A Pruning Algorithm for Optimal Diversified Search
Fei Chen1,2,3,* Yiqun Liu1,2,3, Jian Li4, Min Zhang1,2,3, and Shaoping Ma1,2,3
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
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 computation 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 any performance loss.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Search process

Keywords
Diversified search; exhaustive search; greedy search.

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 pruned 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 $d_i$ 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 $d_k$, $S_{c1}$, $S_{c2}$ and $S_{c3}$ 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 $d_l$ itself because no document exists in the $l$-th slot of List 2 before $d_l$ is added to this position. We denote the score for $d_l$ 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 $d_l$, $d_k$ will not affect the second part of the score change. If $d_k$ is relevant to any subtopic covered by any document after $d_l$, the score of the second part will decrease. We denote the absolute value of this score decrement as $I_{kl}$. 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 pruning strategy.](image)

THEOREM 1. Given $k=l+1$, if there exists a document pair $d_l$ and $d_k$ that satisfies:
$$(G_{kl} - G_{kl}) - (G_{kl} - G_{kl}) > 0$$
(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:

\[ Sc_1 = Sc_2 + G_{c_2} - G_{c_1} - I_{a_2} - B_{a_2} - B \]

Similarly, the score of List 3 is computed as:

\[ Sc_3 = Sc_2 + G_{c_2} - G_{c_1} - I_{a_2} - B_{a_2} - B \]

Because \( B_{a_2} + B_{a_4} + B_{a_3} \) and \( B_3 \) 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: \( B_{a_2} + B_{a_4} + B_{a_3} = B_3 \), \( k = l + 1 \) means there is no document between the \( l \)-th slot and the \( k \)-th slot. Therefore, we obtain \( I_{a_2} = 0 \). Then we subtract \( Sc_1 \) from \( Sc_3 \):

\[ Sc_3 - Sc_1 = G_{c_2} + G_{c_1} - I_{a_2} - G_{s_2} - G_{s_1} + I_{a_2} > 0 \quad (2) \]

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. Pruned exhaustive search

**Input** all the selected documents \( D \), the required number of documents \( L \)

1. \( S \leftarrow \Phi, \max G \leftarrow 0 \)
2. **function recursion_full_search**(curD, leftD, di, curG)
3.   if (leftD is \( \Phi \) or \( |curD| = L \)) and \( curG > \max G \)
4.     maxG \leftarrow curG
5.   \( S \leftarrow curD \)
6.   else
7.     \( n \leftarrow |curD| \)
8.     foreach \( d_i \) in leftD
9.     if \( (G_{c_1} - G_{c_2}) \geq (G_{c_2} - G_{c_1}) \)
10.    \( S \leftarrow \text{recursion_full_search}(curD \cup \{d_i\}, curG, leftD \setminus \{d_i\}) \)
11.   end function
12.   foreach \( d_i \) in \( D \)
13.   \( S \leftarrow \text{recursion_full_search}(\{d_i\}, D \setminus \{d_i\}, di, curG) \)
14. 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 \( \alpha-\text{nDCG} \) [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 respectively comprises of subtopics mined for the 200 Chinese queries and the 50 English queries. Therefore, in total we obtain 200*18=3,600 Chinese query instances and 50*29=1,450 English query instances. The experimental results are compared to the exhaustive search results. However, it is difficult to 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 diversified search experiments.

#### 3.2 Experiment Results

With Theorem 1 we can see that the proposed algorithm could obtain 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 3,800 Chinese queries and the 1,450 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. It shows that by pruning the useless search branches in the exhaustive search, Algorithm 1 decreases the time cost of the exhaustive search. The larger the number of selected documents is, the larger the decrement of the time cost is.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Queries</td>
<td>Exhaustive Search</td>
<td>-1.56</td>
<td>0.63</td>
<td>2.86</td>
<td>5.08</td>
</tr>
<tr>
<td></td>
<td>Algorithm 1</td>
<td>-2.03</td>
<td>-0.12</td>
<td>2.09</td>
<td>4.40</td>
</tr>
<tr>
<td>English Queries</td>
<td>Exhaustive Search</td>
<td>-1.64</td>
<td>0.53</td>
<td>2.74</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>Algorithm 1</td>
<td>-1.63</td>
<td>0.26</td>
<td>2.44</td>
<td>4.50</td>
</tr>
</tbody>
</table>

### 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 3,800 Chinese queries and the 1,450 English queries. The efficiency of our proposed algorithm is proved to be higher than the unpruned exhaustive search in the experimental studies.

### ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205) of China.

### REFERENCES


