A Pruning Strategy for Optimal Diversified Search

Fei Chen[1,2,3,*], Yiqun Liu[1,2,3], Jian Li[4], Min Zhang[1,2,3], and Shaoping Ma[1,2,3]
[1] State Key Laboratory of Intelligent Technology and Systems
[2] Tsinghua National Laboratory for Information Science and Technology
[3] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[4] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
* chenfei27@gmail.com

ABSTRACT

Recommended poster submission from research track, original submission number 111.

Given a number of possible sub-intents (also called subtopics) for a certain query and their corresponding search results, diversified search aims to return a single result list that could satisfy as many users' intents as possible. Previous studies have demonstrated that finding the optimal solution for diversified search is NP-hard. Therefore, several algorithms have been proposed to obtain a local optimal ranking with greedy approximations. In this paper, a pruned exhaustive search algorithm is proposed to decrease the complexity of the optimal search for the diversified search problem. Experimental results indicate that the proposed algorithm can decrease the computation complexity of exhaustive search without performance loss.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation.

Keywords

Diversified search, exhaustive search, NP-Hard

1. INTRODUCTION

Diversified search aims to produce a search result list which could meet the information needs for ambiguous or underspecified queries [1]. However, since search result diversification was proven to be a computational expensive problem [2], it is almost impossible to generate an ideal diversified ranking list for commercial search engines. Several greedy search algorithms such as *IA-Select* [3] were therefore proposed to find an approximation for the ideal diversified ranking list. In this paper, supposing that (1) subtopics and their weights underlying a query are known, and (2) the relevance between the document and subtopics are available, we propose a pruning exhaustive search algorithm for search result diversification to decrease the computation complexity of the exhaustive search without losing any performance.

2. PRUNED EXHAUSTIVE SEARCH

To better describe our algorithm, we first define some symbols in use. Figure 1 shows three different result lists that are composed of the ranked results. The symbol d_l in these lists is used to distinguish a certain document from other documents. It does not stand for the document ranks at the l-th slot of the list. The only difference between List 1 and List 2 is that List 2 contains no documents in either the l-th or the k-th slots ($l<k$). List 3 is the same with List 1 except that we exchange d_l with d_k. Sc_1, Sc_2 and Sc_3 in Figure 1 respectively represent the evaluation scores of *List 1*, *List 2* and *List 3* in terms of a certain diversity metric. If a document d_k is added at the l-th slot of List 2, the total score change of the list may be divided into two parts. The first part of the score change is from d_k itself because no document exists in the l-th slot of List 2 before d_k is added to this position. We denote the score for d_k in the l-th slot as G_{kl}. The second part of the score change results from the documents after the l-th slot. If d_k is not relevant to any subtopic covered by the documents after d_k, d_k will not affect the second part of the total score. If d_k is relevant to any subtopic covered by any document after d_k, the score of the second part will decrease. We denote the absolute value of this score decrement as A_{kl}. Because List 2 in Figure 1 has two "empty" slots at the l-th and k-th slots ($l<k$), we can further divide the score decrement A_{kl} into two subparts: the first subpart of the decrement stems from the documents between the l-th slot and the k-th slot, and the second subpart is from the documents after the k-th slot (the k-th slot in List 2 is empty). We denote their absolute values as I_{kl} and B_{kl}, respectively.

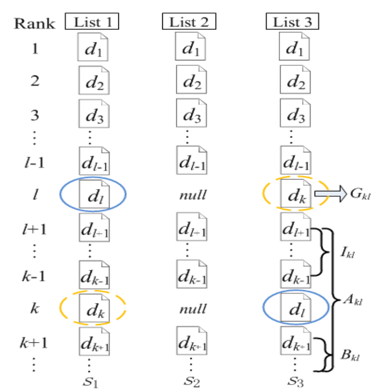

Figure 1. Three diversified ranking lists showing the pruned exhaustive strategy.

THEOREM 1. *Given* $k=l+1$, *if there exists a document pair* d_l *and* d_k *that satisfies*:

$$(G_{kl} - G_{kk}) - (G_{ll} - G_{lk}) > 0 \tag{1}$$

Next, the document list containing d_l *in its* l*-th slot and* d_k *in its* k*-th slot cannot be the optimal diversified search result.*

Proof: We prove this theorem by contradiction. Let us assume there exist d_l and d_k that satisfy Formula (1) and that a document list containing this document pair can be the optimal diversified search result. Let us also assume that List 1 shown in Figure 1 is one optimal result, where d_l is in the l-th slot and d_k is in the k-th slot. To imply contradiction, it is only necessary to exchange these two documents, which results in List 3 in Figure 1.

