Promoting Diversity in Rankings: An MMR Approach

Tabio Romanski, Coen Schoof, Jona te Lintelo

1 INTRODUCTION

Evaluation measures of information retrieval systems do not necessarily take diversity of the retrieved documents into account. Instead, they return the best possible documents, according to the retrieval model. However, when not taking diversity into account, a ranking could contain redundancies such as almost identical documents, which could be regarded as unfavourable by the user. Additionally, diverse rankings should accommodate query ambiguity better since they are likely to present a broader range of documents.

During this project, we use the Washington Post TREC Common Core 2018 Dataset [7] for (re-)ranking. We perform re-ranking using Maximal Marginal Relevance (MMR) [3]. MMR takes two similarity measures as arguments. These similarity measures, in turn, incorporated different fields of the Washington Post documents, with which we have experimented. Additionally, we use three different evaluation metrics upon the re-ranked results with respect to an initial ranking, to determine the effectiveness of the re-ranking with respect to diversity.

2 RELATED WORK

MMR for diversity-based re-ranking was introduced by Carbonell and Goldstein (1998) [3]. Carbonell and Goldstein defined a re-ranking algorithm for information retrieval called MMR for maximizing the difference to the previous ranked search results whilst maintaining the relevance.

Re-ranking for diversity remains a topic receiving considerable interest within the scope of recommender systems (RSs). This interest and importance of diversification started as early as 1980s [6]. One of the earliest introductions of diversification and user relevancy process was by Bradley and Smyth (2001) [2]. More recent approaches to re-ranking exploit customisable evaluation metrics based on lexical, semantic and syntactic features to maximize diversity of a ranking [4] and to solve query ambiguity in search results for users [1] [8].

Many of the approaches from RSs can be used for information retrieval since the two tasks are related. RSs can often rely on more extensive information on the user while that information is only available through the query for simple information retrieval [8].

3 METHOD

With the goal to promote diversity in rankings, we aim to mitigate redundancy in ranking entries. In order to achieve this, we use MMR for re-ranking. This classical approach has been used widely [4] and shown to improve user satisfaction [3]. The approach is defined as follows:

$$MMR = Arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j))] \tag{1}$$

Formula 1 aims to find the maximum diversity and relevance in a given ranking (R) by sequentially adding documents to the re-ranking based on its relevance (Sim_1), subtracted by the maximum similarity (Sim_2) across all other documents in the re-ranking (S). Sim_1 depicts the similarity metric used for the initial ranking (relevance), whereas Sim_2 describes the similarity metric used for re-ranking. The importance of the relevance in comparison to the maximum similarity with documents already contained in the re-ranking is a linear combinations determined by λ. Both similarity metrics are not necessarily the same. Thus, MMR present a flexible approach for re-ranking.

As a consequence of the aim of removing redundancy from the ranking by ranking for diversity, novel evaluation metrics need to be used. These evaluation metrics need to incorporate the diversity of a ranking such that we can compare rankings along that dimension instead of purely relying on relevance of the retrieved documents.

Kunaver et al. [4] have surveyed a number of papers concerned with re-ranking for diversity in RSs including a list of diversity measures. From this survey, we used two diversity measures which are also applicable to information retrieval. First, we used the Bradley and Keith's diversity metric [2]. Formula 2 describes the diversity as the average dissimilarity between all pairs of items in the ranking, based on a given similarity metric. For this, we used the cosine similarity 3.

$$D = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (1 - Similarity(c_i, c_j))}{n/2 \cdot (n - 1)} \tag{2}$$

$$sim(i, j) = \frac{i \cdot j}{||i||||j||} \tag{3}$$

In addition to Bradley and Keith's diversity metric, we also used Vargas' intra-list-diversity (ILD) [8]. It can be understood as an extension of average similarity since it uses pairwise similarity of all documents but also incorporates relevance and distance between documents.

