Supervised Learning for Automated Literature Review

Jason Portenoy[1][0000−0002−3340−2597] and Jevin D. West[1][0000−0002−4118−0322]

University of Washington, Seattle, WA 98105, USA

Abstract. Automated methods to collect papers for literature reviews have the potential to save time and provide new insights. However, a lack of labeled ground-truth data has made it difficult to develop and evaluate these methods. We propose a framework to use the reference lists from existing review papers as labeled data to train supervised classifiers, allowing for experimentation and testing of models and features at a large scale. We demonstrate our method by training classifiers using both citation- and text-based features on 654 review papers. We also demonstrate how this method may be extended to generate a novel review collection for a newly emerging research field.

Keywords: Citation Networks · Scholarly recommendation · Big Scholarly Data · Autoreview

1 Introduction and Background

Conducting a literature review, or survey, is an important part of research. The vast and exponentially growing body of literature makes it increasingly difficult to identify even a slice of the relevant papers for a given topic [12]. The advent of Big Scholarly Data—the availability of data around published research and the techniques and resources to process it—has led to a flurry of activity in finding automated ways to help with this problem.

Many methods have been developed to recommend relevant papers, using features related to textual similarity, keywords, and structural information such as relatedness in a citation network [2]. However, a common problem in developing and evaluating these methods is a lack of ground truth. In this paper, we present an approach to this problem that leverages the references in existing review papers as an approximation to ground truth. Using this abundant labeled data, we are able to frame the collection of a literature survey as a supervised learning problem. In this paper, we derive features from citation clustering and textual similarity of paper titles, but any set of related features (authors, disciplines, etc.) could be incorporated.

We begin by developing methods using the citation list from a single review article as a benchmark (section 3.1). We then show how this method can be applied to a large number of review articles (section 3.2). Finally, we apply these methods as a case study to the emerging field of misinformation studies (section 3.3). We make code and sample data for this project available at https://github.com/h1-the-swan/autoreview.

There have been several previous attempts at automated or semi-automated literature surveys. These approaches have tended to be smaller scale and rely on more qualitative means of evaluations, which are difficult to replicate and compare across studies. For example, Chen [4] developed a system to aid in writing literature reviews, which was evaluated by helping graduate students in their first year of study write and submit papers. A high acceptance rate was reported for these papers, and one student won a best paper award. This evaluation approach, while creative and compelling, does not scale well. Another study acknowledged that alternative approaches "such as those based on supervised learning need the input of annotated corpus ... not commonly available in scientific datasets" [10]. Our approach is an attempt to address this gap by using the considerable body of existing literature reviews as labeled data.

We are aware of two previous attempts that use review articles to test an automated literature review system. Belter used a semi-automated technique to retrieve documents for systematic reviews using citations [3]. Sarol et al. extended Belter's approach to include text-based filtering and additional automation [9]. These studies used a small number of hand-selected systematic review articles. In addition to methodological differences in how we utilize citation structure (e.g., our use of clustering algorithms to provide information about paper relatedness), our experimental approach automates the selection of review papers and allows for a much larger pool of labeled data. Although we share a core idea with this previous work, these differences in implementation mean direct parallels cannot be drawn.

A related problem to the one of identifying papers for surveys is the recommendation of scholarly papers. This topic has been extensively studied; a recent survey paper on research paper recommender systems [2] identified more than 200 articles on the topic published since 1998. The survey notes that the majority of approaches use keywords, text snippets, or a single article as input. Our approach starts with a set of seed papers which is then expanded upon, which is generally more appropriate for literature surveys than using a single article.

2 Data and Methods

The network data used in our analysis came from an October, 2017 snapshot of the Microsoft Academic Graph, an indexing service for scholarly publications consisting of 1.2 billion directed citation links between 77 million papers [11]. The data set also contains metadata relating to the papers, such as titles, abstracts, publication dates and venues, and authors.

We used Infomap to cluster the citation network [8,1]. Clustering is an unsupervised technique to identify groups of related papers in the citation network. We used this clustering information to generate features based on the connections between papers (described below).

Our procedure is presented in Fig. 1. The first step is to randomly split the papers into a set of "seed" papers and a set of "target" papers. We are imagining a researcher who is starting with a set of papers relating to a topic.

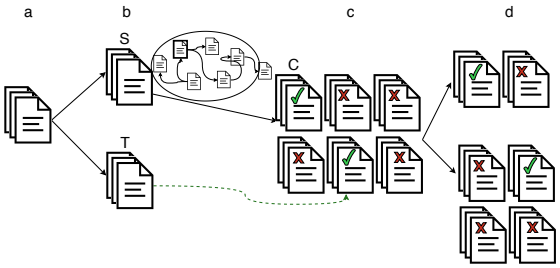

Fig. 1. Schematic of the framework used to collect data for development and testing of a supervised literature review classifier. (a) Start with an initial set of articles (i.e., the bibliography of an existing review article). (b) Split this set into seed papers (S) and target papers (T). (c) Collect a large set of candidate papers (C) from the seed papers by collecting in- and out-citations, two degrees out. Label these papers as positive or negative based on whether they are among the target papers (T). (d) Split the candidate papers into a training set and a test set to build a supervised classifier, with features based on similarity to the seed papers (S).

