were randomly selected from the 50 longest queries in those three years. For each of these queries we created judgments for manually selected aspects. 25 queries were initially selected, however four of the queries were not judged as there were no identifiable aspects. Queries not used included "to be or not to be that is the question" and "all men are created equal." Judgments were assigned on a zero to three scale, where zero was "not relevant" and three was "extremely relevant." The results were generated using the query-likelihood model implemented in the Galago search engine.

Table 1: "TREC Web Track Queries β-NDCG@10"

Query	α=0,β=0	α=0,β=1	α=1,β=0	α=1,β=1
2	0.807	0.794	0.765	0.764
16	0.537	0.504	0.528	0.494
18	0.893	0.762	0.912	0.796
44	0.939	0.852	0.946	0.877
53	0.827	0.671	0.427	0.345
105	0.564	0.564	0.667	0.667
106	0.660	0.652	1.000	1.000
109	0.725	0.598	0.698	0.569
112	0.893	0.893	0.667	0.667
116	0.937	0.936	0.947	0.932
117	0.606	0.606	0.609	0.609
119	0.728	0.728	0.333	0.333
122	0.948	0.872	0.866	0.750
123	0.963	0.905	0.934	0.812
125	0.965	0.882	0.962	0.901
129	0.938	0.859	0.929	0.870
136	0.392	0.338	0.520	0.437
139	0.881	0.665	0.879	0.630
143	0.929	0.745	0.569	0.373
146	0.897	0.912	0.930	0.883
148	0.740	0.492	0.579	0.369
149	0.933	1.000	0.908	0.977
MEAN	0.805	0.738	0.753	0.684

We applied β-NDCG using several different parameter settings. Applying β-NDCG with α and β set to zero is equivalent to NDCG with the document relevance scores being equal to the sum of the aspect scores. β set to zero while varying α rewards list balance without rewarding internal balance. α set to zero while varying β would rewards list balance but not internal balance. Although internal balance may appear to imply list balance and a set of perfectly internally balanced documents would give a perfectly balanced list, if the list contains documents that are imbalanced, then the internal balance penalty will not distinguish between a set of unbalanced documents which together are balanced and a set of unbalanced documents which are together unbalanced. The decrease in performance of the set when measured with α or β greater than zero indicates that the measure does capture list characteristics which are distinct from the combined weight of the aspects.

The performance of query 139 "rocky mountain news" illustrates the effects of α and β. Query 139's two aspects are

Table 2: "Query 139 Aspect Relevance Judgments"

Aspect/Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Rocky Mountain	3	1	1	1	1	2	2	2	2	2
News	1	3	3	3	3	3	3	3	3	3

"rocky mountain" and "news." When the parameters are both set to zero, query 139 has a score of 0.881. Setting $\alpha = 1$ imposes a slight penalty, indicating that the ranked list is slightly list-imbalanced towards one of the two topics (in this case "news"). Increasing β to one, however, imposes a much larger penalty, indicating that although the list is as a whole fairly balanced relative to the idealized cumulative gain list, the documents selected are internally unbalanced. An examination of the result list in Table 2 reveals that none of the documents are internally balanced. "Rocky Mountains" is over represented at rank 1 and "News" is over-represented in the lower ranks. Since the first document over-represents an aspect that is under-represented in the other ranks, the list as whole is relatively balanced which is the cause of the small list balance penalty.

Table 3: "Query 139 Ideal Ranked List α=0,β=1 "

Aspect/Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Rocky Mountain	2	2	2	2	2	2	2	2	2	2
News	2	2	2	2	3	3	3	3	3	3

Table 4: "Query 139 Ideal Ranked List α=1,β=0 "

Aspect/Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Rocky Mountain	2	2	2	2	2	2	2	2	2	3
News	3	3	3	3	3	3	3	3	3	1

Tables 3 and 4 depict the ideal ranked lists for the maximum settings for α and β, respectively. Maximizing β draws balanced documents to the top of the list, however no list balancing occurs once the internally balanced documents have been exhausted. Maximizing α creates a list that attempts to maintain balance without regard to the internal balance of the documents selected.

It is important to note that the normalized scores reflect the best possible list given the available documents. If there are not balanced documents available, then a list would score highly even if it were in-balanced. A higher score for one query over another does thus not indicate that that query is more balanced in an absolute sense.

7. CONCLUSIONS

We introduced a new perspective on retrieval evaluation, query-aspect balancing, and a new measure, β-NDCG, for that addresses it. We demonstrated on a small set of queries that the evaluation measure successfully identifies unbalanced result sets. We also illustrated how variations of β-NDCG can be used to understand the impact of internal balance issues (captured by α) and list balance issues (captured by β).

We have shown that β-NDCG captures a previously unconsidered evaluation issue that is a factor for some complex queries. The challenge of using this measure, however, is that it requires a new type of relevance judgment that has not been collected in the past: a division of a query into aspects and document-level judgments of the relevance of the document to each query aspect.

Future work includes expanding β-NDCG to subtopic balance, striving for balance among subtopics (with multi-level relevance judgments) in the retrieved set. In order to do so, the principles underlying this measure would have to be adjusted to accommodate the notion of ambiguity that is used by subtopic oriented measures. β-NDCG applied to subtopics could be compared to α-NDCG.

8. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] J. Carbonell and J. Goldstein. 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*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[3] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[4] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 75–84, New York, NY, USA, 2011. ACM.

[5] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[6] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94, New York, NY, USA, 2008. ACM.