Analysis of Various Evaluation Measures for Diversity

Praveen Chandar and Ben Carterette

Dept of Computer Science, Univ. of Delaware, Newark, DE 19716 USA
{pcr,carteret}@udel.edu

Abstract. Evaluation measures play a vital role in analyzing the performance of a system, comparing two or more systems, and optimizing systems to perform some task. In this paper, we analyze and highlight the strengths and weaknesses of commonly used measures for evaluating the diversity in search results. We compare MAP-IA, α-nDCG, and ERR-IA using data from TREC'09 web track diversity runs and simulated data. We describe a class of test sets that could be used to compare evaluation measure and systems used for diversifying search results.

1 Introduction

IR researchers have long been interested in optimizing systems to provide results for different users with different information needs when they happen to use the same query. This is known as "novelty and diversity"; the goal of the current research program is to be able to optimize and evaluate retrieval systems by:

1. ability to find relevant material;
2. ability to rank relevant material;
3. ability to satisfy diverse needs;
4. ability to rank documents to satisfy diverse needs.

We evaluate 1 and 3 using simple measures like precision/recall or generalizations to "subtopics" like subtopic precision and subtopic recall [12]. These are set-based measures that, taken alone, do not capture anything about the quality of the ranking. Measures like MAP, DCG, ERR are affected by the ranking as well as the relevance of the documents; generalizations like MAP-IA, α-nDCG, and ERR-IA capture diversity and ranking.

Rank-based diversity measures like MAP-IA, α-DCG, and ERR-IA conflate relevance, diversity, and ranking. They are necessary to have a single value for which to optimize system effectiveness, but the more properties we are evaluating with a single measure, the more likely it is that we mistakenly ascribe an improvement in effectiveness to the wrong cause. Our goal in this work is to investigate the degree to which each of relevance, diversity, and ranking influence the outcome of a measurement of MAP-IA, α-DCG, and ERR-IA.

Amongst the diversity runs submitted to the TREC'09 web track [6] we observed that systems with high relevance nearly always had high diversity scores,

while systems with lower relevance were able to achieve higher diversity. This encouraged us to investigate the sensitivity of MAP-IA, α-DCG, and ERR-IA to relevance, diversity, and ranking. We look at real data (runs submitted to the TREC'09 Web track [6]) as well as simulated data covering more possible cases.

2 Analysis of Evaluation Measures

2.1 Evaluation Measures

As discussed above, evaluation measures for diversity account for both relevance and diversity in the ranking. The degree to which a particular measure is dependent on relevance rather than diversity could potentially have a big impact on system design and optimization. In this section, we briefly discuss commonly used evaluation measures for diversity. We use the values of these measures reported by the `ndeval` utility developed for the TREC Web track.

α-nDCG α-nDCG, an extension of DCG [9], uses a position-based user model [8]. The measure takes into account the position at which a document is ranked along with the subtopics contained in the documents. α-nDCG scores a ranking by rewarding newly-found subtopics and penalizing redundant subtopics geometrically, discounting all rewards with a log-harmonic discount function of rank. α is a parameter controlling the severity of redundancy penalization; we use $\alpha = 0.5$ as done for TREC evaluation.

MAP-IA Mean average precision (MAP) is a very well-known evaluation measure for ad hoc retrieval. The "intent-aware" version computes the MAP for each subtopic separately (assuming the documents relevant to that subtopic are the full set of relevant documents; each subtopic is treated as a distinct interpretation for a given query). MAP-IA is then a weighted average over the subtopics [1].

ERR-IA Chapelle et al. proposed an evaluation measure that is based on interdependent ranking [5]. According to this measure, the contribution of each document is based on the relevance of documents ranked above it. The discount function is therefore not just dependent on the rank but also on the relevance of previously ranked documents. Like MAP-IA, ERR-IA is computed by calculating ERR for each subtopic, then computing a weighted average over subtopics.

2.2 Methods of Analysis

Our primary motivation behind this analysis was to find the relative degree of influence of relevance, diversity, and document ranking on each of α-nDCG, ERR-IA, and MAP-IA. In this section, we describe our determination of categories and the methods used to generate data. Since we use the `ndeval` utility developed for the TREC Web track, the parameters (such as $\alpha = 0.5$ for α-nDCG) for the evaluation measures are the same as used in TREC Web track.

