PaperMiner: An Assistant to Extract Information from Research Papers for Your Knowledge and Career Development

Zan Huang*
Georgia Institute of Technology
huangzan@gatech.edu

Kaiwen Luo*
Georgia Institute of Technology
kluo37@gatech.edu

Kai Li*
Georgia Institute of Technology
kaili@gatech.edu

Jiaji Liu*
Georgia Institute of Technology
jliu808@gatech.edu

Maiqi Ding*
Georgia Institute of Technology
mding41@gatech.edu

ABSTRACT

We [1] propose a question-driven approach to make use of academic paper corpus to answer self development related questions. A lot of great work has been done in last decades including the famous PageRank[18] algorithm and services like Google scholar[22], Microsoft Academic[23] and AMiner[21]. But users are still passive in receiving the feedback which may not meet personal needs. Less is more, in this work, we focus on providing concrete sample on how to actively use available data about research papers to answer questions for self-development by going through data collection, wrangling, mining and visualization phases. It should be more light-weight and reusable for users to actively extract information for their needs in academics and industry. We will test and check success via user studies and experiments. Risks combined are spending time reinventing wheels and hardness of keeping the work reusable for others. Crawling, data wrangling may take more time than expected and our data and computation resource is limited. The cost of this project is 0 since we will finish all the tasks on our computers. We plan to finish our work in two months: In week1, we do the plan and survey; In week2-week4, we will implement our code to collect data from different resources; In week5-week8, we will finish the whole application and the final report. The midterm milestone is to have data collected from arxiv[16] and dblp[12] for selected topics from certain institutes, getting data cleaned and raw statistics extracted. The final deliverable includes technical report, poster and open-source code to answer our proposed questions like what is the must-read-paper and recent trends for a given research topic, along with a guide on how to reuse it to answer users' own questions.

KEYWORDS

Data mining, network analysis, data visualization

1 INTRODUCTION

Academic papers are great resources as the corpus for knowledge discovery[10], data mining[8] and network analysis[11], with potential value on guiding personal knowledge and career development[24] besides the original purpose to share new findings. A lot of great work[12, 18, 21–23] has been done based on the corpus to study the citation network structure, topics, and trends of research. But few

*Authors contributed equally to this work.
[1] Abstract as the answer to HEILMEIER'S CATECHISM

could deliver the outcomes to wider audiences who are in need of insights to help make decisions for self-development in academics or industry. In this work, we propose the question-driven approach to show how to actively and selectively make use of the tremendous corpus for answering specific questions by going through data collection, wrangling, exploration, mining, and visualization. The work could be reused by the public for their own needs.

The corpus of academic papers is huge with hundreds of millions of publications by authors from tens of thousands of institutions published at conferences and journals on a wide range of topics[7]. With popular use of pre-print services[16] and open-access sources, we have easier access to research papers in recent years. At the same time, the number of paper published each year keeps increasing, at least for these related to artificial intelligence[3], which are playing the more and more important role in one's early career for building up personal knowledge base and personal reputation according to our experience and observation.

Meanwhile, the huge number of publications makes it hard for individuals to deal with considering the effort of reading, understanding and following trends of research. Tools[2, 5] have been developed to provide domain specific selective search services to reduce the pain by filtering out more related papers according to the given term. Other tools[4, 6] put more focus on providing categorical information on specific research domains or providing answer to specific questions.

2 OUR WORK

We hope to combine the two approaches in a lightweight approach to demonstrate actively using research paper corpus for answering user-specific questions which in this case, is our proposed questions. By data collections and integration from different sources like arxiv[16] and dblp[12], possibly including specific sites for hosting papers like [1], using both the citation [14, 17, 19] and text [9] data of papers for data mining[20] , and finally present the work by answering the proposed questions and present findings by data visualisation[13].

3 PLAN

We plan to focus on mining the hot topics and research fields in recent years, ranking the papers on those areas and mining the relationship between researchers to discover the expert group on a specific topic. [21] From these data, we can discover the changes and trends in research directions. By looking at the relationship of citations and researchers, we can establish an academic network in a particular field, so that researchers in the same field can design the research path and obtain research resources.

Based on our interests, we are in preparation to collect paper metadata like title, author, publisher, date, abstract and citation from two or more separate sources, considering dblp, arxiv as the major source of data. After data collection, we plan to do data cleaning and integration, which may take relative more time.