Compared to List 2 in Figure 1, we can compute the score of List 1 as following:

$$Sc_1 = Sc_2 + G_{ll} + G_{kk} - I_{ll} - B_{ll} - B_{kk}$$

Similarly, the score of List 3 is computed as:

$$Sc_3 = Sc_2 + G_{kl} + G_{lk} - I_{kl} - B_{kl} - B_{lk}$$

Because B_{ll}, B_{kk}, B_{kl} and B_{lk} represent the decrements derived from the documents after the k-th slot and are only a function of subtopic coverage (in diversity evaluation, the existing measures only take the current subtopic coverage into account as the influence derived from previous documents when assessing a document), we obtain: $B_{ll}+B_{kk}=B_{kl}+B_{lk}$. $k=l+1$ means there is no document between the l-th slot and the k-th slot. Therefore, we get $I_{kl}=0$. Then we subtract Sc_1 from Sc_3:

$$Sc_3 - Sc_1 = G_{kl} + G_{lk} - I_{kl} - G_{ll} - G_{kk} + I_{ll} > I_{ll} \geq 0 \qquad (2)$$

Formula (2) shows $Sc_3>Sc_1$, which means that List 3 is a better result than List 1. This is in contradiction to the assumption that List 1 is one of the optimal results. Therefore, lists containing document pairs that satisfy Formula (2) could not be the optimal result. ■

THEOREM 1 shows that when performing an exhaustive search, we can simultaneously determine whether Formula (1) is satisfied. If Formula (1) is satisfied, we can stop searching the current branch and continue searching the next branch. Therefore, we can propose an algorithm to prune the branches that must not achieve the optimal solution when performing the exhaustive search.

Algorithm 1. Pruning_exhaustive_search

Input *all the selected documents D, the required number of*

$S \leftarrow \Phi$, $maxG \leftarrow 0$

function ***recursion_full_search***(*curD*, *leftD*, d_i, *curG*)

    if (*leftD* is Φ or $|curD| = L$) and $curG > maxG$

        $maxG \leftarrow curG$

        $S \leftarrow curD$

    else

        $n \leftarrow |curD|$

        foreach d_j in *leftD*

            if $(G_{in} - G_{i(n+1)}) - (G_{jn} - G_{j(n+1)}) \geq 0$

                recursion_full_search($curD \cup \{d_j\}$, $leftD / \{d_j\}$,

//end function

foreach d_i in *D*

    recursion_full_search($\{d_i\}$, $D/\{d_i\}$, d_i, G_{i1})

return *S*

3. EXPERIMENTS

3.1 Datasets

To demonstrate the effectiveness of our proposed algorithm, we collect all the subtopics submitted in the subtopic mining tasks of NTCIRs 9 and 10. Based on the metric *α-nDCG* [4], we generated the diversified search results using Algorithm 1 for each query. Altogether 200 Chinese queries and 50 English queries in the subtopic mining tasks are used. We take the subtopics submitted by different participants as different query instances because they are mined using different methods. Totally 18 Chinese runs and 29 English runs are submitted, which comprises of subtopics mined for 200 Chinese queries and 50 English queries. Therefore, in total we obtain 200*18=3600 Chinese query instances and 50*29=1450 English query instances. The experimental results are compared to the exhaustive search results. However, it is difficult to exhaustively search for the optimal result when a large number of documents are selected for exhaustive search. Therefore, we change the number of selected documents from 2 to 5 to construct the diversification experiments.

3.2 Experiment Results.

With THEOREM1 we can see that the proposed algorithm could get optimal ranking lists with respect to a given evaluation metric. By investigating into the percentages of queries from different datasets that obtain the optimal results using Algorithm 1, we find that Algorithm 1 obtains the optimal results for both the 3800 Chinese and the 1450 English queries, which means that Algorithm 1 perform lossless pruning on all the query instances.

Table 1 presents the corresponding time costs of the experiments. The values of columns 3-6 are the log(t)s with the number of selected documents changing from 2 to 5. It shows that by pruning the useless search branches in the exhaustive search, Algorithm 1 can effectively decrease the time cost of the exhaustive search. The larger the number of selected documents is, the decrement of the time cost is larger.

Table 1 The logarithm of the time costs of different algorithms.

Dataset	Algorithm	2	3	4	5
Chinese Queries	Exhaustive Search	-1.56	0.63	2.86	5.08
	Algorithm 1	-2.03	-0.12	2.09	4.40
English Queries	Exhaustive Search	-1.64	0.53	2.74	4.98
	Algorithm 1	-1.63	0.26	2.44	4.50

4. CONCLUSIONS

In this paper, we propose a pruned exhaustive search algorithm for search result diversification. We prove that our algorithm can cut the useless branches of exhaustive search without losing any performance. Experimental results also show that our proposed algorithm can obtain the optimal results for all the 3800 Chinese queries and the 1450 English queries. Compared to the exhaustive search, the time cost of our algorithm is effectively decreased.

REFERENCES

[1] C. L. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. In 2nd International Conference on the Theory of Information. Cambridge, UK. pages 188-199. 2009.

[2] B. Carterette. An Analysis of NP-Completeness in Novelty and Diversity Ranking. In Proceedings of ICTIR. pages 200-211. 2009.

[3] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM, Barcelona, Spain. pages 5-14. 2009.

[4] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Blzttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In Proceedings of ACM SIGIR 2008. ACM, Singapore. pages 659-666. 2008.