An Optimization Method for Proportionally Diversifying Search Results

Lin Wu, Yang Wang, John Shepherd, and Xiang Zhao

School of Computer Science and Engineering
The University of New South Wales, Sydney, Australia
{flinw,wangy,jas,xzhaog}@cse.unsw.edu.au

Abstract. The problem of diversifying search results has attracted much attention, since diverse results can provide non-redundant information and cover multiple query-related topics. However, existing approaches typically assign equal importance to each topic. In this paper, we propose a novel method for diversification: proportionally diversifying search results. Specifically, we study the problem of returning a top-k ranked list where the number of candidates in each topic is proportional to the popularity degree of that topic with respect to the query. We obtain such a top-k proportionally diverse list by maximizing our proposed objective function and we prove that this is an NP-hard problem. We further propose a greedy heuristic to efficiently obtain a good approximate solution. To evaluate the effectiveness of our model, we also propose a novel metric based on the concept of proportionality. Extensive experimental evaluations over our proposed metric as well as standard measures demonstrate the effectiveness and efficiency of our method.

Keywords: Diversity, Optimization, Proportions.

1 Introduction

Diversification models for search results [14,3,7,23,20,22] have attracted much attention since they can effectively identify possible aspects of the query and return documents for each aspect. In many cases, this is more useful than conventional search methods which focus on finding the top-k most relevant documents, often favouring (near) duplicates in the top positions of the ranked list at the expense of topic diversity. Although methods for finding a diverse search result list have been well studied, they primarily address the problem from the perspective of *minimizing redundancy*, and promoting lists that contain documents covering multiple topics. One limitation of these approaches is that they treat each document equally while overlooking the fact that some topics are more popular than others; this can result in too much prominence being given to topics that are unlikely to be interesting to a majority of searchers. Ideally, the number of documents from each topic should reflect the popularity degree of that topic. Consider the case of recommending a set of 10 musical documents in a music recommendation system where two topics are considered (e.g., *rock*

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 390–401, 2013.

and *classical*) with 90% popularity voting for the topic *rock* and 10% for the topic *classical*. For most users, it would be more useful to return a list which included mainly results related to *rock* with less results for *classical* (e.g. 9 rock, 1 classical). Existing approaches to diversification would return roughly equal numbers of results for each topic (i.e. 5 rock, 5 classical), which is less than ideal.

Motivated by this, we aim to better solve the problem of diversification by considering it from a different viewpoint: *proportionally diversifying search results*. Specifically, we study the problem of diversifying the top-k representative search results by respecting the overall popularity degree for each topic; we acheive this by determining the number of representative documents on each topic proportional to the topic's popularity by maximizing a novel objective function. Since the computation of this objective function is NP-hard, the final proportionally representative results are obtained by using an effective greedy heuristic to approximately maximize the objective function.

We evaluate both our method and state-of-the-art approaches by conducting comparison experiments over standard metrics [7,8,6] for diversity based on redundancy penalization, and our proposed metric, which considers proportional diversification.
Our principal contributions are as follows.

- We present a novel method for diversification: proportionally diversifying search results. Specifically, a novel objective function is proposed to obtain the top-k diverse list by considering the popularity degree over each topic.
- We show that finding the optimal diversified top-k results by our objective function is NP-hard. To address that, an efficient greedy heuristic is proposed with good approximation ratio.
- A novel metric for diversity is proposed to verify our technique from the perspective of proportion. To demonstrate the efficiency and effectiveness of our approach, extensive experiments are conducted on a real-world database, which are evaluated by standard metrics and our proposed metric.

The rest of the paper is organized as follows: Related work is briefly reviewed in Section 2. We formulate the problem into a combinatorial optimization problem and show its potential to find a proportionally diverse ranking list in Section 3. We present the objective function and near-optimal algorithm in Section 4. Then, in Section 5, we report the experimental studies. Section 6 provides the conclusion.

2 Related Work

There has been rising interest in incorporating diversity into search results to meet the diverse requirements of users by both covering a large range of topics as well as preserving relevance. Standard diversification techniques [3,2,21,17,16] attempt to form a diverse ranked list by repeatedly selecting documents that are different to the selected items. One of the typical techniques is "Maximal

Marginal Relevance" (MMR) proposed by Carbonell *et al.*[3], where each candidate is iteratively selected by the MMR objective function until a given condition is met. MMR was extensively studied by Rafiei *et al.*[17], who aimed to find the best strategy for ordering the results such that many users would find their relevant documents in the top few slots by formulating a weight vector as facets of the query were discovered.