This researcher wants to expand this set to find the other relevant and important papers in the topic. Ideally, we would like to search for these target papers within the total set of papers in our data set. However, it is infeasible to generate features and train models using the total set of 77 million papers. To narrow the total set to a more reasonable number of candidate papers, we collect all of the papers that have either cited or been cited by the seed papers. We then go one more degree out, taking all of the papers that have cited or been cited by all of those. This process of following in- and out-citations imitates the recommended practice for a researcher looking for papers to include in a survey, but at a larger scale [13]. The resulting set of papers, while large (generally around 500K to 2M), is manageable enough to work with. We have found that this method, using different samples for the seed papers, reliably generates sets of papers that contain all or nearly all of the target papers. We label each candidate paper positive or negative depending on whether it is one of the target papers. The goal is to identify the positive (target) papers among the many candidate papers. At this point, we split the candidate papers into training and test sets in order to build classifiers.

Our next step is to generate features to use in a classification model. To incorporate the clustering information we have, one feature we use is the average cluster distance between a paper and the 50 seed papers. Distance for two papers i and j is defined as $(D_i + D_j - 2D_{LCA})/(D_i + D_j)$ where D_i and D_j represent the depth in the clustering tree hierarchy of i and j, and D_{LCA} represents the depth of the lowest common ancestor of the two papers' clusters. The feature for paper i is the average distance to each of the seed papers. We also use pagerank as a measure of citation-based importance [7].[1]

[1] Code and sample data available at https://github.com/h1-the-swan/autoreview

3 Preliminary Results

Table 1. Comparison of R-Precision scores in pilot study for Logistic Regression (LR) and Random Forest (RF) classifiers for five random splits of the data into seed and target sets, using only network-based features (average cluster distance and pagerank), or network features + text features from paper titles

		Network Features		Network + Text	
Seed	Num Candidates	LR	RF	LR	RF
1	598,117	0.378	0.297	0.196	0.612
2	1,209,241	0.403	0.322	0.227	0.607
3	804,110	0.421	0.297	0.237	0.579
4	1,604,360	0.388	0.302	0.181	0.559
5	1,432,785	0.426	0.312	0.199	0.537
avg	1,129,722	**0.403**	0.306	0.208	**0.579**

3.1 Pilot Study

For our initial pass at this problem, we used a review article on community detection in graphs [5]. We chose this paper because we are familiar with it, and believe it to be a good review of a specific topic with a large number of references. This paper cites 447 papers in its bibliography; we randomly sampled 50 of these to get our set of "seed papers"—i.e., the small set of papers that our imagined researcher above starts with. The remaining 397 papers are "target" papers that we would like to identify.

Table 1 shows the results from five splits, each using a different random seed. The "random seed" is an integer that the sampler uses as a starting point; each different random seed leads to a different split of the initial set of papers into seed and target sets. For each run, we split the 447 papers into a set of 50 seed papers and 397 target papers. After collecting candidate papers, we cleaned the data by removing the seed papers, papers for which we did not have titles, and papers published after the year the review paper was published (2010). Each seed (i.e., each row of Table 1) represents one instance of the process in Fig. 1. We report the number of candidate papers in the final set for each run. These sets of candidate papers range in size from 600K to 1.6M papers. In each case, only 397 of these papers are in the positive class. This parallels the experience of a researcher trying to do an effective survey of a topic—the goal is to find the right papers in a vast sea of literature.

We report the performance of the models as the *R-Precision*, the fraction of target papers found in the top N papers, where N is the total number of target papers—397 in this case [6]. Using two network-based features—the average distance between a paper's cluster and those of the seed papers, and the pagerank score—a logistic regression classifier identified on average 160 of the target papers (40.3%). We also ran the same experiments using a simple text-based feature: the average cosine-similarity of the TF-IDF vector of the paper title to those of the seed paper titles. Including this feature hurt the performance of the Logistic Regression model, but increased considerably the performance of the Random

Forest model. The latter identified on average 230 of the target papers (57.9%).[2] In the Appendix, we include some examples of papers ranked by the classifier.

3.2 Larger-scale study on multiple review papers

Our next step was to apply these same methods to more review papers. In order to identify a set of review papers from which we could pull bibliographies, we turned to the Web Of Science (WoS), which identifies review articles in its citation index data. In order to smoothly apply the same method as above, we limited our sample of review papers to those that could easily be linked to the Microsoft Academic Graph using a Document Object Identifier (DOI). We tested a sample of 648 review articles, choosing papers with the largest bibliographies in order to limit artifacts from insufficient input data.