Real Systems In order to observe the levels of relevance and diversity on the current systems, we first looked at the 48 runs submitted to the diversity task

	relevace		
diversity	low	medium	high
high	0	4	12
medium	0	14	2
low	15	1	0

Table 1. Classification of TREC 2009 Web diversity runs into a 3-by-3 table of increasing ability to find relevant documents and increasing ability to find diverse documents.

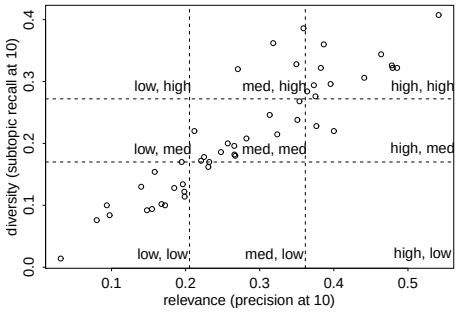

Fig. 1. Precision@10 vs s-recall@10 for 48 systems submitted to the TREC 09 Web track's diversity task. The dashed lines shows relevance and diversity class boundaries.

of the TREC 2009 Web track [6]. We categorized these systems into three levels of relevance based on precision at rank 10 (with a document judged relevant to any subtopic considered relevant for precision@10) and three levels of diversity based on subtopic recall (S-recall) at rank 10 [12] (which is the ratio of unique subtopics retrieved in the top 10 to total unique subtopics). With three levels of each factor, there were nine categories in total. Table 1 gives the number of systems observed in each category. Figure 1 plots S-recall@10 vs precision@10 for these 48 systems to show the breakdown of categories in more detail.

Note that relevance and diversity among these systems are highly correlated. None of the Web track systems have high relevance and low diversity, nor low relevance and high diversity, though both situations are theoretically possible—high relevance/low diversity could be achieved by a system finding many redundant documents, while low relevance/high diversity could be achieved by a system that finds a few relevant documents covering many subtopics. The fact that few systems fall off the diagonal in the Figure 1 suggests that current systems confound relevance and diversity in their ranking approach and therefore may not be good for analyzing general properties of measures.

Simulated Systems Since the real systems do not account for all possible scenarios that we may want to investigate using our measures, we generate several systems in each category using simulations. Since the dependent variable here is the MAP-IA, α-DCG, and ERR-IA scores, the simulated data must be obtained by varying independent variables such as relevance, diversity, document order-

ing, and subtopic distribution. We generated two kinds of simulated systems to study the effect of independent variables on the evaluation measures.

Rel+Div: First, we randomly sample documents from the full Web 2009 *qrels* to create random rankings that satisfy one of our nine experimental conditions: low/medium/high precision@10 and low/medium/high S-recall@10, with labels corresponding to values between $0-0.3$ for low, $0.3-0.6$ for medium, and $0.6-1$ for high. We sampled until we had 10 random rankings in each condition.

Rel+Ord: Next, we controlled diversity ranking in the following way: ten different rankings in each of the same nine relevance/diversity conditions were carefully chosen by varying the minimum rank at which maximum S-recall is obtained. In each category we generate ten rankings in which the documents are re-ordered such that maximum S-recall is obtained only at rank i, where i ranges from 1 to 10. The first ranking (ranking 1, i.e $i = 1$) would attain maximum S-recall at rank 1, the second (ranking 2, i.e $i = 2$) attains max S-recall at rank 2, and so on. In this way we model degrading ability of a system to rank documents.

2.3 Re-ranking Methods

A common way to achieve diversity in a ranking is to first rank by relevance, then re-rank those documents to achieve greater diversity. We briefly describe two re-ranking approaches that we will investigate in this work.

Maximal Marginal Relevance linearly combines a typical bag-of-words relevance score of a document with the amount of "novelty" the document adds to the ranking [2]. The degree of novelty in ranking can be controlled as MMR is a linear combination of relevance and novelty scores. The algorithm prefers documents relevant to the query and least similar to previously ranked documents.

Similarity Pruning is a greedy approach that diversifies the result set by iterating through the initial ranking and removing similar documents [4]. The algorithm iterates over an initial ranking sorted by relevance and prunes documents with similarity scores above a threshold θ.

3 ANOVA

Our goal is to decompose the variance in an evaluation measure into components:

1. variance due to changes in the system's ability to find relevant documents;
2. variance due to changes in the ability of a system to satisfy diverse needs;
3. variance due to changes in the system's ability to rank relevant and diverse documents;
4. variance due to interactions among the above;
5. variance due to topics;
6. variance due to other attributes of a system or other factors.