Other than the work discussed above, there are many recent works studying result diversification [22,20,1,14,19]. For instance, in [22], the authors proposed a random-walk-based approach to encourage diversity by assigning absorbing states to the items that have been selected, which effectively "drags down" the importance of similar unranked states. In a similar vein, a model based on a reinforced random-walk is proposed in [14] to automatically balance the relevance and diversity in the top returned answers. Tong *et al.*[20], propose to address diversity from an optimization viewpoint which considers relevance and diversity optimally. Although the experimental results in [20] show improved performance in terms of diversity, it is still less than ideal in applications where the awareness of proportional popularity is desirable. The work most relevant to our own is proposed in [10], where an election-based method is proposed to address the problem of diversifying searched results proportionally. The method is divided into two phases. First, it diversifies the topics of all candidates by an election-based strategy, and then the final ranked list is yielded by selecting an appropriate number of candidates for each topic. However, this method can lead to some documents in popular topics being irrelevant to the query due to the separation of topic selection and candidate selection. It aims at diversifying the topics of all candidate documents rather than the candidate documents in essence. In contrast, our technique unifies the above phases, and effectively obtains a diverse top-k ranked list taking into account both the popularity degree of each topic and the relevance of each document to the query.

In this paper, we propose a novel objective function where the final top-k proportionally diversifying search results are obtained by achieving the optimal set of the function. To the best of our knowledge, our work is the first to obtain an effective solution for proportionally diversifying search results in an optimizing environment.

3 Problem Formulation

In this section, we formulate a description of the problem of *proportionally diversifying search result* as follows. Let $Q = \{w_1, \ldots, w_n\}$ $(n \geq 1)$ be a set of keywords comprising a query, let $T = \{t_1, \ldots, t_m\}$ be the set of all m topics in the result of Q, and let $\mathcal{U}$ denote the set of all documents. We denote p_i to be the popularity degree of topic $t_i \in T (1 \leq i \leq m)$ in $\mathcal{U}$.

Definition 1. *The ranked list R is a* ***proportional representation*** *of $\mathcal{U}$ iff the number of documents in R within topic $t_i \in T (1 \leq i \leq m)$ is proportional to*

its popularity degree p_i. Suppose $N(i)$ is the number of candidates from t_i in R, then we have

$$\frac{N(i)}{\sum_{j=1}^{m} N(j)} \approx \frac{p_i}{\sum_{j=1}^{m}(p_j)} \tag{1}$$

We normally present the top-k elements of R as the result for query Q; the proportion of documents for each topic in the query result should roughly follow the the popularity degree for that topic Note that Eq.1 shows that the number of candidates for each topic in the final ranked list is not required to exactly match the proportion of the popularity degree for that topic. This is because the relevance between query and each candidate could degenerate if we strictly adhere to the precise proportions (this is demonstrated in section 5).

Example 1. Consider a document collection $\mathcal{U}$ where we assume that 80% of the documents in $\mathcal{U}$ about the "Apple" computer company and 20% are about the fruit "apple". In this case, $\mathcal{U}$ is associated with two topics. Let $R = [R_1, R_2]$ where R_1 denotes the set of documents about "Apple" (the company) and R_2 is the set of documents about "apple" (the fruit). In a top-10 ranked result list for the query "apple", we would expect roughly 8 results from R_1 and 2 results from R_2.

Challenges. There are two challenges to be solved in our framework. The first challenge is how to design an objective set function where the optimal or near-optimal set can best describe the proportionally diverse ranked list, which is proportionally representative of the document set. The second challenge is developing an effectiveness measure; that is, given a proportionally ranking list, how to quantify its goodness. To solve the above two challenges, we propose a novel objective set function as well as a metric, both of which are shown in section 4.

4 Proportionally Diversifying Search Results

In this section, we first introduce the preliminaries and then describe our novel objective set function to obtain the top-k ranked list proportionally to the popularity degree of each topic.

