# A Meta-Evaluation of Evaluation Methods for Diversified Search

Anonymous Authors

**Abstract.** For the evaluation of diversified search results, a number of different methods have been proposed in the literature. Prior to making use of such evaluation methods, it is important to have a good understanding of how diversity and relevance contribute to the performance metric of each method. In this paper, we use the statistical technique ANOVA to analyse and compare three representative evaluation methods for diversified search, namely $\alpha$-nDCG, MAP-IA, and ERR-IA, on the TREC-2009 Web track dataset. It is shown that the performance scores provided by those evaluation methods can indeed reflect two crucial aspects of diversity — richness and evenness — as well as relevance, though to different degrees.

## 1 Introduction

The same query could be submitted to a search engine by users from different backgrounds and with different information needs. When this occurs, the search engine should present users with relevant and diversified results that can cover multiple aspects or subtopics of the query. For more than a decade, there has been a surge of research in the diversification of search results [3,13,17,19]. The main objective of such research is to deal with the ambiguity of query or the multiplicity of user intent.

To evaluate the performance of diversified search, a variety of metrics have been proposed in recent years, such as $\alpha$-nDCG [9], MAP-IA [1], and ERR-IA [5] which generalise the corresponding traditional IR metrics [15] to capture both the *diversity* and the *relevance* of search results. In this paper, we aim to investigate exactly how the above mentioned three representative performance metrics for diversified search are determined by diversity and relevance, using the Analysis of Variance (ANOVA) [10].

## 2 Related Work

The widely used IR performance metric nDCG [12] measures the accumulated usefulness ("gain") of the ranked result list with the gain of each relevant document discounted at lower positions. Clarke et al. proposed its extended version $\alpha$-nDCG [9] to evaluate diversified search results. It takes into account not only the position at which a relevant document is ranked but also the subtopics contained in that document, and uses a parameter $\alpha \in [0, 1)$ to control the severity of redundancy penalisation. Specifically, $\alpha$-nDCG for the top-$k$ search results is

the discounted cumulative gain $\alpha$-DCG[$k$] normalised by its "ideal" value, and DCG[$k$] can be calculated as:

$$\alpha\text{-DCG}[k] = \sum_{i=1}^{k} \frac{\sum_{s=1}^{N} g_{i,s}(1-\alpha)^{\sum_{j=1}^{i-1} g_{j,s}}}{\log_2(i+1)} \ , \tag{1}$$

where $N$ is the total number of distinct subtopics, and $g_{i,s}$ is the human judgement for whether subtopic $s$ is present or not in document $i$.

Agrawal et al. [1] proposed an approach to generalising traditional IR performance metrics for the search results of a query with multiple subtopics (user intents). The idea is to calculate the given performance metric for each subtopic separately, and then aggregate those scores based on the probability distribution of subtopics for the query. Extending the traditional IR performance metrics MAP [15] and ERR [6] in this way, we get their diversified versions:

$$\text{MAP-IA} = \sum_{s=1}^{N} P(s) \cdot \text{MAP}_s \quad \text{and} \quad \text{ERR-IA} = \sum_{s=1}^{N} P(s) \cdot \text{ERR}_s \ , \tag{2}$$

where $N$ is the total number of distinct subtopics, $P(s)$ is the probability or weight of subtopic $s$, while $\text{MAP}_s$ and $\text{ERR}_s$ are the MAP and ERR scores for subtopic $s$ respectively.

The previous studies most similar to our work are those from Clarke et al. [8] and Chandar et al. [4] which attempt to compare evaluation methods in the context of diversified search. The former assumes simple cascade models of user behaviour, while the latter measures diversity just by the subtopic recall — *s-Recall* [18] — which may not reveal the full picture of diversity.

## 3 Meta-Evaluation

### 3.1 Factors

To examine the diversity of search results for a query, it is important to consider not only the number of distinct subtopics but also the relative abundance of the subtopics present in the search result set. Drawing an analogy between subtopics and species, we would like to borrow two measures from ecology [2, 14, 16] — *richness* and *evenness* — to describe the above two complementary dimensions of diversity respectively. The measure of richness on its own cannot provide a complete picture of diversity, as it does not account for the varying proportions of different species in a population. For example, intuitively, one wild-flower field with 500 daisies and 500 dandelions should be more diverse than another wild-flower field with 999 daisies and 1 dandelions — although they both have the same richness (two species), evidently the first field has much higher evenness than the second field.

