Result Diversification in Automatic Citation Recommendation

Onur Küçüktunç, Erik Saule, Kamer Kaya, Ümit V. Çatalyürek
The Ohio State University
kucuktunc.1@osu.edu, {esaule,kamer,umit}@bmi.osu.edu

Abstract

The increase in the number of published papers each year makes manual literature search inefficient and furthermore insufficient. Hence, automatized reference/citation recommendation have been of interest in the last 3-4 decades. Unfortunately, some of the developed approaches, such as keyword-based ones, are prone to ambiguity and synonymy. On the other hand, using the citation information does not suffer from the same problems since they do not consider textual similarity. Today, obtaining the desired information is as hard as looking for a needle in a haystack. And sometimes, we want that small haystack, e.g., a small result set containing only a few recommendations, cover all the important and relevant parts of the literature. That is, the set should be diversified enough. Here, we investigate the problem of result diversification in automatic citation recommendation. We enhance existing techniques, which were designed to recommend a set of citations with satisfactory quality and diversity, with direction-awareness to allow the users to reach either old, well-cited, well-known research papers or recent, less-known ones. We also propose some novel techniques for a better result diversification. Experimental results show that our techniques are very useful in automatic citation recommendation.

Keywords: diversity, direction awareness, automatic citation recommendation, random walks

Introduction

Automatic citation recommendation has been a popular problem since '60s. There are methods that only take local neighbors (i.e., citations and references) into account, e.g., bibliographic coupling (Kessler, 1963), cocitation (Small, 1973), and CCIDF (Lawrence, Giles, & Bollacker, 1999). Recent studies, however, employ graph-based algorithms, such as Katz (Liben-Nowell & Kleinberg, 2007), random walk with restarts, or the well-known PageRank (PR) algorithm to investigate the whole citation network. These algorithms treat the citations and references in the same way. With this approach, old and well cited papers have an advantage over the recent ones since they usually have more incoming edges in the citation graph. Hence, the graph tends to have more and shorter paths from the papers of interest (hereafter called *seed* papers) to old papers. We previously defined the class of *direction aware* algorithms which can be tuned to reach a variety of citation patterns, allowing them to match the patterns of recent or traditional documents (Kucuktunc, Saule, Kaya, & Catalyurek, 2012). Here, we enhance existing result diversification techniques with direction awareness. The proposed algorithms will be integrated to our publicly available citation recommendation system the**advisor**[1].

Result Diversification on Graphs. Diversity in ranking schemes has been discussed in various data mining fields including recommender systems (Ziegler, McNee, Konstan, & Lausen, 2005). The topic is often addressed as a multi-objective optimization problem and is NP-hard (Carterette, 2009).

Result diversification in random walk-based methods recently attracted attention. For example, GRASSHOPPER addresses diversified ranking on graphs by vertex selection with absorbing random walks (Zhu, Goldberg, Gael, & Andrzejewski, 2007). It greedily selects the highest ranked vertex at each step and turns it into a sink for the next steps. The algorithm has a high time complexity and does not scale to large graphs. Another algorithm, DIVRANK combines the greedy vertex-selection process with the vertex reinforced random walk (VRRW) model (Mei, Guo, & Radev, 2010). It updates the transition matrix at each iteration with respect to the current/cumulative ranks of the nodes to introduce a *rich-gets-richer* mechanism to the ranking. The shortcomings of those techniques were discussed in (Li & Yu, 2011). Tong, He, Wen, Konuru, and Lin (2011) proposes the *goodness* measure to combine relevancy and diversity, and presents an algorithm called DRAGON which produces solutions of near-optimal goodness.

Acknowledgments: This work was partially supported by the U.S. Department of Energy SciDAC Grant DE-FC02-06ER2775 and NSF grants CNS-0643969, OCI-0904809 and OCI-0904802.

