Learning to Improve Affinity Ranking for Diversity Search

Yue Wu[1], Jingfei Li[1], Peng Zhang[1], and Dawei Song*[1,2]

[1] School of Computer Sci & Tec, Tianjin University, Tianjin, China
[2] The Computing Department, The Open University, UK
{yuewuscd,jingfl}@foxmail.com,{pzhang,dwsong}@tju.edu.cn

Abstract. Search diversification plays an important role in modern search engine, especially when user-issued queries are ambiguous and the top ranked results are redundant. Some diversity search approaches have been proposed for reducing the information redundancy of the retrieved results, while do not consider the topic coverage maximization. To solve this problem, the Affinity ranking model has been developed aiming at maximizing the topic coverage meanwhile reducing the information redundancy. However, the original model does not involve a learning algorithm for parameter tuning, thus limits the performance optimization. In order to further improve the diversity performance of Affinity ranking model, inspired by its ranking principle, we propose a learning approach based on the learning-to-rank framework. Our learning model not only considers the topic coverage maximization and redundancy reduction by formalizing a series of features, but also optimizes the diversity metric by extending a well-known learning-to-rank algorithm LambdaMART. Comparative experiments have been conducted on TREC diversity tracks, which show the effectiveness of our model.

Keywords: Search diversification, Affinity Ranking, Learning-to-rank

1 Introduction

Search diversification plays an important role in modern search engine, especially when user-issued queries are ambiguous and the top ranked results are redundant. Some diversity search approaches have been proposed (e.g., Maximal Marginal Relevance (MMR) [1] and its numerous variants [5, 7, 9]) for reducing the information redundancy of the retrieved results, while do not consider the topic coverage maximization.

In order to address the aforementioned drawbacks of traditional implicit diversity approaches, Zhang et al. [8] proposed an innovative method named Affinity Ranking (AR) model which pursues the query subtopics coverage maximization and information redundancy reduction simultaneously. Specifically, AR applies a content-based document graph to compute the information coverage score for each document and imposes a penalty score to the information coverage score in order to reduce the information redundancy, then ranks documents according to the final document score which linearly combines the query relevance information score and the diversity information (i.e., topic coverage information and redundancy reduction information) score of the document. However, the

original Affinity ranking model model uses a predefined heuristic ranking function which can only integrate limited features and has many free parameters to be tuned manually. A direct idea to solve this problem is to borrow machine learning methods to train the Affinity ranking model. Intuitively, the Affinity ranking model is similar to the traditional retrieval ranking model (e.g., query likelihood Language Model) which ranks documents in descending order according to document scores. Therefore, improving the Affinity ranking model with learning-to-rank technique is reasonable and feasible. To do this, in this paper, we addressed three pivotal problems, i.e., (i) how to redefine the ranking function which can incorporate both relevance information and diversity information within an unified framework; (ii) how to learn the ranking model by optimizing the diversity evaluation metric directly; (iii) how to extract diversity features (i.e., topic coverage features and redundancy reduction features) inspired by the Affinity ranking model. Particularly, we propose a learning based Affinity ranking model by extending a well-known Learning-to-Rank method (i.e., LambdaMART). Extensive comparative experiments are conducted on diversity tracks of TREC 2009-2011, which show the effectiveness of our method.

2 Model Construction

2.1 Overview of Affinity Ranking Method

This subsection gives a brief description of the Affinity Ranking model [8] which maximizes the topic coverage and reduces the information redundancy. At first, they introduce a directed link graph named Affinity Graph to compute the information richness score which represents how many the query subtopics have been covered for each document. Similar to the PageRank, the information richness score for each document is obtained through running the random walk algorithm. The documents with largest information richness score (subtopic coverage information) will be returned to users. Meanwhile, in order to reduce the information redundancy, they compute the Affinity ranking score by deducting a diversity penalty score for each document as described in the Algorithm 1. However, improving the diversity may bring harm to the relevance quality. In order to balance the diversity ranking and relevance ranking, their final ranking function linearly combines both original relevance score and Affinity ranking score, and then they sorts the documents in descending order according to the final combination score.

