What We Talk About When We Talk About Diversity

Peter B. Golbus Virgil Pavlu Javed A. Aslam
College of Computer and Information Science, Northeastern University
{pgolbus, vip, jaa}@ccs.neu.edu

ABSTRACT

Novelty and diversity are functions of three things: a system's performance at ad-hoc retrieval, its ability to order documents diversely, and the collection over which the system is run. Ideally, our diversity evaluation framework (test collections and evaluation measures) would sort systems by their skill at ordering documents. Unfortunately, the current framework as exemplified by TREC is actually dominated by ad-hoc performance and insensitive to document ordering.

In this paper, we define a measure of *diversity difficulty* for a query and subtopics irrespective of any ranked list, and suggest cases for future failure analysis of diversity evaluation frameworks. We conclude with a suggested retrieval task that will sort systems by their ability at ordering documents given a query's diversity difficulty while controlling for their ability to perform ad-hoc retrieval.

Categories and Subject Descriptors

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

General Terms

Theory, Measurement

Keywords

Information retrieval, evaluation, diversity, diversity difficulty

1. INTRODUCTION

It is impossible to evaluate a system's diversity independent of its performance at ad-hoc retrieval. The problem of ad-hoc performance—ranking documents from highest to lowest probability of relevance—is not easily separable from the problem of ordering documents—for any pair of documents, which is presented to the user first to maximize diversity. Furthermore, poor performance implies poor diversity: a system that returns few relevant documents cannot present a diverse ranked list to the user. Previous work has attempted to quantify the impact of document ordering versus ad-hoc performance on diversity evaluation using ANOVA [2], by comparing diversification strategies [7], and by appeal to intuition [5]. However, these only measure the impact on absolute system scoring; we wished to measure the effect on relative system ranking.

It is also impossible to evaluate a system's diversity independent of the corpus and query. Unambiguous or specific queries are naturally less diverse than queries that are ambiguous or broad. For example, ranked lists retrieved in response to the query "2004 world series winners" will be less diverse than those retrieved for "world series." For judged collections, we can only measure the diversity with respect to a small number of pre-defined subtopics. If our subtopics were defined differently, the documents would be relevant for different subsets of them. The relevant documents themselves also impact the diversity. If all of the documents relevant to a query only discuss one aspect, then even the best system will be unable to create a diverse ranked list. Alternatively, if a random sample of relevant documents display a high level of diversity, then even the most simplistic systems should produce a diverse list. Finally, some corpora may simply be overall more diverse than others.

Because of this, measures of diversity necessarily conflate system performance, document ordering and collection characteristics. Ideally, the most important factor in system evaluation would be a system's skill at ordering documents diversely—collection characteristics affect all systems equally, while ad-hoc performance is a better understood problem and is not the focus of diversity research. Examining the current test collection and diversity evaluation measure framework as exemplified by the TREC2009 and TREC2010 Web tracks [3][4] shows that diversity plays a relatively small role in the ranking of systems. System ranking is dominated by performance; the potential impact of document ordering is, with few exceptions, actually quite small.

We believe that this is due to one or more of the following causes:

1. the measures used do not accurately measure diversity—the measures could be overly dependent on ad-hoc performance or insensitive to the ordering of documents in a list,
2. the corpus does not contain diversity—for any given topic, the documents are preponderantly about a single subtopic,
3. the queries chosen are not diverse—the corpus exhibits diversity, but the documents relevant to the chosen queries are not,
4. the subtopics are flawed—the documents do discuss

DDR '12, Feb 12, 2011 Seattle WA, USA.

multiple subtopics, but not the ones defined for the test collection, or

5. the runs submitted were not able to find diverse documents—there are documents that cover many of the defined subtopics, but they were not submitted to the pool for judging.

In order to distinguish between the first cause (measures) and the remaining causes (collection), we define the *diversity difficulty* (albeit not the novelty) of a query and its subtopics, irrespective of any ranked list. Our measure combines the maximum amount of diversity achievable by any ranked list with the ease of creating a diverse ranked list. We conclude by presenting a diversity task that controls for system performance and diversity difficulty, forcing relative system evaluation to be determined by document ranking.

2. AD-HOC PERFORMANCE

Observe that all diversity measures can be used to measure ad-hoc performance by collapsing all sub-topics into a single aspect and tuning parameters as necessary. For example, $\alpha-$nDCG measured over a single subtopic with $\alpha = 0$ is equivalent to nDCG. Similarly, ERR-IA over a single subtopic is simply ERR.