Küçüktunç, O., Saule, E., Kaya, K. & Çatalyürek, Ü. V. (2013). Result Diversification in Automatic Citation Recommendation. *IConference 2013 Proceedings). Workshop on Computational Scientometrics: Theory and Applications*

[1] `http://theadvisor.osu.edu/`

Problem Formulation and Proposed Methods

Let $G = (V, E)$ be a directed citation graph where $V = \{v_1, \ldots, v_n\}$ is the vertex set and E, the edge set, contains an edge (u, v) if paper u cites paper v. Let $\delta^+(u) = |\{(u, v) \in E\}|$ and $\delta^-(u) = |\{(v, u) \in E\}|$ be the number of references of and citations to paper u, respectively. We assume that the researcher has already collected a list of papers of interest. Therefore, the objective is to return papers that extend that list: given a set of m seed papers $\mathcal{Q} = \{q_1, \ldots, q_m\}$ s.t. $\mathcal{Q} \subseteq V$, and a parameter k, return top-k papers which are relevant to the ones in $\mathcal{Q}$. With the diversity objective in mind, we want to recommend papers to be not only relevant to the query set $\mathcal{Q}$, but also covering different topics around the query set.

We use the direction aware random walk with restart algorithm (DARWR) (Kucuktunc, Saule, et al., 2012) to rank the papers based on a given query. The method is iterative, and in each iteration a fraction of each paper's score is distributed towards its citations and references, proportional to κ and $1 - \kappa$, respectively, where κ is the direction awareness parameter. A diversified result set is expected to be somewhat different than the top-k relevant set. Because the highly ranked nodes increase the ranks of their neighbors (Mei et al., 2010), the top-k results, recommended by the original DARWR (Kucuktunc, Saule, et al., 2012), is not diverse enough as shown in (Radlinski & Dumais, 2006) and in our experiments.

Method 1: Diversity based on local maxima. We argue that computing the vertices which are locally maximum and returning the k most relevant ones will guarantee that the nodes returned are recommended by taking the smoothing process of random walks into account. Once the ranks are computed, the straightforward approach for getting the local maximas is to iterate over each node in the graph and check if its rank is greater than all of its neighbors' with a $\mathcal{O}(|E|)$ algorithm. However, the algorithm runs much faster in practice since every rank comparison between two unmarked nodes (either local maxima or not) will mark one of them. The LM algorithm is given in Alg. 1 where π denotes the scores of the nodes in V, and k denotes the required number of recommendations.

Method 2: Diversity based on relaxed local maxima. The drawback of diversifying with local maximas is that for large k's (i.e., $k > 10$), the results of the recommendation algorithm are generally no longer related to the queried seed papers. Popular papers in unrelated fields can be returned, e.g., a set of well-cited physics papers for a computer science related query. Although this might improve the diversity, it hurts the relevancy, hence, the results will no longer be useful to the user.

In order to keep the results within reasonable relevancy to the query and the diversify them, we relax the algorithm by incrementally getting local maximas within the top-γk results until $|S| = k$, and removing the selected vertices from the subgraph for the next local maxima selection. We refer to this algorithm as parameterized relaxed local maxima (γ-RLM). Note that 1-RLM reduces to DARWR and ∞-RLM reduces to LM. The outline of the algorithm is given in Alg. 2. In the experiments, we select $\gamma = k$ and refer to this algorithm as k-RLM. However, we will devise experiments to see the effects of γ with respect to different measures.

ALGORITHM 1: Diversify with local maximas (LM)

Input: $G' = (V, E')$, π, k
Output: A list of recommendations S
$L \leftarrow$ empty list of (v, π_v)
for each $v \in V$ **do** $lm[v] \leftarrow$ LOCALMAX
for each $v \in V$ **do**
  if $lm[v] =$LOCALMAX **then**
    for each $v' \in adj[v]$ **do**
      if $\pi_{v'} < \pi_v$ **then** $lm[v'] \leftarrow$ NOTLOCALMAX
      else $lm[v] \leftarrow$ NOTLOCALMAX; **break**
    if $lm[v] =$LOCALMAX **then** $L \leftarrow L \cup \{(v, \pi_v)\}$
SORT(L) w.r.t π_i non-increasing
return $S \leftarrow L[1..k].v$, i.e., top-k vertices

ALGORITHM 2: Diversify with relaxed LMs (γ-RLM)