In order to obtain a good diversity performance (in term of the diversity evaluation measures) and incorporate more features, we propose a learning approach which is illustrated in the following parts.

2.2 Learning Diversity Ranking Method

We build a learning based Affinity ranking model with the help of learning-to-rank technique (the LambdaMART [6] algorithm) to improve the diversity ability of Affinity ranking model. In following parts, we will redefine the ranking function, label, the objective function of learning algorithm, and then describe the features of our learning model.

Algorithm 1 : The greedy algorithm for diversity penalty.

Input: $InfoRich(d_i)$: information richness score, D: candidate document set, $\hat{M}_{ji}$: the weight of link in the graph, $\hat{M}_{ji}InfoRich(d_i)$: penalty score.
Output: Affinity ranking score $AR(d_i)$ for every document d_i

```
for d_i ∈ D do
    AR(d_i) = InfoRich(d_i)
end for
while D ≠ empty do
    d_i = arg max_{d_i ∈ D}(AR(d_i))
    D ← D − d_i
    for d_j ∈ D do
        AR(d_j) = AR(d_j) − M̂_ji InfoRich(d_i)
    end for
end while
```

Learning algorithm for diversity search For the original LambdaMART, the ranking score of each document can be computed by ranking function $f(x) = w^T x$ where x is the document feature vector which only consider the relevance. However, for diversity task, we need to incorporate both relevance, redundancy reduction and topic coverage maximization. Inspired by the ranking function of the Affinity Ranking model, we can extend the ranking function as described in the Eq.1

$$f(w_1, w_2, w_3, x, y, z) = w_1^T x + w_2^T y + w_3^T z \tag{1}$$

where the w_1, w_2 and w_3 encodes the model parameters, the x, y is topic coverage maximization and redundancy reduction feature vector respectively while the z is relevance feature vector. Even if we have the reasonable ranking function, it is still a big challenge to redefine the objective function for using the diversity metric to guide the training process.

Unlike others, the LambdaMART algorithm defines the derivatives of objective function with the respect to document score rather than deriving them from the objective function. For the document pair $< i, j >$ (the document i is more relevant than document j), the derivatives λ_{ij} is

$$\lambda_{ij} = sigmoid(s_i - s_j)|\Delta Z_{ij}| \tag{2}$$

where s_i is the model score of document i and the $|\Delta Z_{ij}|$ is the change value of evaluation metric when swapping the rank positions of document i and j. We know that LambdaMART can be extended to optimize any IR metric by simply replacing $|\Delta Z_{ij}|$ in Eq.2. However, the evaluation metric needs to satisfy the property that if irrelevant document ranks before the relevant document after swapping (that is, wrong swapping), the metric should decrease (ie., $\Delta Z_{ij} < 0$). So if we extend derivatives λ_{ij} by using the current diversity metric (e.g., α-$NDCG$ [3] or ERR-IA [2]), some adjustments should be made. The relevance label of a document is one value in original LambdaMART to decide the relevant-irrelevant document pair used in the Eq.2, while our label should a multiple values (in which each value represents whether the document is relevant to the each query subtopic) in order to compute the change value of diversity metric. So we assume that the document covering at least one query subtopic is more relevant than documents covering no any query subtopics. Thus