In order to determine the impact of ad-hoc performance on relative system ranking, we compare the ranking of systems induced by the performance version of a measure against the ranking induced by the actual measure. For example, we compare the rankings induced by nDCG to the rankings induced by α-nDCG. We call such a comparison a "direct" comparison. If a system's diversity is an important factor in its ranking, we would expect these comparisons to give us poorly-correlated rankings. Indeed, the more these two rankings are correlated, the less impact diversity can possibly have.

We can also compare the ranking of one measure's performance version against a different measure, e.g. we can compare nDCG to ERR-IA. We refer to these comparisons as "cross." The purpose of these cross comparisons is to show that whatever diversity is being measured by the diversity measures is completely over-powered by performance; diversity as measured by α-nDCG may or may not be correlated with diversity as measured by ERR-IA but our current framework is too dominated by performance to tell.

Using the TREC-supplied `ndeval` script, we evaluate the systems submitted to TREC2009 and TREC2010 with both the diversity measures (e.g. α-nDCG) and their underlying performance measures (e.g. nDCG). `ndeval` computes twenty-one different scores, each of which induces two ranked lists. For each such pair of diversity and underlying performance measures, we can compute a Kendall's τ rank correlation score between the ranks induced by the two scores. Figure 1 (left column) shows a histogram of the Kendall's τs produced by these comparisons. There are also $2 \times \binom{21}{2} = 420$ different diversity performance measure pairs, each of which (e.g. nDCG to ERR-IA and ERR to α-nDCG) induces a ranked list over which one can compute a Kendall's τ rank correlation score (Figure 1—right column).

In TREC2009 (Figure 1—top row), performance is clearly the dominant factor in a system's evaluation. The situation is much improved in TREC2010 (Figure 1—bottom row), but it is still the case that performance alone is a far better than random predictor of diversity. While we would never expect that these Kendall's τs would be zero, we believe that if our framework measures what we want to measure, these numbers should be much lower than they are.

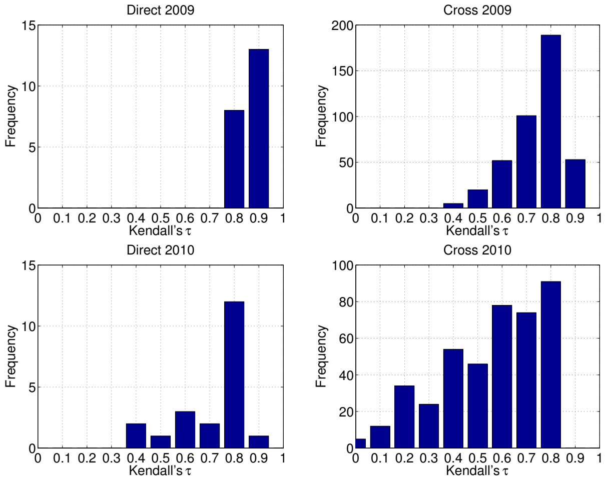

Figure 1: Kendall's τ correlation of performance rankings against diversity rankings in TREC2009 (top row) and TREC2010 (bottom row). The left column shows "direct" comparisons (e.g. α-nDCG to nDCG). The right column shows "cross" comparisons (e.g. α-nDCG to ERR).

3. DOCUMENT RANKING

In this section we explore the sensitivity of the standard novelty and diversity evaluation framework to document ordering. We know that there are document orderings that produce exceptionally large and small scores. What we wish to know is how likely arbitrary choices of orderings are to produce noticeably different scores. The more sensitive our framework is to the choice of document ordering, the more we will be able to rank systems by their ability to diversely order documents.

Given a run, we wish to consider all possible runs with the same performance. Fixing the ranks at which relevant and non-relevant documents appear, we populate the list with a random permutation of all documents relevant to the query. This gives us many ranked lists with many different document orderings, all having the same performance. Viewing this as a random experiment, the scores the various measures assign to these ranked lists are random variables with measurable means and variances. Computing the coefficient of variation of the scores (the standard deviation divided by the mean) produces a normalized measure of the variance of randomly created ranked lists with fixed performance. A low coefficient of variation means that it is unlikely that a system which assigns relevant documents at random will be different than the mean.[1] The larger this number, the more sensitive our framework is to the ordering of documents within ranked lists.