Input: $G' = (V, E')$, π, k, γ: relaxation parameter
Output: A list of recommendations S
$T \leftarrow$ SORT(V) w.r.t. π_i non-increasing
$R \leftarrow T[1 : \gamma k]$
while $|S| < k$ **do**
  $R' \leftarrow$ FINDLOCALMAXIMAS(G, R, π)
  if $|R'| > k - |S|$ **then**
    SORT(R') w.r.t. π_i non-increasing
    $R' \leftarrow R'[1 : (k - |S|)]$
  $S \leftarrow S \cup R'$
  $R \leftarrow R \setminus R'$
return S

Experiments

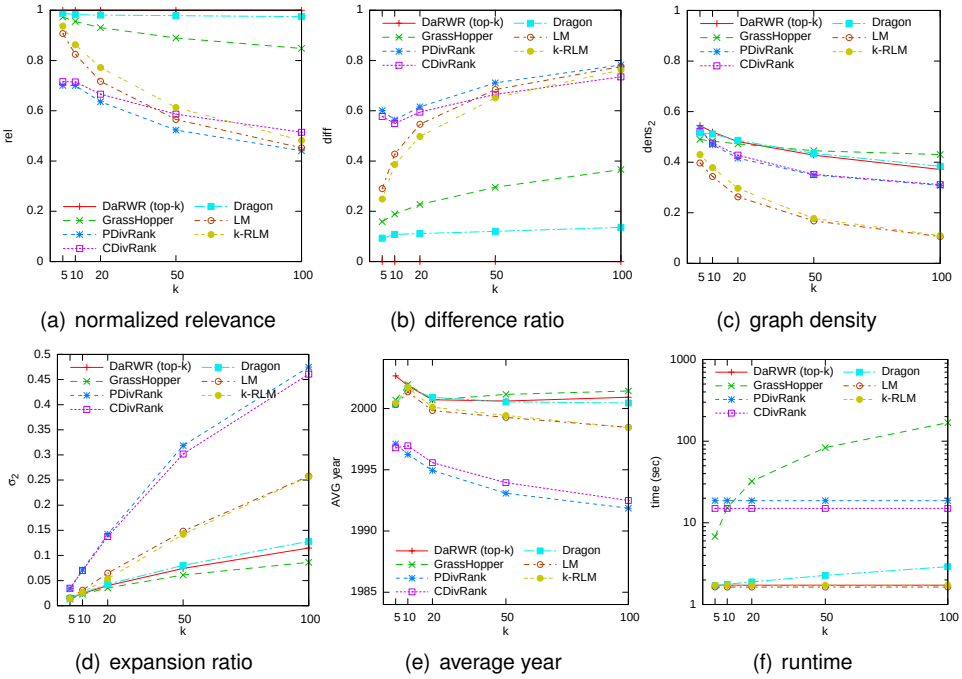
Dataset. We retrieved information about over 3 million physics, mathematics, and computer science articles from DBLP (http://dblp.uni-trier.de), arXiv (http://arxiv.org), and HAL-Inria (http://hal.inria.fr) open access library. Data gathered from CiteSeer$^{\chi}$ (Giles, Bollacker, & Lawrence, 1998) (http://citeseerx.ist.psu.edu) are also used to increase the number of paper-to-paper relations of computer science publications. We mapped each CiteSeer$^{\chi}$ document to at most one document with the title information (using an inverted index on title words and Levenshtein distance) and publication years. We then merged the papers and their corresponding metadata from these datasets. Papers without any references or incoming citations are discarded. The final graph has about 1 million unique papers and 6 million references, and is currently being used in our service.

Queries. The query set is composed of actual queries submitted to the**advisor** service. We selected about 240 queries where each query is a set $\mathcal{Q}$ of paper ids obtained from the bibliography files submitted by the users of the service who agreed to donating their queries for research purposes. $|\mathcal{Q}|$ varies between 1 and 130, with an average of 24.35.

Relevance measures. For both diversity and relevancy parts, we evaluate the quality of the results with a number of measures. The **normalized relevance** score (*rel*) is computed by comparing the original ranking scores of the resulting set with the top-k ranking list (Tong et al., 2011), defined as $\textit{rel}(S) = (\sum_{v \in S} \pi_v) / (\sum_{i=1}^{k} \hat{\pi}_i)$, where $\hat{\pi}$ is the sorted ranks in non-increasing order. Note that the original result set has the utmost relevancy, which can mislead the evaluation since the objective is to improve the diversity of the results. We decided to measure the difference of each result set from the set of original top-k nodes. Given $\hat{S}$ to be the top-k relevant set, the **difference ratio** (*diff*) is computed with $\textit{diff}(S, \hat{S}) = 1 - (|S \cap \hat{S}|)/|S|$.