Formula 4 describes the ILD. The ILD is calculated with respect to a document i_k, given a re-ranked list R (in which i_k is contained) and a query u. Calculating the ILD is achieved by multiplying two variables; C'_k and a summation. In turn, the summation is comprised of three variables: $disc$, p, and $dist$.

$disc$ acts as a discount where a document in the re-ranking is penalized when its position is higher than than the position of i_k. Thus, it penalises if two documents are close together in the ranking.

p represents the probability for relevance of a given document in the re-ranking (i_l), given query u (originally, this is the information on the user). p is calculated using a utility function and the maximal relevance value for a query. The utility function is based on the probability of relevance given i_l, and a query and constant τ.

Finally, $dist$ describes the distance between i_k and i_l. $dist$ is defined as 1 - cosine similarity of the TF-IDF vectors of a document pair and represents the cosine dissimilarity of the text body of a document pair.

As the ILD only outputs the diversity of a list with respect to a single document - and needing an output across the whole re-ranking - we decided to calculate the ILD for every document in the re-ranking, and taking the mean across all results (mILD). This would yield divisions by zero if there is no or only one relevant document. Thus, we set the average ILD to zero in case of no relevant documents. We exclude the one problematic summation in case where there is only one relevant document and average over the remaining calculations.

$$ILD(i_k|u, R) = C'_k \sum_l disc(l|k)p(rel|i_l, u)dist(i_k, i_l) \quad (4)$$

$$C'_k = 1/\sum_{l \neq k} disc(l|k)p(rel|i_l, u) \quad (5)$$

$$disc(l|k) = disc(max(1, l - k)) \quad (6)$$

$$p(rel|i_l, u) = \frac{2^{g(u,i)} - 1}{2^{g_{max}}} \quad (7)$$

$$dist(i_k, i_l) = 1 - sim(i_k, i_l) \quad (8)$$

In addition to the evaluation metrics used from the survey by Kunaver et al. [4], we also used topic recall at k. Topic recall is determined by the number of unique topics belonging to relevant documents in the re-ranking@k divided by the total number of unique topics belonging to relevant documents for a given query across the whole corpus. It adds an additional metric, which is not based on the body but on a high level semantic description of the text. Such an evaluation is particularly insightful given that the re-ranking itself is also at least partially based on similarity of the text.

Furthermore, when diversifying search results, relevance could be compromised. Therefore, it is imperative to incorporate relevance into the evaluation. We have used precision@k to determine relevance of the ranking as it captures well how many relevant documents are retrieved.

4 EXPERIMENTAL SETUP

The data used is the Washington Post TREC Common Core 2018 dataset. This dataset consists of articles and blog posts published between 2012-2017 on the Washington Post. The data contains the following features: document id, URL, title, author, published date, kicker, text body, source paragraphs and multimedia. In total, the dataset consists of 595037 text documents. Additionally, the dataset includes 50 queries for which a subset of the dataset has relevancy labels.

Several data features were altered before performing any (re-)ranking. All stop words and punctuation were removed from the text body and title of the article. Additionally, all text was transformed to be lowercase. For the kicker and author features, values that contained more than one word were appended such that every entry only consists of one category and token. The assumption is made that an article written by an author A is different in terms of the text from an article where author A is a co-author. This assumption can be useful when the author is incorporated into the re-ranking. Also, the queries were pre-processed using PyTerrier [5]. Additionally, documents have been removed, if they do not have a title, kicker or annotated relevance since then, our evaluation metrics would have invalid input.

The initial rankings for each query were produced using the BM-25 ranking algorithm. For each query, the initial ranking used for re-ranking consists of the top 50 documents. The scores obtained from BM-25 are normalised using the maximum value and represent the first similarity for MMR (see 1. This normalization ensures that we can compare with the similarity metrics chosen for the re-ranking.