the label of the document covering at least one query subtopic is bigger than the document covering no any query subtopics. Therefore, the document label used in the training procedure contains two part. And then after defining the relevant-irrelevant document pair, we should show diversity metrics satisfy the above property. We choose the α-$NDCG$ as the representative because ERR-IA is same in rewarding the relevant document ranking before the irrelevant document. In the top k results of a return list for query q, for example, there are m documents which covers at least one query subtopic where four documents among the m is relevant to the query subtopic t. We denote the ranking positions of the four documents as p_1, p_2, p_3, p_4 where $0 < p_1 < p_2 < p_3 < p_4 < k$. If one relevant document (we use d_{p_2} in the following proof case, which means the document at the position p_2) swaps with another irrelevant document which ranking position is beyond k, we have proved that $\Delta Z < 0$. Let Z is the α-$NDCG@k$ before the swapping while $\tilde{Z}$ is the α-$NDCG@k$ after the swapping (the value of α is between 0 and 1). When only considering the query subtopic t, we have $\Delta Z_t = \frac{\tilde{Z}_t - Z_t}{ideaDCG@k} = \frac{1}{ideaDCG@k}\sum_{j=1}^{k}\frac{\tilde{G_t[j]} - G_t[j]}{\log_2(1+j)} = \frac{1}{ideaDCG@k}((\frac{(1-\alpha)}{\log_2(1+P_1)} - \frac{(1-\alpha)}{\log_2(1+P_1)}) + (0 - \frac{(1-\alpha)^2}{\log_2(1+P_2)}) + (\frac{(1-\alpha)^2}{\log_2(1+P_3)} - \frac{(1-\alpha)^3}{\log_2(1+P_3)}) + (\frac{(1-\alpha)^3}{\log_2(1+P_4)} - \frac{(1-\alpha)^4}{\log_2(1+P_4)}))$ $< (\frac{(1-\alpha)^2}{\log_2(1+P_3)} - \frac{(1-\alpha)^2}{\log_2(1+P_2)}) + (\frac{(1-\alpha)^3}{\log_2(1+P_4)} - \frac{(1-\alpha)^3}{\log_2(1+P_3)}) - \frac{(1-\alpha)^4}{\log(1+P_4)} < 0$. The same is true for every subtopics. Through above adjustments, it is suitable for using the diversity metric as part of objective function to guide the training process.

Feature extraction For topic coverage maximization features, we use the information richness score used in the Affinity ranking model. For information redundancy reduction features, we formalize it according to Algorithm 1:

$$f(q, d_i, D_q) = t(q, d_i) - \sum_{d_j \in D_q} p(d_j, d_i) \tag{3}$$

where document set D_q is the already selected document set, $t(q, d_i)$ measures how many query topics has been covered by the document d_i, $p(d_j, d_i)$ denotes the penalty score that the document d_j deploy to d_i for the information redundancy. In this paper, we can use different form of $t(q, d_i)$ and $p(d_j, d_i)$ to produce diversity features for capturing redundancy reduction information. Then, for relevance features, we use some common features. Detailed features used are shown in Table 1.

3 Experiments and Results

3.1 Experimental setting

We evaluate our method using the diversity task of the TREC Web Track from 2009-2011, which contains 148 queries. We use the ClueWeb09 category-B as the document collection and the official evaluation metrics of diversity task (α-$NDCG$ [3] where α is 0.5 and ERR-IA [2]). All approaches are tested by re-ranking the original top 1000 documents retrieved by the Indri search engine (implemented with the query likelihood Language Model abbreviated with LM) for each query. For all approaches with free parameters, 5-fold cross validation

Table 1. Diversity and Relevance features for learning on ClueWeb09-B collection

Feature	Description
TopicCovFea0	information richness score in the [8]
RedReduceFea1	$t(q, d_i)$ is information richness score, $p(d_j, d_i)$ is penalty score in the [8]
RedReduceFea2	$t(q, d_i)$ is TF-IDF score, $p(d_j, d_i) = \sqrt{t(q, d_i)}\sqrt{t(q, d_j)}f(d_i, d_j)$, $f(d_i, d_j)$ denotes cosine similarity between documents d_i and d_j represented with TF-IDF vectors
RedReduceFea3	$t(q, d_i)$ is BM25 score, $p(d_j, d_i)$ is same to RedReduceFea2
RedReduceFea4	$t(q, d_i)$ is LMIR with ABS smoothing, $p(d_j, d_i)$ is same to RedReduceFea2
RedReduceFea5	$t(q, d_i)$ is LMIR with DIR smoothing, $p(d_j, d_i)$ is same to RedReduceFea2
RedReduceFea6	$t(q, d_i)$ is LMIR with JM smoothing, $p(d_j, d_i)$ is same to RedReduceFea2
RelFea7	sum of query term frequency for every document
RelFea8	length for the every document
RelFea9-13	sum,min,max,mean,variance of document term frequency in collection
RelFea14-18	sum,min,max,mean,variance of document tfidf in collection
RelFea19-23	tfidf score, BM25 score, LMIR score with ABS, DIR, JM smoothing