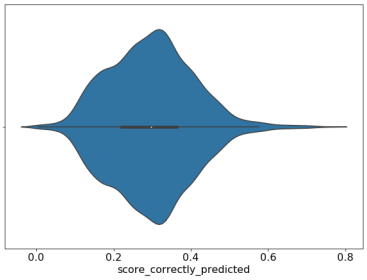

Fig. 2. Violin plot showing the distribution of R-Precision scores for 3,240 classifiers

For each review article, we gathered the cited papers from MAG, and trained models for 5 different random seeds, representing 5 different splits of the data into seed and target papers. We chose the best-performing model for each split—invariably a random forest classifier using the network and title-text features described above. Fig. 2 shows the distribution of *R-Precision* scores (number of correctly predicted target papers divided by total number of target papers) for 3,259 classifiers, each trained and tested on one of the 648 review articles. The average score was 0.30 (standard deviation 0.11); the highest score was 0.76.

3.3 Exploring an emerging field using automated literature review

The method we introduce can be adapted as a tool for exploring key papers in an emerging field. In this use case, it is the papers the classifier "misses" that we are interested in. The classifier, attempting to predict the target papers, assigns a confidence score to each of the candidate papers. We are interested in those candidate papers which received a high score, yet were not actually target papers. In the classic classification task, these would be considered misidentified,

[2] Machine learning experiments were conducted using scikit-learn version 0.19.1 running on Python 3.5.2. Trying a variety of classifiers, we saw the best performance with logistic regression and random forest models.

but in this task we consider the possibility that their similarity to the seed papers may make them relevant papers for this field. This is consistent with Belter's suggestion of "supplement[ing] the traditional method by identifying relevant publications not retrieved through traditional search techniques" [3]. As a case study, we applied this method to papers in the emerging field of misinformation studies, which pulls research from psychology, risk assessment, science communication, computer science, and others.

As part of this case study and in collaboration with the National Academy of Sciences, we curated a collection of important papers in this field[3] and used this collection as a seed set to identify other related papers that might have been missed by our more manual methods. Evaluating these results brings us back to shaky territory where we do not have ground truth. However, conversations with domain experts interested in formally characterizing these fields have been encouraging, suggesting the utility of these methods in identifying relevant papers.

4 Discussion

Our preliminary results suggest that it is possible using these automated methods to identify many of the most relevant papers for a literature review from a large set of candidate papers. We believe that, by trying new features and tuning model parameters, we can increase performance and learn more about what distinguishes these papers. We have also seen promise in using these methods to build novel surveys of topics from a set of seed papers.

Furthermore, we see potential in using this framework to develop and evaluate methods for literature survey generation and related problems such as scholarly recommendation and field identification. The objective we propose for our modeling task—accurately finding all of the remaining references from a review paper given a held out sample of seed papers from those references—is not a perfect one. We assume that the references in a review paper represent domain experts' best attempt to collect the relevant literature in a single research topic; however, there exist several different types of review article (systematic review, meta-analysis, broad literature survey, etc.), and our current method ignores potential nuance between them. Additionally, we assume that every article in a review paper's bibliography is a relevant article to be included in a field's survey; in practice, an article can be cited for many different reasons, even within a review article. Despite these limitations, the large amount of available data allows our framework to provide a means of experimenting with and developing methods for automated literature surveys. There are many review articles similar to the ones we used that have their bibliographies available and so it will be possible to do this development and analysis on a large scale across many domains. Using this framework, it will be possible to empirically evaluate novel features for their use in identifying papers relevant to a survey in a given topic.