The re-rankings are produced using a similarity matrix for Sim_2 of the MMR equation, which represents the similarity of a candidate document with the documents in the re-ranking. We created two different terms for defining a similarity matrix. The first term uses text body and title. For both of these features, the cosine similarity 3 is calculated, which, in turn, is based upon TF-IDF text representation (i and j are stand for these representations). These two pairwise similarities are combined by adding them, and having a hyperparameter β which influences the trade-off between text and title. The combined similarity is a linear combination of the two values with a higher value of beta leading to more impact of the body. The second term calculates the binary cosine coefficient 9 using the kicker and author represented as binary vectors where i and j are vectors representing one document each. Finally, the aforementioned two terms, being the binary cosine coefficient and the cosine similarity, are added together using linear combinations determined by α. A higher α puts more emphasis on the binary cosine coefficient (title and author). The introduction of the hyperparameter adds additional flexibility tailored to the dataset.

$$sim_{binary}(i, j) = \frac{|i \cap j|}{\sqrt{|i||j|}} \quad (9)$$

The above described setup enables us to perform a number of experiments resulting in an ablation study. We have used combinations of the hyperparameters α and β ranging between 0 and 1 in 0.25 increments. Additionally, λ varies between 0.25 and 0.75 in 0.25 increments. This hyperparameter determines the weight given to the similarity and the estimated relevance.

Finally, we evaluate our ranking using the above presented evaluation metrics. The average ILD and similarity are obtained using the body only since this represents the actual content of the article, on which we would like to diversify. All metrics are evaluated for the top ten documents. The results for the whole ranking (all 50 documents) would always be identical except for ILD since it takes positions into account. Thus, we have decided for the threshold of ten as we deem it a suitable baseline given that one is unlikely to search for results outside of the top ten.

5 RESULTS

Results of our ablation study are shown in Table 1-3 below. Our baseline concerns the initial ranking by BM-25, without re-ranking. The baseline yields the following results for topic recall (TR), average dissimilarity, mILD and Precision@k respectively: 0.194, 0.854, 0.719, 0.431. In the tables, bold text show the highest results for a given metric and λ.

The best topic recall was produced with $\lambda = 0.5$, $\alpha = 0.5$ and $\beta = 0.25$. The results show that $\alpha = 0$ and $\beta = 1$ returned the best value for the average dissimilarity for all λ. The smaller the value for λ, the higher the average dissimilarity. Overall, the best average dissimilarity was for $\lambda = 0.25$, $\alpha = 0$ and $\beta = 1$. Furthermore, the best mILD is given by $\lambda = 0.25$, $\alpha = 0.25$ and $\beta = 0.75$. mILD tends to lower when λ becomes higher. For these settings of λ, a low α and high β also yielded the highest results (0 and 1, respectively). For

Method (λ = 0.25)	**TR**	**Avg. dis.**	**mILD**	**Prec@k**
α = 0, β = 0	0.160	0.894	0.752	0.339
α = 0, β = 0.25	0.164	0.916	0.759	0.321
α = 0, β = 0.5	0.165	0.932	0.756	0.310
α = 0, β = 0.75	0.172	0.937	0.769	0.295
α = 0, β = 1	0.167	**0.949**	0.735	0.283
α = 0.25, β = 0	0.204	0.896	0.773	0.313
α = 0.25, β = 0.25	0.201	0.915	0.777	0.317
α = 0.25, β = 0.5	0.190	0.929	0.787	0.299
α = 0.25, β = 0.75	0.180	0.930	**0.796**	0.283
α = 0.25, β = 1	0.181	0.942	0.762	0.277
α = 0.5, β = 0	0.210	0.894	0.770	0.317
α = 0.5, β = 0.25	0.207	0.909	0.756	0.313
α = 0.5, β = 0.5	0.210	0.920	0.779	0.317
α = 0.5, β = 0.75	0.201	0.926	0.786	0.297
α = 0.5, β = 1	0.199	0.936	0.773	0.297
α = 0.75, β = 0	0.218	0.884	0.746	0.333
α = 0.75, β = 0.25	0.216	0.897	0.750	0.329
α = 0.75, β = 0.5	0.219	0.904	0.770	0.335
α = 0.75, β = 0.75	0.213	0.911	0.754	0.323
α = 0.75, β = 1	**0.221**	0.872	0.758	0.325
α = 1, β = 0	**0.221**	0.872	0.707	**0.343**
α = 1, β = 0.25	**0.221**	0.872	0.707	**0.343**
α = 1, β = 0.5	**0.221**	0.872	0.707	**0.343**
α = 1, β = 0.75	**0.221**	0.872	0.707	**0.343**
α = 1, β = 1	**0.221**	0.872	0.707	**0.343**