4.1 Preliminaries

As our diversification algorithm is developed based on the availability of pair-wise similarities between documents, we adopt the personalized PageRank technique to compute the values [11]. Suppose there are n documents in the database and q_i is ith query. We represent q_i by a $n \times 1$ vector $\mathbf{q}_i$ such that $\mathbf{q}_i(i) = 1$ and $\mathbf{q}_i(j) = 0$ $(j \neq i)$. The pair-wise similarities from each document d_j (for $1 \leq j \leq n$) to the query d_i (i.e., q_i) can be precalculated by Eq.(2) below and are denoted by a $n \times 1$ vector $\mathbf{r}_i$.

$$\mathbf{r}_i = c\mathbf{P}'\mathbf{r}_i + (1-c)\mathbf{q}_i \tag{2}$$

$\mathbf{P}$ is the row-normalized adjacency matrix (i.e. $\sum_{j=1}^{n} P(i,j) = 1$) of the similarities, $\mathbf{P}'$ is the transpose of $\mathbf{P}$, c is a damping factor, and $\mathbf{r}_i(j)$ is the similarity of j to i. Note that $\mathbf{r}_i(j)$ is not necessarily equal to $\mathbf{r}_j(i)$. For each pre-computed $n \times 1$ vector $\mathbf{r}_i$, we use $\|\mathbf{r}_i\|$ to denote the sum of all elements in $\mathbf{r}_i$ except $\mathbf{r}_i(i)$; that is, $\|\mathbf{r}_i\| = \sum_{j=1, j\neq i}^{n} \mathbf{r}_i(j)$.

4.2 Objective Function

Given a set R of k documents, we propose to measure the quality of R, regarding the relevance to a given query q_{i_0} and the proportional diversity based on the topic popularity, as follows.

$$g(R) = \sum_{i \in R} (1 - \frac{w_i}{\|\mathbf{r}_i\|} \sum_{j \in R, j \neq i} \mathbf{r}_i(j)) \mathbf{r}_{i_0}(i) \tag{3}$$

where $\mathbf{r}_{i_0}(i)$ is a relevance score; the more relevant each individual document d_i is to the query, the higher the value of $g(R)$. Nevertheless, the inclusion of d_i in R is penalized against the pair-wise similarities ($\mathbf{r}_i(j)$) from document d_i to other documents d_j in R; that is, subtract $\frac{w_i}{\|\mathbf{r}_i\|} \sum_{j \in R, j \neq i} \mathbf{r}_i(j)$ $(0 \leq w_i \leq 1)$ where $\mathbf{r}_i(j)$ is large when d_i and d_j have the same topic, which will further reduce the value of $g(R)$. Thus, $g(R)$ is expected to capture simultaneously the high closeness and the great diversity by maximizing its value while confirming the proportionality. Thereby, we aim to efficiently retrieve a set R of k documents such that $g(R)$ is maximized.

The question here is how to proportionally diversify top-k results? We argue that it is implemented by w_i, which indicates the importance of discounting the pair-wise similarity to include d_i in R. Herein, w_i determines the topic to be selected; we call this the *topic coefficient*. In fact, the proportion for the number of documents in each topic is guaranteed by automatically updating the topic coefficient w_i, which manages the possibility of declining d_i provided that many items belonging to the same topic as d_i have already been included. Specifically, we define w_i as

$$w_i = e^{1 - \frac{z_i}{u_i + 1}} \tag{4}$$

where z_i denotes how many documents that belong to the same topic as d_i have been included in R and u_i is the number of documents with the topic t_i. We assume that z_i is always less then u_i in our setting. It is natural to observe that the larger w_i is, the heavier penalty on document d_i, and it becomes more difficult for d_i to be selected.

We now prove that the problem of maximizing $g(R)$ is NP-hard even when all $w_i = 1$, which is a special case of this optimization problem. To deal with this, we then propose a near-optimal algorithm with a performance guarantee (i.e., accuracy guarantee and time-complexity) regarding a general $g(R)$.

Theorem 1. *The problem of retrieving a set R of k documents to maximize $g(R)$ is NP-hard with respect to k even if all $w_i = 1$ in Eq. (3).*

Proof. We convert the decision problem of maximum clique to a special case of the decision problem of the optimization problem in Theorem 1.