[3] See Data and Methods at http://www.misinformationresearch.org for details

References

1. Bae, S.H., Halperin, D., West, J., Rosvall, M., Howe, B.: Scalable Flow-Based Community Detection for Large-Scale Network Analysis. In: 2013 IEEE 13th International Conference on Data Mining Workshops. pp. 303–310 (Dec 2013). https://doi.org/10.1109/ICDMW.2013.138
2. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries **17**(4), 305–338 (Nov 2016). https://doi.org/10.1007/s00799-015-0156-0, http://link.springer.com/10.1007/s00799-015-0156-0
3. Belter, C.W.: Citation analysis as a literature search method for systematic reviews. Journal of the Association for Information Science and Technology **67**(11), 2766–2777 (2016). https://doi.org/10.1002/asi.23605, http://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23605
4. Chen, T.T.: The development and empirical study of a literature review aiding system. Scientometrics **92**(1), 105–116 (Jul 2012). https://doi.org/10.1007/s11192-012-0728-3, https://link.springer.com/article/10.1007/s11192-012-0728-3
5. Fortunato, S.: Community detection in graphs. Physics Reports **486**(35), 75–174 (Feb 2010). https://doi.org/10.1016/j.physrep.2009.11.002, http://www.sciencedirect.com/science/article/pii/S0370157309002841, 05400
6. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, 1 edition edn. (Jul 2008)
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999), http://ilpubs.stanford.edu:8090/422
8. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008), http://www.pnas.org/content/105/4/1118.short, 01699
9. Sarol, M.J., Liu, L., Schneider, J.: Testing a Citation and Text-Based Framework for Retrieving Publications for Literature Reviews (Mar 2018), http://hdl.handle.net/2142/99900
10. Silva, F.N., Amancio, D.R., Bardosova, M., Costa, L.d.F., Oliveira, O.N.: Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics **10**(2), 487–502 (May 2016). https://doi.org/10.1016/j.joi.2016.03.008, http://www.sciencedirect.com/science/article/pii/S1751157715301966
11. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. pp. 243–246. ACM Press (2015). https://doi.org/10.1145/2740908.2742839, http://dl.acm.org/citation.cfm?doid=2740908.2742839
12. Tsafnat, G., Dunn, A., Glasziou, P., Coiera, E.: The automation of systematic reviews: Would lead to best currently available evidence at the push of a button. BMJ: British Medical Journal **346**(7891), 8–8 (2013), http://www.jstor.org/stable/23493904
13. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly **26**(2), xiii–xxiii (2002), http://www.jstor.org/stable/4132319

Appendix

Example of autoreview results

Below is a sample of results (random samples of true positives, false positives, true negatives, and false negatives) from the autoreview classifier using the references from Fortunato et al. [5]—a review on Community Detection in Graphs—with a random seed of 5. The "Rank" represents the position of the candidate paper when ordered descending by the classifier's score. Some of the false positives, while not in the original reference list, still seem to be relevant to the topic (e.g., "Clustering Algorithms"), others less so ("Handbook of Mathematical Functions"). The true negatives tend to have lower scores than the false negatives, suggesting that the assigned score does tend to predict relevant documents, even if they are below the cutoff.[4]

True Positives

Rank	Title	Year
0	Modularity and community structure in networks.	2006
32	Optimization by simulated annealing.	1983
83	An iteration method for the solution of the eigenvalue problem of linear differential and integral operators	1950
136	Maps of random walks on complex networks reveal community structure	2008
145	An efficient heuristic procedure for partitioning graphs	1970
179	Near linear time algorithm to detect community structures in large-scale networks	2007
187	Graphs over time: densification laws, shrinking diameters and possible explanations	2005
323	Evolutionary spectral clustering by incorporating temporal smoothness	2007
341	The Elements of Statistical Learning	2001
408	Community detection by signaling on complex networks	2008

[4] These results use a slightly different version of the input data than our original pilot study in section 3.1, which is why there are more target papers (411) than in the pilot study.

False Positives

Rank	Title	Year
45	Elements of information theory	1991
85	The complexity of theorem-proving procedures	1971
176	Clustering Algorithms	1975
190	The Concept and Use of Social Networks	1969
218	Fundamental statistics in psychology and education	1979
235	Line graphs of weighted networks for overlapping communities	2010
272	Some simplified NP-complete graph problems	1976
283	Quantizing for minimum distortion	1960
306	The advanced theory of statistics	1958
374	Handbook of Mathematical Functions	1966

True Negatives

Rank	Title	Year
50738	Linguistic Bayesian Networks for reasoning with subjective probabilities in forensic statistics	2003
61089	Comparison of Sensor Management Strategies for Detection and Classification.	1996
121773	Testing goodness of fit for the distribution of errors in multivariate linear models	2005
151627	Low-cost, bounded-delay multicast routing for QoS-based networks	1998
192168	4 Cross-language facilitation, repetition blindness, and the relation between language and memory: Replications of Altarriba and Soltano (1996) and support for a new theory	2002
624287	Developing visual sensing strategies through next best view planning	2009
1011214	Mis-generalization: An Explanation of Observed Mal-rules.	1984
1057264	Global fixed-priority scheduling of arbitrary-deadline sporadic task systems	2008
1099562	Facilitation Catalyst for Group Problem Solving	1989
1122428	Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder	2005

False Negatives

Rank	Title	Year
468	Community Structure in Congressional Cosponsorship Networks	2008
1029	Local method for detecting communities.	2005
1222	On Modularity - NP-Completeness and Beyond	2006
4337	A method for finding communities of related genes	2004
4394	Self-similar community structure in a network of human interactions.	2003
5742	Spectral coarse graining and synchronization in oscillator networks	2008
10726	Modular organization of cellular networks	2003
15739	Sequential algorithm for fast clique percolation	2008
46734	Categorical Data Analysis of Single Sociometric Relations	1981
1014712	Cliques, clubs and clans	1979