is conducted. We tested 5 baseline approaches including the original query likelihood Language Model (LM), MMR [1], quantum probability ranking principle (QPRP)[9], RankScoreDiff [4] and Affinity Ranking model (AR) [8].

3.2 Result and Analysis

In this section, we report and analyze the experiment results to investigate the effectiveness of the proposed diversity model. If our model uses α-$NDCG$ in objective function, it is denoted as LAR(α-$NDCG$) while it is denoted as LAR(ERR-IA) for using ERR-IA. At first, we compare AR model with other baselines to show the diversity ability of Affinity ranking model. From Table 2, we find AR model has better performance than other baselines. Moreover, we find that the result list does not achieve good diversity ability in term of diversity evaluation α-$NDCG$ and ERR-IA for two approaches[1, 9] which only reduce the redundancy. The experiment result shows that a group of document with low redundancy can not achieve large subtopic coverage. For RankScoreDiff approach [4], one only considers the subtopic coverage maximization, also outperform MMR and QPRP[1, 9]. The experiment results illustrate that query subtopic coverage maximization is more important than low information redundancy for diversity search. Secondly, we compare our model with AR model to prove that our model (both LAR(α-$NDCG$) and LAR(ERR-IA) model) improve the diversity ability of AR model significantly. For our proposed learning model, for using α-$NDCG$ as evaluation metric, the improvement percentages of compared with the AR model is 26.96% for LAR(α-$NDCG$) and 31.31% for LAR(ERR-IA) respectively. When uses ERR-IA as evaluation metric, the improvement percentage is 43.68% for LAR(α-$NDCG$) and 50.42% for LAR(ERR-IA) respectively.

Table 2. Diversification performance of the models

Metric	LM	MMR	QPRP	RankScoreDiff	AR	LAR (α-$NDCG$)	LAR (ERR-IA)
α-$NDCG$	0.2695	0.2681↓	0.1663↓	0.2705↑	0.2711↑	0.3442↑	0.3560↑
ERR-IA	0.1751	0.1715↓	0.1266↓	0.1767↑	0.1765↑	0.2536↑	0.2655↑

4 Conclusions and Future Work

In this paper, we build a learning diversity model within the framework of learning-to-rank to improve the diversity ability of Affinity Ranking model. Our motivation comes from that the Affinity Ranking model can reduce the redundancy and make topic coverage maximization. Beyond that, the ranking principle of Affinity Ranking model makes it possible to build learning model with help of learning-to-rank approach. The final comparative experiments have shown that our approach is effective. In the future, we will propose better topic coverage representation technique to formalize the better diversity features.

Acknowledgements: This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2013CB329304, 2014CB744604), the Chinese 863 Program (grant No. 2015AA015403), the Natural Science Foundation of China (grant No. 61272265, 61402324), and the Tianjin Research Program of Application Foundation and Advanced Technology (grant no. 15JCQNJC41700).

References

1. J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336. ACM, 1998.
2. O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630. ACM, 2009.
3. 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 *SIGIR*, pages 659–666. ACM, 2008.
4. S. Kharazmi, M. Sanderson, F. Scholer, and D. Vallet. Using score differences for search result diversification. In *SIGIR*, pages 1143–1146. ACM, 2014.
5. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR*, pages 115–122. ACM, 2009.
6. Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.
7. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17. ACM, 2003.
8. B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR*, pages 504–511. ACM, 2005.
9. G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Advances in information retrieval*, pages 357–369. Springer, 2010.