Table 1: Results of the ablation study where λ = 0.25

Method (λ = 0.5)	**TR**	**Avg. dis.**	**mILD**	**Prec@k**
α = 0, β = 0	0.195	0.883	0.736	**0.385**
α = 0, β = 0.25	0.188	0.901	0.743	0.379
α = 0, β = 0.5	0.187	0.912	0.767	0.355
α = 0, β = 0.75	0.184	0.916	0.761	0.351
α = 0, β = 1	0.184	**0.925**	**0.770**	0.343
α = 0.25, β = 0	0.219	0.884	0.741	0.353
α = 0.25, β = 0.25	0.221	0.898	0.741	0.355
α = 0.25, β = 0.5	0.223	0.904	0.748	0.357
α = 0.25, β = 0.75	0.217	0.911	0.752	0.341
α = 0.25, β = 1	0.212	0.915	0.758	0.337
α = 0.5, β = 0	0.227	0.880	0.747	0.353
α = 0.5, β = 0.25	**0.230**	0.888	0.748	0.359
α = 0.5, β = 0.5	0.226	0.900	0.745	0.351
α = 0.5, β = 0.75	0.222	0.904	0.750	0.341
α = 0.5, β = 1	0.222	0.909	0.753	0.343
α = 0.75, β = 0	0.217	0.874	0.709	0.339
α = 0.75, β = 0.25	0.221	0.879	0.711	0.349
α = 0.75, β = 0.5	0.220	0.878	0.712	0.345
α = 0.75, β = 0.75	0.226	0.888	0.746	0.345
α = 0.75, β = 1	0.223	0.893	0.746	0.341
α = 1, β = 0	0.221	0.872	0.707	0.343
α = 1, β = 0.25	0.221	0.872	0.707	0.343
α = 1, β = 0.5	0.221	0.872	0.707	0.343
α = 1, β = 0.75	0.221	0.872	0.707	0.343
α = 1, β = 1	0.221	0.872	0.707	0.343

Table 2: Results of the ablation study where λ = 0.5

all cases of λ and β where α = 1, the values are identical, for any metric.

Finally, for precision@k, the best values were found when λ = 0.75, α = 0 and β = 0. For λ = 0.5, precision@k was also highest at α = 0 and β = 0. Remarkably, this was not the case for λ = 0.25, where the highest value was found at α = 1, regardless of the value of β. Precision@k tends to become higher when λ becomes higher.

6 DISCUSSION

Overall, the results indicate that the precision@k worsens with re-ranking whereas the diversity by any metric increases during re-ranking. Nevertheless, there are several subtle influences of the three hyperparameters.

Quite surprising was to see the relatively bad results of topic recall. This, however, is to be expected considering the overall poor precision@k, which indicates that we retrieve only few relevant documents. Since we only consider such topics for the topic recall which belong to relevant documents, it is obvious that the topic recall is heavily influenced by the number of relevant documents in the re-ranking@k.

The poor precision in turn, is explained by the rather poor performance of the initial ranking using BM-25. The subsequent rankings can only be as good as the initial ranking unless, by coincidence, a highly diverse document happens to be relevant and substitutes a non-relevant one. The re-rankings perform even poorer in terms of precision because we use the original BM-25 ranking score and similarity between documents to reorder the documents. This procedure is unlikely to improve relevance as the only added information for the re-ranking are the similarity scores. The explanation is consistent with the fact that more emphasis on the initial ranking (high λ) yields better precision.