	SSE in measure (and %age)		
component	ERR-IA	α-nDCG	MAP-IA
relevance	819.0 (22%)	639.9 (16%)	386.2 (11%)
diversity	1075.7 (29%)	1979.3 (52%)	648.8 (20%)
interaction	48.7 (1%)	75.6 (2%)	19.6 (1%)
topic	482.7 (13%)	567.5 (15%)	1362.8 (42%)
residual	1282.5 (35%)	561.5 (15%)	822.1 (25%)

Table 2. Variance decomposition for components affecting the value of each measure. The first three are independent variables we control. The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

Multi-way analysis of variance (ANOVA) is the statistical tool that we will use. In each of our experiments we have at least two independent factors from numbers 1–3 above, as well as one random effect (TREC 2009 topics) for which we have repeated measures on every independent factor. We will not go into details on computing ANOVA, since they can be found in standard statistics textbooks. The numbers we report are derived from the ANOVA procedures in the statistical programming environment R [10]; they are meant to provide intuition about how much we can distinguish between systems that are different on one factor when the rest are held constant.

There are many ways to evaluate evaluation measures; this is one way, but others include detailed examination of single-topic rankings [11], examination of mathematical properties of measures [3], or other data analytic approaches [7].

4 Results

4.1 Varying relevance and diversity

As described above, our first set of simulated data uses two independent factors—relevance as measured by precision@10 and diversity as measured by S-recall@10—with three levels each. We have 47 topics (after dropping those with two or fewer subtopics) and 10 random rankings at each pair of levels. Thus we have $3 \cdot 3 \cdot 45 \cdot 10 = 4050$ total data points for our ANOVA.

Table 2 shows ANOVA variance decomposition for our three measures of interest. From this table we conclude the following:

1. α-nDCG does a much better job at distinguishing between systems that provide different levels of diversity, with 52% of its variance being explained by diversity level as compared to 29% for ERR-IA and 20% for for MAP-IA.
2. MAP-IA is dominated by random variance due to topic sample. This is because the range of achievable MAP-IAs for a given topic depends heavily on the number and distribution of subtopics in documents [3].

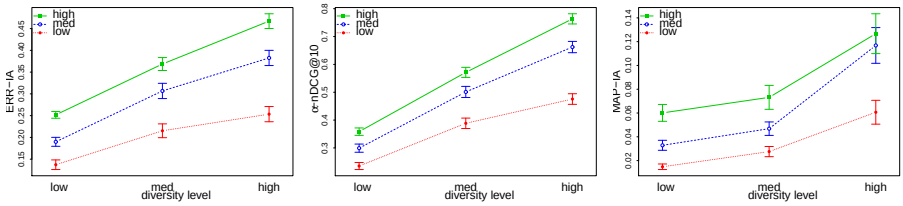

Fig. 2. Effect of increasing diversity and relevance independently on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

	SSE in measure (and %age)		
component	ERR-IA	α-nDCG	MAP-IA
relevance	682.3 (16%)	586.0 (16%)	386.2 (9%)
diversity	891.7 (22%)	1174.6 (47%)	648.8 (14%)
ranking alg	1174.6 (29%)	593.7 (16%)	19.6 (3%)
interactions	477.5 (12%)	298.3 (8%)	152.9 (3%)
topic	347.9 (9%)	497.2 (13%)	1362.8 (35%)
residual	375.0 (12%)	288.1 (7%)	822.1 (35%)

Table 3. Variance decomposition for components affecting the value of each measure. The first four are the independent variables we control (interactions between the first three are aggregated together). The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables or the random effect. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

3. ERR-IA is more strongly affected by unmodeled factors captured in residual error than the other two measures. This may imply that ERR-IA is more sensitive to the ranking of documents than α-nDCG or MAP-IA.
4. Interaction between relevance and diversity plays relatively little role in any of the three measures (though these effects are significant). Our classification of Web track runs suggests interaction effects play a much bigger role in system optimization, however.

Figure 2 shows the mean value of each measure increasing with diversity level for each relevance level, with standard error bars showing randomness due to topic sample. This shows that each measure can distinguish between both different levels of relevance and diversity (as ANOVA analysis suggests). Interestingly, standard error tends to increase with diversity and relevance; this suggests that other factors are affecting the measures more when the systems are better.

4.2 Varying relevance, diversity, and ranking algorithm

The fact that there was so much residual error in the previous results suggests that the ranking algorithm may play a role in determining the measure value (which is not surprising considering that all three use information about ranks).