We then intend to implement algorithms to cluster and rank papers. Both data mining and machine learning models will be applied in our works by using the network and cluster algorithms on bibliometrics analysis in Madani's study [15] for reference.

Finally, to present our analysis, the methods we learned in our course will be applied.

Our group is well designed to assign the same workload to each team member based on our backgrounds. So far our group had completed the topic choosing, proposal writing, presentation preparation. The topic was discussed by the entire group. Zan, Jiaji and Kaiwen were responsible for writing the proposal. Kai and Maiqi were responsible for preparing the presentation.

All group members then will participate in every part of the whole projects and each part has a leader. Zan will be in charge of data collection, Kaiwen will be in charge of data cleaning and integration, Kai and Maiqi will be in charge of data mining and data modeling, Jiaji will be in charge of data visualization. The entire group will complete the report and presentation together.

REFERENCES

[1] 1987. Electronic Proceedings of the Neural Information Processing Systems Conference. http://papers.nips.cc/. Accessed: 2020-02-21.

[2] 2017. arxiv-sanity. http://arxiv-sanity.com/. Accessed: 2020-02-20.

[3] 2017. NIPS Accepted Papers Stats. https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0. Accessed: 2020-02-20.

[4] 2017. A web app for ranking computer science departments according to their research output in selective venues. https://github.com/emeryberger/CSrankings. Accessed: 2020-02-21.

[5] 2018. sotawhat. https://github.com/chiphuyen/sotawhat. Accessed: 2020-02-20.

[6] 2019. Browse State-of-the-Art. https://paperswithcode.com/sota/. Accessed: 2020-02-21.

[7] 2020. Microsoft Academic. https://academic.microsoft.com/home. Accessed: 2020-02-21.

[8] Shane Dawson, Dragan Gašević, George Siemens, and Srecko Joksimovic. 2014. Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge*. 231–240.

[9] Mark Gerstein, Michael Seringhaus, and Stanley Fields. 2007. Structured digital abstract makes text mining easy. *Nature* 447, 7141 (2007), 142–142.

[10] Zhen Guo, Zhongfei Zhang, Shenghuo Zhu, Yun Chi, and Yihong Gong. 2009. Knowledge discovery from citation networks. In *2009 Ninth IEEE international conference on data mining*. IEEE, 800–805.

[11] Norman P Hummon and Patrick Dereian. 1989. Connectivity in a citation network: The development of DNA theory. *Social networks* 11, 1 (1989), 39–63.

[12] Michael Ley. 2009. DBLP: some lessons learned. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1493–1500.

[13] Loet Leydesdorff. 2008. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology* 59, 1 (2008), 77–85.

[14] Michael H MacRoberts and Barbara R MacRoberts. 1989. Problems of citation analysis: A critical review. *Journal of the American Society for information Science* 40, 5 (1989), 342–349.

[15] Farshad Madani. 2015. 'Technology Mining'bibliometrics analysis: applying network analysis and cluster analysis. *Scientometrics* 105, 1 (2015), 323–335.

[16] Gerry McKiernan. 2000. arXiv. org: the Los Alamos National Laboratory e-print server. *International Journal on Grey Literature* (2000).

[17] Henk F Moed. 2006. *Citation analysis in research evaluation*. Vol. 9. Springer Science & Business Media.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The pagerank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[19] Linda C Smith. 1981. Citation analysis. (1981).

[20] Daria Sorokina, Johannes Gehrke, Simeon Warner, and Paul Ginsparg. 2006. Plagiarism detection in arXiv. In *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 1070–1075.

[21] Jie Tang. 2016. AMiner: Toward understanding big scholar data. In *Proceedings of the ninth ACM international conference on web search and data mining*. 467–467.

[22] Rita Vine. 2006. Google scholar. *Journal of the Medical Library Association* 94, 1 (2006), 97.

[23] Kuansan Wang, Zhihong Shen, Chi-Yuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data* 2 (2019), 45.

[24] Stephen L Wright, Michael A Jenkins-Guarnieri, and Jennifer L Murdock. 2013. Career development among first-year college students: College self-efficacy, student persistence, and academic success. *Journal of Career Development* 40, 4 (2013), 292–310.