The observation that the average dissimilarity is highest whenever β = 1 was expected. This is because it puts all emphasis in the re-ranking on the body instead of the title. The body, in turn, is used to calculate the average dissimilarity. Thus, it shows that the re-ranking is indeed effective. Additionally, we presume that the title (which is of higher importance for lower betas) does not contain sufficient information to re-rank for diversity effectively given its length compared with the body. Since dissimilarity is also relevant for the calculation of mILD, higher β values tend to yield good results for mILD. These trends are further emphasised by low λ values since it emphasises the re-ranking.

Furthermore, if α is set to 0, the author and the kicker are not relevant for the similarity score anymore. Thus, the combination of low α and high β yields the highest scores of mILD and average dissimilarity. Thus, topic, title and author are not sufficient to obtain good results for these two metrics, especially given the distance to the other re-rankings.

However, given the dependence of the similarity-based metrics and the re-ranking algorithm, it is important to consider our third metric. Here we can observe some interesting dynamics. Firstly, the hyperparameter λ, which determines the weight of original ranking

Method ($\lambda = 0.75$)	**TR**	**Avg. dis.**	**mILD**	**Prec@k**
$\alpha = 0, \beta = 0$	0.200	0.873	0.725	**0.413**
$\alpha = 0, \beta = 0.25$	0.195	0.880	0.732	0.405
$\alpha = 0, \beta = 0.5$	0.191	0.886	0.752	0.395
$\alpha = 0, \beta = 0.75$	0.191	0.893	0.754	0.391
$\alpha = 0, \beta = 1$	0.192	**0.909**	**0.759**	0.387
$\alpha = 0.25, \beta = 0$	0.203	0.875	0.731	0.387
$\alpha = 0.25, \beta = 0.25$	0.205	0.880	0.733	0.395
$\alpha = 0.25, \beta = 0.5$	0.207	0.878	0.738	0.393
$\alpha = 0.25, \beta = 0.75$	0.213	0.888	0.756	0.389
$\alpha = 0.25, \beta = 1$	0.211	0.892	**0.759**	0.383
$\alpha = 0.5, \beta = 0$	0.221	0.873	0.736	0.389
$\alpha = 0.5, \beta = 0.25$	0.222	0.874	0.738	0.385
$\alpha = 0.5, \beta = 0.5$	**0.223**	0.874	0.740	0.385
$\alpha = 0.5, \beta = 0.75$	0.221	0.879	0.742	0.379
$\alpha = 0.5, \beta = 1$	0.218	0.879	0.743	0.375
$\alpha = 0.75, \beta = 0$	0.213	0.871	0.703	0.357
$\alpha = 0.75, \beta = 0.25$	0.214	0.874	0.703	0.359
$\alpha = 0.75, \beta = 0.5$	0.217	0.876	0.705	0.365
$\alpha = 0.75, \beta = 0.75$	0.216	0.878	0.705	0.363
$\alpha = 0.75, \beta = 1$	0.217	0.879	0.706	0.361
$\alpha = 1, \beta = 0$	0.216	0.872	0.704	0.353
$\alpha = 1, \beta = 0.25$	0.216	0.872	0.704	0.353
$\alpha = 1, \beta = 0.5$	0.216	0.872	0.704	0.353
$\alpha = 1, \beta = 0.75$	0.216	0.872	0.704	0.353
$\alpha = 1, \beta = 1$	0.216	0.872	0.704	0.353

Table 3: Results of the ablation study where $\lambda = 0.75$

vs re-ranking, should not emphasise the re-ranking to preserve topic recall. This can be explained by the fact that relevance determines topic recall to a certain extent (see above) and the initial ranking is responsible for the relevance. Overall, this means that relevance is sacrificed for dissimilarity, from which topic recall suffers.