For every choice of system, measure, and query, we create 5000 ranked lists with random relevant documents at the

[1] For normally-distributed data, a coefficient of variation of x% means that roughly 68% of the population is within +/- x% of the mean.

Query ID	Title	# Rel Docs	Subtopic 1	Subtopic 2	Subtopic 3	Subtopic 4	Subtopic 5	d_{mean}	d_{max}	dd
95	earn money at home	0	0	0	0	N/A	N/A	0.000	0.000	0.000
6	kcs	3	3	0	0	0	0	0.200	0.200	0.200
62	texas border patrol	103	89	9	8	0	N/A	0.260	0.500	0.342
25	euclid	120	115	3	2	0	N/A	0.278	0.750	0.406
50	dog heat	89	54	33	29	N/A	N/A	0.684	1.000	0.812
84	continental plates	224	210	160	112	N/A	N/A	0.706	1.000	0.828

Table 1: Examples of subtopic coverage and diversity difficulty in TREC2009 and TREC2010 queries.

relevant ranks. In Figure 2, we present histograms of the coefficients of variation over all query, system, measure triples. For the vast majority of triples, this number was negligible, indicating that the measure was indifferent to the ordering of documents used by that system for that query. While this number could be almost as high as 25%, the histograms show that such occurrences are rare enough to be ignored.

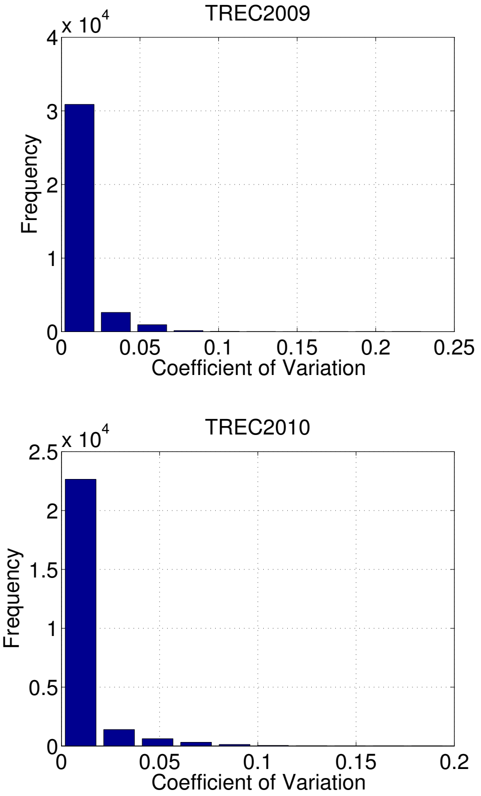

Figure 2: Coefficients of variation of ranked lists with random document orderings and fixed performance. Top: TREC2009. Bottom: TREC2010.

Again, this does not mean that the measure did not evaluate some lists as substantially better than others: the maximum and minimum scores achieved by any of the randomly generated ranked lists were different. However, the vast majority of ranked lists have incredibly similar scores. For a fixed performance, almost all document orderings produce scores that are not appreciably different from the mean.

4. DIVERSITY DIFFICULTY

For a specific query and corpus, query difficulty is a measure of how successful the average search engine should be at ad-hoc retrieval. In this section, we introduce an analogous notion for diversity. Like query difficulty, *diversity difficulty* is defined with respect to a query and a corpus, independent of any ranked list. It must be noted that diversity difficulty describes diversity—the number of subtopics which are covered by a list—only; novelty—which is inversely proportional to the number of times a list repeats a subtopic—cannot be defined independently of a ranked list.

Imagine a query with 10 subtopics, 1,000 documents relevant to only the first subtopic, and each of the remaining subtopics covered by a single, unique document. This query and set of documents exhibits diversity, but it is difficult to generate a diverse list: a system would need to order those nine documents high in the list: the equivalent of finding a handful of “needles” in the “haystack” of 1,000 documents relevant to subtopic 1. On the other hand, if there are large numbers of docs relevant to each subtopic, or large numbers of documents relevant to multiple subtopics, it would be easy to produce a diverse list—almost any list with good performance would exhibit diversity. However, both of these sets of documents contain the same maximum amount of diversity—each of the 10 aspects can be covered by some ranked list.

Diversity difficulty should be a function of these two things: the maximum amount of diversity achievable by any ranked list, and the ease with which a system can produce a diverse ranked list. When the maximum amount of diversity achievable by any system is small, the number should be small. When the maximum amount of diversity is large but it is hard to create a diverse list, this number should increase somewhat. Finally, if the maximum amount of diversity is large and a system created at random will come close to achieving it, the number should be large. The harmonic mean is a function that exhibits such behavior and has been used extensively by the IR community.