Formally, we define the two measures, richness and evenness, in the context of diversified search, as follows. The richness of the search result set for a query

(topic) could be just defined as the amount of distinct subtopics appeared in the set. In order to make the value of richness comparable across queries, we choose to use not the absolute number of distinct subtopics but the relative proportion of distinct subtopics:

$$richness = R/N \ , \tag{3}$$

where $R$ is the number of distinct subtopics covered by the given search result set for a query, while $N$ is the total number of distinct subtopics relevant to that query. This proportionate version of richness is actually equivalent to the *s-Recall* proposed by Zhai et al. [18]. The value of (proportionate) richness is obviously between 0 and 1. The evenness of the search result set for a query (topic) refers to how close in numbers each subtopic in the set is, i.e., it quantifies how evenly the search results are spread over the subtopics. For example, a search result set having 5 results from subtopic $u$ and 5 results from subtopic $v$ should have greater evenness than a search result set having 2 results from subtopic $u$ and 8 results from subtopic $v$. Mathematically, the value of evenness is calculated as the normalised diversity:

$$evenness = D/D_{\max} \ , \tag{4}$$

where $D$ is a diversity index, and $D_{\max}$ is the maximum possible value of $D$. Here, we use the well-known *inverse Simpson's diversity index* [11]:

$$D = \left( \sum_{s=1}^{R} p_s^2 \right)^{-1} \ . \tag{5}$$

where $R$ is the number of distinct subtopics covered by the search result set, and $p_s$ is the proportion of subtopic $s$ within the search result set. In this case, it can be proved that $D_{\max}$ is equal to $R$, which happens when all the subtopics appear in the search result set with equal frequencies $\frac{1}{R}$. The value of evenness is greater than 0, and less than or equal to 1.

For the purpose of assessing the *relevance* of search results, we can simply use the Precision@$k$ measure [15], as in [4].

### 3.2 Data

The dataset used for our experiments comes from TREC-2009 Web track diversity task [7] which have also been used in previous studies [4, 8]. This dataset includes 50 topics, each of which consists of a set of subtopics representing different user needs.

### 3.3 Experiments

The evaluation methods for diversified search, including $\alpha$-nDCG, MAP-IA, and ERR-IA, must be able to capture not only the relevance of search results but also the diversity of search results in terms of both richness and evenness. The

statistical technique, Analysis of Variance (ANOVA) [10], provides the perfect tool to gain insight into how each of these three factors (richness, evenness, and relevance) contributes to the overall performance measured by an evaluation method.

In our experiments, the dependent variable for the ANOVA would be the performance score given by $\alpha$-nDCG[1], MAP-IA, or ERR-IA. Regarding the independent variables (richness, evenness, and relevance), since the real IR system outputs submitted to the TREC-2009 Web track could not account for all the possible scenarios that we would like to investigate, we generated a number of synthetic search result sets via a simulation process similar to the "*Rel+Div*" setting in [4]. Given a query (topic) in our dataset, we randomly sampled 10 documents form the full *qrels* file [7] to create such artificial document rankings that satisfy one of the $3^3 = 27$ different experimental conditions for top-10 search results: low/medium/high *richness*, low/medium/high *evenness*, and low/medium/high *relevance*, where the category labels low, medium, and high correspond to the value ranges 0.0–0.3, 0.3–0.6, and 0.6–1.0 respectively. The simulation process would continue until for each of the 50 queries (topics) we had generated 10 search result sets (rankings) per experimental condition. Therefore, the ANOVA for each evaluation method would have $50 \times 10 \times 27 = 13500$ data points to analyse.

### 3.4 Results

The statistical significance results of the ANOVA are shown in Table 1. It can be seen that all those performance metrics, $\alpha$-nDCG, MAP-IA, and ERR-IA, would be influenced heavily by the individual factors — richness, evenness, and relevance — with almost zero $p$-values, but not so much by their interactions. This confirms that the chosen three factors are relatively independent (untangled) aspects of a system's performance for diversified search.

Furthermore, Table 2 shows the variance decomposition results of the ANOVA, where SSE stands for the sum of squared errors. It seems that MAP-IA reflects more richness than the other two performance metrics, as the change of richness accounts for 13% of the total variability in MAP-IA which is substantially higher than 8% in $\alpha$-nDCG and 6% in ERR-IA. On the other hand, evenness is probably reflected better by $\alpha$-nDCG or MAP-IA than ERR-IA, as the change of evenness accounts for 11% of the total variability in $\alpha$-nDCG and MAP-IA but only 7% in ERR-IA. In terms of relevance, $\alpha$-nDCG looks the most accurate indicator, because 10% of its total variability is attributed to the change of relevance, which is followed by 9% in ERR-IA and 6% in MAP-IA. The "residual" component which comprises everything about the performance metric unexplained by the proposed independent variables (factors) occupies a high proportion of the total variability, which suggests that the difficulty of the query (topic) and also the specific ranking algorithm still play the major roles in determining performance scores.

[1] The parameter $\alpha$ for $\alpha$-nDCG was set to 0.5, the default value used in the TREC-2009 Web track diversity task.