Decision Problem of Maximum Clique (MC)

INSTANCE: Given a graph G with n vertices, and an integer $k \leq \frac{n}{2}$.
QUESTION: Is there a complete subgraph of G with size k?

It is well known that the maximum clique problem is NP-hard [13]; thus, its decision problem, above, is NP-complete regarding k.

Proof of Theorem 1. For each instance (i.e. G and k) in MC, we construct $\mathbf{r}$ as follows. Suppose that the query has label 0 and each vertex $v_i \in G$ corresponds to a document i. For each vertex $v_i \in G$ $(1 \leq i \leq n)$, we assign $\mathbf{r}_0(i) = \frac{1}{n}$ and $\mathbf{r}_i(0) = 0$ (note $\mathbf{r}$ is not always symmetric); clearly, $\mathbf{r}_0(0)$ should be 1. Then, for each edge $(v_i, v_j) \in G$, we assign $\mathbf{r}_i(j) = \mathbf{r}_j(i) = 0$, and for each pair of vertices v_i and v_j which are not connected by an edge in G, we assign $\mathbf{r}_i(j) = \mathbf{r}_j(i) = \frac{1}{n^2}$. Then, based on a preliminary calculation, it can be immediately verified that $g(R) \geq 1$ with $|R| = k$ if and only if the following two conditions hold:

1. R contains the query with label 0; and
2. R contains a complete subgraph, with $(k-1)$ vertices, of G. Note that the $(k-1)$ vertices correspond to the $k-1$ documents.

4.3 Efficient Near-Optimal Algorithm

Theorem 1 shows that retrieving a set of k documents to maximize $g(R)$ is NP-hard. The function $g(R)$ is *submodular* and [15] states that a greedy algorithm to maximize a submodular function has the approximation ratio $(\frac{e-1}{e})$. Our algorithm (see Algorithm 1) is a greedy algorithm, for which we increase the value of topic coefficient w_j if some documents belonging to the same topic t_j have already been included in R. The set R resists document d_j with higher value of w_j while it prefers document $d_h (1 \leq h \leq n)$ if its topic coefficient is lower. Furthermore, when the maximum number of a topic is reached (e.g. $z_i = u_i$), the corresponding topic coefficient is set to be a prohibitive value Ω. Setting $w_i = \Omega$ in Eq. (3) ensures that we reject any further documents which have topic t_i. A suitable value of Ω was determined via our empirical studies.

According to the *Proposition 4.3* in [15], the greedy algorithm of diversification has the following accuracy guarantee.

Theorem 2. *The greedy algorithm achieves an approximation ratio of* $(\frac{e-1}{e})$ *for the submodular function of diversification with proportionality constraint.*

Proof. Omitted for brevity. Refer to [15] for details.

4.4 Proposed Metric

Observing that most existing metrics measure diversity by explicitly penalizing redundancy over each returned document while maintaining relevance, we propose a novel metric that considers the proportionality on the diversified search

Algorithm 1. Diversification by Popularity-based Proportionality.

Input: $\mathbf{r}_i$ (for $1 \leq i \leq n$); k; query i_0.
Output: A list R of k documents.
set initial R as i_0;
set both of initial w_i and z_i (for $1 \leq i \leq n$) as 0;
set initial u_i (for $1 \leq i \leq n$) as ceil($\frac{1}{k}$);
for *looper=1:k* **do**
  choose the document d_j such that $g(R)$ is maximized;
  if $z_j \leq u_j$ **then**
    add d_j into R;
    $z_j = z_j + 1$;
    $w_i = e^{1-\frac{z_i}{u_i+1}}$;
  else
    $w_j = \Omega$;
    discard d_j;
Return R

results by extending the metric in [10]. The metric proposed in [10] considers the following principles: First, each document need not belong to just one aspect of the query; that is, a document might be related to multiple aspects. Second, selecting a document that is related to some topics which already have enough relevant documents should be evaluated better than a non-relevant document. In other words, non-relevant documents should not be evaluated as highly as over-selection. However, the metric in [10] ignores the importance of rank positions of documents. Therefore, another critical property should be added: non-relevant documents appearing at earlier rank positions should be evaluated worse than relevant documents in later positions.