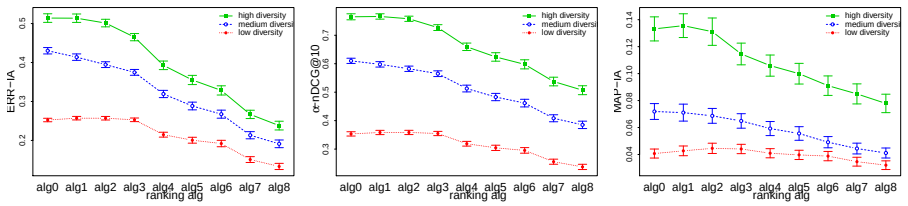

Fig. 3. Effect of degrading a ranking algorithm at independent diversity levels on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

To investigate that, we used our data simulating different ranking algorithms; our 10 random rankings above are now non-random levels of a "ranking" factor. Table 3 summarizes the ANOVA analysis; we see the same trends as before regarding diversity, relevance, and topic effects, but now we see ranking accounts for a large amount of variance in the measure. Residual variance decreases, except in MAP-IA; this suggests that MAP-IA is dominated by undesirable factors.

Figure 3 shows the effect of degrading the simulated ranking algorithm on measure value at different diversity levels (averaged over all relevance levels). Note that the maximum ERR-IA values here are much higher than those shown in Figure 2; this is because the ranking of documents is much more important to ERR-IA than either relevance or diversity alone.

4.3 Effect of reranking algorithm

Finally, we looked at whether the initial level of relevance and diversity affect the efficacy of the reranking-for-diversity approaches we describe above. We reranked results for the random systems using the approaches, then looked at the effect of each of our components on variance in the difference in a measure from the initial ranking to the re-ranked results.

Figure 4 shows that MMR and SimPrune work best when there's high relevance and medium diversity in the initial ranking, and worst when there is already high diversity in the initial ranking, likely because both tend to exclude documents from the original ranking. The wide range in the error bars shows that in general relevance is not a strong factor, only being significant at $p < 0.1$.

5 Conclusions

In this paper, we perform a thorough analysis on various evaluation measures for diversity. We observe that ERR-IA is more sensitive to document ranking and α-nDCG is more sensitive to the diversity among documents retrieved. Further, it is interesting to note that MAP-IA is more sensitive to the topic sample and other factors, which is not desirable in any evaluation measure. The re-ranking approaches were found to be influenced more by diversity in the initial ranking than relevance, with only a medium level of diversity being conducive to improving results after re-ranking.

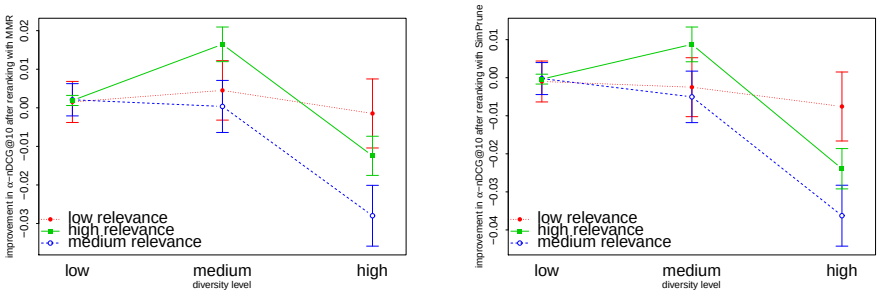

Fig. 4. Effect on α-nDCG@10 of reranking an initial set of results with the given relevance and diversity levels using MMR or SimPrune.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of WSDM '09. pp. 5–14 (2009)
2. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR '98. pp. 335–336 (1998)
3. Carterette, B.: An analysis of NP-completeness in novelty and diversity ranking. In: Proc. ICTIR. pp. 200–211 (2009)
4. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proceeding of CIKM'09. pp. 1287–1296 (2009)
5. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceeding of CIKM '09. pp. 621–630 (2009)
6. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of TREC (2009)
7. Clarke, C.L., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: Proc. WSDM. pp. 75–84 (2011)
8. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of SIGIR '08. pp. 659–666 (2008)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20, 422–446 (October 2002)
10. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010), http://www.R-project.org, ISBN 3-900051-07-0
11. Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., Lin, C.Y.: Simple evaluation metrics for diversified search results. In: Proc. EVIA (2010)
12. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of SIGIR '03. pp. 10–17 (2003)