We measure the amount of diversity using S-recall [6] at rank k, the number of subtopics covered by a ranked list at rank k divided by the number of subtopics. Given a query, the S-recall of a set of documents is the same for any ranked list of those documents. The maximum amount of diversity (d_{max}) of a query is then easy to define (though NP-hard to compute [1]): it is the maximum possible S-recall for any set of documents in the corpus. Fix k as the size of the minimal set that achieves d_{max}. To measure how easy it is to create a diverse list (d_{mean}) imagine the random experiment of selecting k relevant documents from the corpus and measuring the S-recall. The expectation of this experiment is analogous to the S-recall of a system that is good at ad-hoc retrieval and gives no thought to diversity. The diversity difficulty dd is the harmonic mean of these two numbers.

$$dd = \frac{2d_{max}d_{mean}}{d_{max} + d_{mean}}$$

Since S-recall is a percentage of subtopics, diversity difficulty ranges between zero for difficult queries and one for easy queries.

	Min	Max	Mean
TREC2009	0.143	0.812	0.468
TREC2010	0	0.828	0.581

Table 2: TREC2009 and TREC2010 collection diversity difficult.

Measuring the diversity difficulty of TREC2009 and TREC2010 queries[2] we can see that *dd* behaves as desired. Table 1 shows several queries and the number of relevant documents for each subtopic. Two TREC2010 queries have no relevant documents. For these queries, the *dd* score of 0 is indisputably correct. Query 6, "kcs," has five subtopics, but only one is covered by relevant documents. Its *dd* score of 0.2 accurately reflects its lack of diversity. Query 25, "euclid," has 120 relevant documents and four subtopics. There are 115 documents relevant to one subtopic, while the remaining three have two, three and zero relevant documents. No document is relevant for more than one subtopic. The *dd* score of 0.41 indicates that there do exist some diverse ranked lists, but that such ranked lists are rare. Query 62, "texas border patrol," also has three of four subtopics covered and one dominant subtopic; its *dd* score of 0.34 also indicates the presence and scarcity of diversity. Queries 50, "dog heat," and 84, "continental plates," each have three subtopics, each covered by a large and roughly equal number of relevant documents. These queries with abundant diversity have *dd* scores of 0.81 and 0.83 respectively.

5. CONCLUSION

Diversity is a function of ad-hoc performance, document ordering, and diversity difficulty. We have shown that our current evaluation framework is dominated by performance and insensitive to document ordering. An ideal evaluation framework would rank systems primarily by their ability to order documents diversely. We suggest that the problem is due to one of the following causes:

1. the measures used do not accurately measure diversity,
2. the corpus does not contain diversity,
3. the topics chosen are not diverse,
4. the subtopics are flawed, or
5. the runs submitted were not able to find diverse documents.

We offer a definition for diversity difficulty in the hopes that it can be used to distinguish between the first cause and the remaining causes, and also because we feel that it is a useful and overlooked factor to consider when discussing diversity.

Finally, we observe that our current set-up is not well-suited to measuring systems based on document ordering independent of performance. If this is something that we indeed want to do, a much better task would be to present a system with a topic (and no subtopics) and a number of documents that are either known to be relevant or mostly relevant, and ask the system to select a small number of documents and order them diversely. This would allow us to measure system's relative abilities at diversity while controlling for the much better understood question of ad-hoc performance.

6. AKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1017903. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

7. REFERENCES

[1] B. Carterette. An analysis of np-completeness in novelty and diversity ranking. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 200–211, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] P. Chandar and B. Carterette. Analysis of various evaluation measures for diversity. In *Proceedings of Diversity in Document Retrieval 2011*, 2011.

[3] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *TREC*, 2009.

[4] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the trec 2010 web track. In *TREC*, 2010.

[5] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1043–1052, New York, NY, USA, 2011. ACM.

[6] C. X. Zhai, W. W. Cohen, and J. Lafferty. 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 informaion retrieval*, SIGIR '03, pages 10–17, New York, NY, USA, 2003. ACM.

[7] W. Zheng and H. Fang. A comparative study of search result diversification methods. In *Proceedings of Diversity in Document Retrieval 2011*, 2011.

[2] d_{max} is estimated via greedy approximation.