Considering the above three principles and least square index (LSq) [12], which is a standard metric for measuring *dis-proportionality*, we formulate our metric as Eq.(5) for penalizing the dis-proportionality for each rank position $L(1 \leq L \leq k)$:

$$DP@L = \sum_{t_i} c_i \| \frac{u_i - v_i}{v_i} \|^2 + \frac{1}{L} \cdot Y^2 \tag{5}$$

where u_i indicates the number of documents relevant to topic t_i, v_i is the number of documents that are actually found for t_i, Y denotes the number of non-relevant documents at positions 1..L. The coefficient c_i on topic t_i is defined as follows:

$$c_i = \begin{cases} 1, u_i \geq v_i; \\ 0, \text{otherwise.} \end{cases} \tag{6}$$

We now briefly discuss how our metric satisfies the aforementioned three principles. Our metric addresses the first principle associated with metric design by penalizing a list for under-selecting ($v_i \leq u_i$) on some topics but not for over-selecting ($v_i \geq u_i$) on it. At the same time, non-relevant aspects are penalized

($Y \geq 0$) while over-selecting is not, which meets the second principle. Finally, the third principle of rank positions is implemented by considering the positions that are occupied by the non-relevant documents in the top-k diverse ranked list. To make the metric comparable across queries, we normalize the proportionality measure as follows:

$$PM@L = 1 - \frac{DP@L}{\sum_{t_i} u_i^2 + L} \tag{7}$$

In the end, the proportionality diversification metric for a ranked list R can be computed as follows:

$$PM(R) = \frac{1}{R} \sum_{L=1}^{|R|} PM@L \tag{8}$$

5 Experimental Evaluations

In this section, we conduct extensive experiments to evaluate the effectiveness and efficiency of our algorithms. The setting of experiment is introduced in section 5.1, followed by the study of parameter learning in section 5.2. Then elaborate evaluations are presented in section 5.3.

5.1 Experimental Setup

Baseline Diversity Models. We implemented the model described above, along with four other diversity models as baselines. The first diversity model is MMR [3], which has been widely considered standard in diversity literature. Another model, xQuAD [18], uses a probabilistic framework which determines how well each document satisfies each topic and outperforms many others in the task of diversitfication. The third model, proposed by Dang *et al.*[10], is referred to as Election in our experiment, and uses an election-based approach to address the problem of search result diversification. Finally, we implemented the approach of Dragon, which captures relevance and diversity in an optimized way [20].

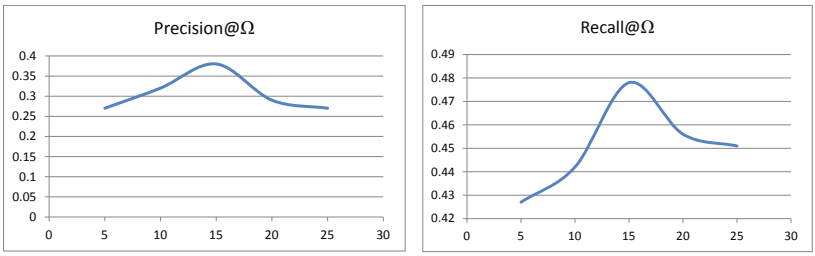
Query and Topic Collection. There are 50 queries in our query set, which come from the diversity task of the TREC 2009 Web Track [5]. To obtain the relevant documents for each query, we adopt the query-likelihood framework to conduct the relevance search [9]. The evaluation is conducted on the ClueWeb09 Category B retrieval collection [1], which contains 50 million webpages. As our approach and xQuAD require the availability of query topics and their popularity, we utilize the sub-topics provided by TREC as aspects for each query. Since the popularity of each topic is not available in TREC data, we follow the model in [18] by adopting suggestions provided by a search engine as topic representation.

[1] `http://boston.lti.cs.cmu.edu/Data/clueweb09`

Evaluation Metrics. We evaluate our approach and baseline models in terms of the proportionality metric proposed in Section 4. Considering that the proportion metric is specialized towards our model, we also report performance using several standard metrics including α-NDCG[7], ERR [4] and NRBP [8].