**Table 1.** The statistical significance results of the ANOVA.

| Component | $\alpha$-nDCG | | MAP-IA | | ERR-IA | |
|---|---|---|---|---|---|---|
| | $F$ | $p$-value | $F$ | $p$-value | $F$ | $p$-value |
| *richness* | 362.4 | 0.00 | 590.9 | 0.00 | 253.7 | 0.00 |
| *evenness* | 480.0 | 0.00 | 521.7 | 0.00 | 282.7 | 0.00 |
| *relevance* | 465.7 | 0.00 | 285.0 | 0.00 | 397.0 | 0.00 |
| *richness* * *evenness* | 10.8 | 0.00 | 2.9 | 0.03 | 0.9 | 0.46 |
| *richness* * *relevance* | 3.5 | 0.01 | 5.3 | 0.00 | 5.8 | 0.00 |
| *evenness* * *relevance* | 4.3 | 0.00 | 0.3 | 0.91 | 3.2 | 0.01 |
| *richness* * *evenness* * *relevance* | 0.8 | 0.53 | 1.4 | 0.24 | 2.4 | 0.05 |

**Table 2.** The variance decomposition results of the ANOVA.

| Component | $\alpha$-nDCG | | MAP-IA | | ERR-IA | |
|---|---|---|---|---|---|---|
| | SSE | (%) | SSE | (%) | SSE | (%) |
| *richness* | 13.1 | (8%) | 8.2 | (13%) | 2.0 | (6%) |
| *evenness* | 17.4 | (11%) | 7.2 | (11%) | 2.3 | (7%) |
| *relevance* | 16.9 | (10%) | 3.9 | (6%) | 3.2 | (9%) |
| *richness* * *evenness* | 0.6 | (0%) | 0.1 | (0%) | 0.0 | (0%) |
| *richness* * *relevance* | 0.3 | (0%) | 0.1 | (0%) | 0.1 | (0%) |
| *evenness* * *relevance* | 0.3 | (0%) | 0.0 | (0%) | 0.1 | (0%) |
| *richness* * *evenness* * *relevance* | 0.1 | (0%) | 0.0 | (0%) | 0.0 | (0%) |
| residual | 117.0 | (71%) | 44.6 | (70%) | 25.9 | (77%) |

## 4 Conclusion

In this paper, we have shown using the ANOVA that the three representative evaluation methods for diversified search, $\alpha$-nDCG, MAP-IA and ERR-IA, do reflect two crucial aspects of diversity — richness and evenness — as well as relevance, though to different degrees.

## References


1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of the 2nd International Conference on Web Search and Web Data Mining (WSDM). pp. 5–14. Barcelona, Spain (2009)
2. Begon, M., Harper, J.L., Townsend, C.R.: Ecology: Individuals, Populations, and Communities. John Wiley & Sons, 3rd edn. (1996)
3. Carbonell, J.G., Goldstein, J.: 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). pp. 335–336. Melbourne, Australia (1998)
4. Chandar, P., Carterette, B.: Analysis of various evaluation measures for diversity. In: Proceedings of the DDR Workshop. pp. 21–28. Dublin, Ireland (2011)

5. Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., Wu, S.L.: Intent-based diversification of web search results: Metrics and algorithms. Information Retrieval 14(6), 572–592 (2011)
6. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18$^{th}$ ACM Conference on Information and Knowledge Management (CIKM). pp. 621–630. Hong Kong, China (2009)
7. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of The 18$^{th}$ Text REtrieval Conference (TREC). Gaithersburg, MD, USA (2009)
8. Clarke, C.L.A., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: Proceedings of the 4$^{th}$ International Conference on Web Search and Web Data Mining (WSDM). pp. 75–84. Hong Kong, China (2011)
9. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Buttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31$^{st}$ Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 659–666. Singapore (2008)
10. Gamst, G., Meyers, L.S., Guarino, A.: Analysis of Variance Designs: A Conceptual and Computational Approach with SPSS and SAS. Cambridge University Press (2008)
11. Hill, M.O.: Diversity and evenness: A unifying notation and its consequences. Ecology 54(2), 427–432 (1973)
12. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
13. Kingrani, S.K., Levene, M., Zhang, D.: Diversity analysis of web search results. In: Proceedings of the ACM Web Science Conference (WebSci). pp. 43:1–43:2. Oxford, UK (2015)
14. Magurran, A.E.: Ecological Diversity and Its Measurement. Princeton University Press (1988)
15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
16. Pielou, E.C.: An Introduction to Mathematical Ecology. Wiley-Interscience (1969)
17. Santos, R.L., Macdonald, C., Ounis, I.: Search result diversification. Foundations and Trends in Information Retrieval 9(1), 1–90 (2015)
18. Zhai, C., Cohen, W.W., Lafferty, J.D.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: Proceedings of the 26$^{th}$ Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 10–17. Toronto, Canada (2003)
19. Zuccon, G., Azzopardi, L., Zhang, D., Wang, J.: Top-k retrieval using facility location analysis. In: Proceedings of the 34$^{th}$ European Conference on IR Research (ECIR). pp. 305–316. Barcelona, Spain (2012)