Diversity measures. For evaluating the diversity of the results, **ℓ-step graph density** ($\textit{dens}_\ell$), a variant of graph density that takes the effect of non-immediate neighbors into account, is commonly used (Tong et al., 2011). It is computed with $\textit{dens}_\ell(S) = \left(\sum_{u,v \in S, u \neq v} d_\ell(u,v)\right) / \left(|S| \times (|S|-1)\right)$, where $d_\ell(u,v) = 1$ when v is reachable from u within ℓ steps, i.e., $d(u,v) \leq \ell$, and 0 otherwise. As an alternative to density, *expansion ratio* and its variant **ℓ-expansion ratio** (Li & Yu, 2011) measure the coverage of the graph by the solution set, computed with $\sigma_\ell(S) = |N_\ell(S)|/n$, where the *$\ell$-step expansion set* is defined in (Li & Yu, 2011) as $N_\ell(S) = S \cup \{v \in (V - S) : \exists u \in S, d(u,v) \leq \ell\}$. We also compute the average publication year of the recommendation set and running times of the algorithms in order to confirm that the direction-awareness property is kept and to see whether the method is practical to be used in the service.

Results. We run the algorithms on the citation graph used in the**advisor** with varying k values ($k \in \{5, 10, 20, 50, 100\}$) and with the following parameters: α in DIVRANK is selected as 0.25 as suggested by the authors. For the DARWR ranking, we use the default settings of the service, which are $d = 0.9$ for damping factor, and $\kappa = 0.75$ to get more recommendations from recent publications. In each run, the selected algorithm gives a set of recommendations S, where $S \subseteq V$, $|S| = k$, and $S \cap \mathcal{Q} = \emptyset$. The relevancy and diversity measures are computed on S, and the average of each measure is displayed.



(a) normalized relevance (b) difference ratio (c) graph density

(d) expansion ratio (e) average year (f) runtime

Figure 1

Evaluation of the results of the algorithms under various relevance (rel, diff) and diversity ($\textit{dens}_2$, σ_2) measures with respect to top-k results. rel $= 1$ and diff $= 0$ for top-k results since we compare $\mathcal{R}$ against itself. Average publication year of the papers and the runtime of the algorithms are also displayed.

Fig. 1(a-b) shows the normalized relevance and difference ratio of the recommendations compared to top-k results. We argue that a diversity-intended algorithm should actively maximize the relevancy since the top-k results will always get the highest score, yet those have almost no value with respect to diversity. On the other hand, having a very low relevancy score would tell us that the vertices are selected randomly, having no connection to the query at all. Since the normalized relevancy does not give us a clear idea of what is expected, we compare the set difference of the results from top-k relevant recommendations. Fig. 1(b) clearly shows that DRAGON gives a result set that is only 10% to 15% different than the top-k. In other words, the results of DRAGON differs in only one element when $k = 10$.

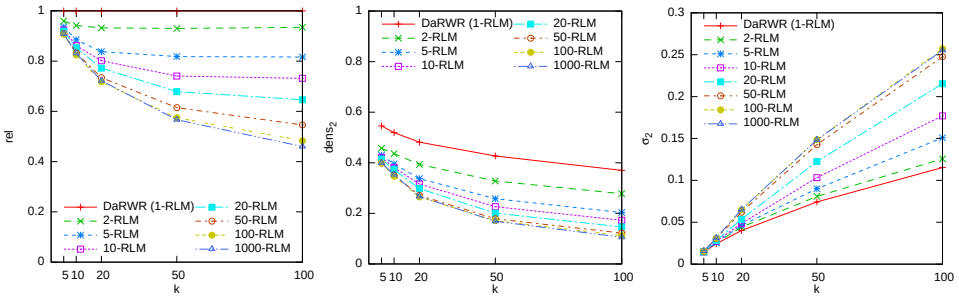
Graph density is frequently used as a diversity measure in the literature (Tong et al., 2011; Li & Yu, 2011). LM, k-RLM, and DIVRANK variants seem very promising for such a diversity objective. The same

algorithms also perform good on ℓ-step expansion ratio (see Fig. 1(d)), which is related to the coverage of the graph with the recommendations. GRASSHOPPER performs convincingly worse in these diversity metrics, meaning that its results are more dense than the results of DARWR in the citation graph.