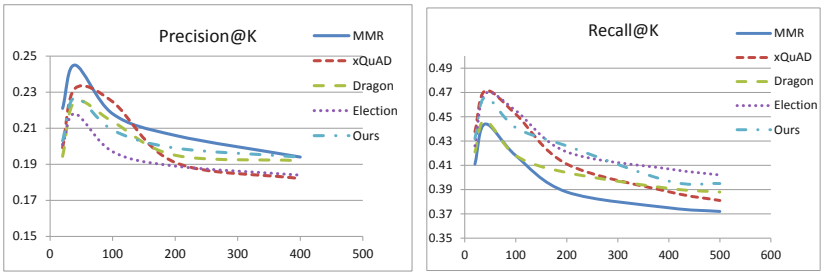
5.2 Parameter Learning

Parameter learning aims to determine "optimal" values for Ω and k. The Ω measure is specific to our approach, and we evaluate precision and recall using Ω values ranging from 5 to 25. Fig.1 shows that our model achieves the best results when Ω has the value of 15. The k measure applies to all algorithms, and we need to ensure that we do not choose a k value that is biassed towards any particular approach. Fig.2 shows that all approaches perform best with a value of k around 40. Thereby, we conduct the diversification search with a ranked list of 40 documents.



(a) (b)

Fig. 1. Parameter learning on Ω



(a) (b)

Fig. 2. Parameter learning on k

5.3 Performance Evaluations

Metrics and Proportionality Evaluations. We first compared our proposed technique to MMR, xQuAD, Dragon and Election using our proportional metric $PM(R)$ for a list of R. From Fig.3 (a), we can see that our technique outperforms the other four, which demonstrates the effectiveness of our method at preserving proportionality. Secondly, we conducted comparisons in terms of three standard metrics from the diversity literature: α-NDCG, ERR and NRBP. The results are reported in Fig.3 (b) to (d), from which we can observe the similar result as in the previous example with proportional metric. Specifically, MMR is the least effective because of its ignorance of query topics. On the other hand, our method outperforms greatly over all the other method on almost all metrics. Note that these measures are computed using top 20 documents retrieved by each model, which is consistent with the standard TREC evaluations [5].

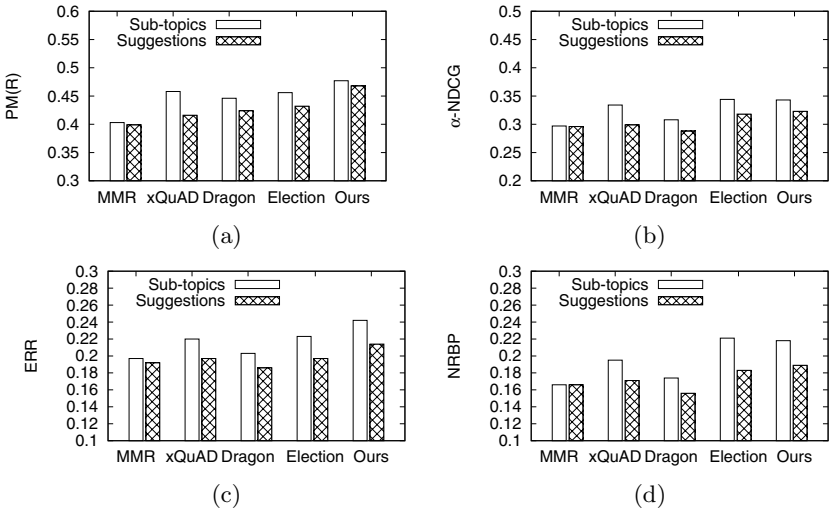


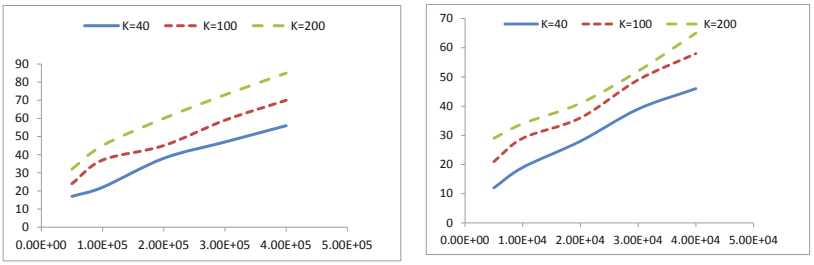
Fig. 3. Performance of diversity models in standard measures and our proposed metric

As all algorithms rank the top-k retrieved documents according to different principles, we examine the performance in both precision and sub-topic recall. We summarize the results in Table 1, which suggest that documents returned by MMR are more relevant and Election covers more topics than others. However, in terms of suggestions, i.e., the representation of popularity on retrieved documents, our model achieves better performance than the other four.