Besides, topic recall is least poor when both α and β are roughly equal and not to the extremes for two out of three λ values. This is somewhat surprising since we would expect to see the highest topic recall when diversifying based on topic and author (high α). This expectation is confirmed for $\lambda = 0.25$. Hence, when contrasting topic recall with the similarity-based metrics (mILD and average dissimilarity), it becomes apparent that also title and author should receive some attention in the re-ranking since the metrics disagree.

The last observation made was that for $\alpha = 1$ all values are identical. This phenomenon occurs because we disregard the title and body completely for $\alpha = 1$, thus, β does not have any impact on the re-ranking.

6.1 Limitations

Our research also introduces some limitations. Firstly, the TREC Washington Post dataset was incomplete to an extent, which could have influenced our results. E.g., many documents were not annotated with a relevance, had no title/kicker/body/author. Due to this, we were forced to regard documents without authors as being written by the same author, which could have impacted the re-ranking. Moreover, after the pre-processing, we were left with a small fraction of the original dataset, which could be detrimental to the generalizability of our results.

Secondly, our initial ranking, based on BM-25, yielded poor results. This, in turn, influenced the results of the re-ranking. Moreover, such poor results may not be representative of common ranker performance.

Thirdly, due to restrictions in time as well as resources, we were forced to limit our choice in hyper-parameters, therefore lowering the granularity of our results. This, in turn, could have had impact upon our interpretations derived from the results.

Finally, due to our choice to only regard relevant documents when calculating the topic recall, we diluted the metric by also emphasizing the overall recall at the expense of accurately measuring the topic recall. The alternative would have been to include both relevant as well as irrelevant documents when counting relevant topics, which would have been a more accurate measure of topic recall since all topics are counted. However, since the overarching goal of promoting diversity in re-ranking was to maintain overall ranking relevance, we chose for the former design choice, despite its limitations.

6.2 Future Work

Future work within aiming at diversifying rankings should implement re-ranking for multiple corpora and subject domains in order to verify generalizability of our findings. Additionally, other re-ranking and evaluation methods should be compared with each other such that more aspects of diversity and more aspects of the documents have influence on the diversity measure than just the body, title, author and kicker.

7 CONCLUSION

In this report, we have re-ranked a ranking of Washington Post articles using MMR. For the similarity metrics needed for MMR, we used the cosine similarity and the binary cosine coefficient. For evaluating the diversity of the re-ranking, we used the topic recall, average dissimilarity and mILD. For evaluating the relevance of the resulting re-ranking, we used precision@k. The best results for diversity in terms of similarity-based diversity were achieved by diversifying based on text bodies, which is a consequence of the evaluation metric. However, topic recall was best by taking a broad range of factors into account when re-ranking. This emphasises that the re-ranking should be carefully designed based on the aims of the final ranking i.e. which notion of diversity one applies. Increasingly improving diversity comes at the cost of decreasing relevance of the ranking and should thus only be carefully selected.

REFERENCES

[1] G. Adomavicius and YoungOk Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (May 2012), 896–911. https://doi.org/10.1109/tkde.2011.15

[2] Keith Bradley. 2001. Improving Recommendation Diversity. *Proc. AICS '01* (09 2001).

[3] Jaime Carbonell and Jade Goldstein. 1998. 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*. ACM Press. https://doi.org/10.1145/290941.291025

[4] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (May 2017), 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

[5] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation inInformation Retrieval using PyTerrier. In *Proceedings of ICTIR 2020*.

[6] Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

[7] TREC. 2020. https://trec.nist.gov/data/wapost/

[8] Súul Vargas. 2011. New Approaches to Diversity and Novelty in Recommender Systems. In *Electronic Workshops in Computing*. BCS Learning & Development. https://doi.org/10.14236/ewic/fdia2011.2

A PROJECT NOTEBOOK

The project notebook can be found following the link below:

https://colab.research.google.com/drive/1L9olWSJRTfmHw22DfEubwI568k-oRofN?usp=sharing