After evaluating the results on various relevancy and diversity metrics, we are left with only a couple of methods that performed well on almost all of the measures: LM, k-RLM, and DIVRANK variants. In Fig. 1(e), however, we observe that PDIVRANK and CDIVRANK methods give a set of publications that are not recent although the κ parameter is set accordingly. Since we are searching for an effective diversification method that runs on top of DARWR, DIVRANK variants are no longer good candidates.

Efficiency. We ran the experiments on a cluster which has a 2.4GHz *AMD Opteron* CPU and 32GB of main memory. DARWR method and the dataset are also optimized based on the techniques given in (Kucuktunc, Kaya, Saule, & Catalyurek, 2012). As seen in Fig 1(f), GRASSHOPPER has the longest running time, even though it is faster than DIVRANK variants for $k \leq 10$. This behavior was also mentioned in (Mei et al., 2010). The running time of DRAGON is slightly higher than LM and k-RLM since it updates the goodness vector after finding each result. In short, query refinement-based methods (i.e., GRASSHOPPER) have linearly increasing runtimes. DIVRANK variants (PDIVRANK, CDIVRANK) requires more iterations of the random walk process, therefore, more time to converge. Finally, DRAGON, and especially LM and k-RLM are extremely efficient compared to other methods. k-RLM is slightly better than LM since it also improves the relevancy of the set to the query.

Parameter Test. The evaluations on different metrics show that γ-RLM is able to sweep through the search space between all relevant (results of DARWR) and all diverse (results of LM) with a varying γ parameter. Therefore, this parameter can be set depending on the data and/or diversity requirements of the application. We display the results of γ-RLM with varying γ parameters in Figure 2.



(a) normalized relevance (b) graph density (c) expansion ratio

Figure 2
Results of γ-RLM *with varying parameters.*

Conclusion

In this work, we addressed the diversification in citation recommendation. We enhanced the existing methods to our direction-aware problem, and proposed some new ones based on local maxima. Our experiments with various relevancy and diversity measures show that the proposed γ-RLM algorithm can be preferred for both its efficiency and effectiveness.

References

Carterette, B. (2009). An analysis of NP-completeness in novelty and diversity ranking. In *Proc. International Conference on Theory of Information Retrieval* (p. 200-211).

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). Citeseer: An automatic citation indexing system. In *Proc. ACM/IEEE Joint Conference on Digital Libraries.*

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*, 10–25.

Kucuktunc, O., Kaya, K., Saule, E., & Catalyurek, U. V. (2012). Fast recommendation on bibliographic networks. In *Proc. IEEE/ACM International Conference on Social Networks Analysis and Mining.*

Kucuktunc, O., Saule, E., Kaya, K., & Catalyurek, U. V. (2012). Direction awareness in citation recommendation. In *Proc. International Workshop on Ranking in Databases (DBRank'12) in conjunction with VLDB'12.*

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, *32*, 67–71.

Li, R.-H., & Yu, J. (2011). Scalable diversified ranking on large graphs. In *Proc. IEEE International Conference on Data Mining.*

Liben-Nowell, D., & Kleinberg, J. M. (2007). The link-prediction problem for social networks. *JASIST*, *58*(7), 1019–1031.

Mei, Q., Guo, J., & Radev, D. (2010). DivRank: the interplay of prestige and diversity in information networks. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Radlinski, F., & Dumais, S. (2006). Improving personalized web search using result diversification. In *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 691–692).

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Tong, H., He, J., Wen, Z., Konuru, R., & Lin, C.-Y. (2011). Diversified ranking on large graphs: an optimization viewpoint. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1028–1036).

Zhu, X., Goldberg, A. B., Gael, J. V., & Andrzejewski, D. (2007). Improving diversity in ranking using absorbing random walks. In *Proc. of HLT-NAACL.*

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proc. International Conference on World Wide Web* (pp. 22–32).