Scalability. Fig.4 gives our evaluation on the scalability of our algorithm (using synthetic data). The number of edges are fixed when we evaluate the scalability with respect to the number of nodes and vice versa. Fig.4 shows that our model

Table 1. Precision and recall for top-40 results

	Precision@40					Recall@40				
	MMR	xQuAD	Dragon	Election	Ours	MMR	xQuAD	Dragon	Election	Ours
Sub-topics	**0.2231**	0.1907	0.1775	0.2107	0.1902	0.4673	0.4724	0.4550	**0.4820**	0.4644
Suggestions	0.1801	0.1891	0.1609	0.1576	**0.2133**	0.4341	0.4410	0.4122	0.3978	**0.4522**



(a) Time w.r.t. number of nodes (b) Time w.r.t. number of edges

Fig. 4. Scalability of our model

increases linearly with respect to nodes and edges, which demonstrates that it can be applied to large-sized databases.

6 Conclusion

In this paper, we present a novel technique to address the problem of proportionally diversifying search results. A novel objective function is proposed to obtain a top-k ranked list by maximizing the value of the function. We prove that obtaining the optimal maximal value with respect to the proposed objective function is NP-hard, and resolve this by proposing an efficient greedy heuristic. We also propose a metric ($PM(R)$) to measure how effectively a diversification algorithm captures proportionality. Our experimental studies, evaluated on both standard metrics and our proposed metric, validate that our algorithm is not only able to effectively balance the relevance and diversity of search results, but is also capable of keeping approximate proportionality of the top-k search results according to the popularity degree of the various topics.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM (2009)
2. Bahmani, B., Chowdhury, A., Goel, A.: Divdb: A system for diversifying query results. In: PVLDB, pp. 1395–1398 (2011)
3. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR, pp. 335–336 (1998)

4. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM (2009)
5. Clarke, C., Craswell, N., Soboroff, I.: Overview of the trec 2009 web track. In: TREC (2009)
6. Clarke, C., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: WSDM (2011)
7. Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Buttcher, S., MacKinnon, I.: Novelty and diversity in information retreival evaluation. In: SIGIR (2008)
8. Clarke, C.L.A., Kolla, M., Vechtomova, O.: An effectiveness measure for ambiguous and underspecified queries. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 188–199. Springer, Heidelberg (2009)
9. Croft, W., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice (2009)
10. Dang, V., Croft, W.B.: Diversity by proportionality: an election-based approach to search result diversification. In: SIGIR (2012)
11. Fogaras, D., Rácz, B., Csalogány, K., Sarlós, T.: Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. Internet Mathematics 2(3), 333–358 (2005)
12. Gallagher, M.: Proportionality, disproportionality and electoral systems. Electoral Studies 10(1), 33–51 (1991)
13. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness (1979)
14. Mei, Q., Guo, J., Radev, D.: Divrank: the interplay of prestige and diversity in information networks. In: ACM SIGKDD, pp. 1009–1018 (2010)
15. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions. Mathematical Programming 14, 265–294 (1978)
16. Radlinski, F., Dumais, S.: Improving personalized web search using result diversification. In: SIGIR (2006)
17. Rafiei, D., Bharat, K., Shukia, A.: Diversifying web search using result diversification. In: WWW (2010)
18. Santos, R., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: WWW (2010)
19. Slivkins, A., Radlinski, F., Gollapudi, S.: Learning optimally diverse rankings over large document collections. In: ICML (2010)
20. Tong, H., He, J., Wen, Z., Konuru, R., Lin, C.-Y.: Diversified ranking on large graphs: An optimization viewpoint. In: ACM SIGKDD, pp. 1028–1036 (2011)
21. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: ACM SIGIR, pp. 10–17 (2003)
22. Zhu, X., Goldberg, A.B., Gael, J.V., Andrzejewski, D.: Improving diversity in ranking using absorbing random walks. In: HLT-NAACL, pp. 97–104 (2007)
23. Ziegler, C.-N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: WWW, pp. 22–32 (2005)