Finding Foundations

Jack Beasley, jbeasley@stanford.edu Kristine Guo, kguo98@stanford.edu

October 18th, 2018

Abstract

The fields of scientometrics and bibliometrics seek to understand and measure science through quantitative measures of scientific output. Within these fields, determining which papers are most important and impactful within a given field of science is a key problem that requires broad domain knowledge and exhaustive research.

As an alternative to manually seeking out foundational papers, researchers have turned towards citation network analysis in order to examine the evolutionary structure of scientific knowledge throughout time. However, while literature exists on applying these methods to understanding the development of specific fields fo science, such findings are limited to specific fields and fail to apply to the broader day-to-day inquiries a researcher has.

To bridge this gap between research and practice, we seek to create a recommendation system that can tell researchers not simply which papers are related to the current selected paper, but which papers contribute the most important knowledge to the paper. We seek to apply methods based on main path analysis to find articles central to the development of the paper at hand, apply those methods to a large citation dataset that draws from many fields and do so in an efficient manner such that recommendations could feasibly be returned to a user interactively.

Literature Review

Single-Score Methods

While reviewing the literature, we found many different methods of determining node importance in a graph. Specifically for paper and journal importance applications, Clarivate Analytics has published journal importance statistics based on two distinct measures: the Journal Impact Factor, which is mostly based on citation counts[1] and the Eigenfactor score,[2] which is essentially a modified PageRank algorithm for citation networks rather than web networks.

While these methods provide effective rankings given the properties they attempt to rank by, these systems do not provide value to the researcher in her day to day work of reviewing literature, as they provide nothing but a single score number that can be hard to interpret. Because these scores distill complex graph phenomena into a single number, they offer little help in determining the actual relationships between articles and understanding the larger development of science. Thus, single-score methods do not effectively help researchers with the problem of determining what the foundations of a paper are.

Because of this lack of information from single-score methods, we seek algorithms that preserve the citation network graph structure. Thus, our research pointed us towards the usage of main path analysis, which we will explore more in the next three papers.

Combining mapping and citation network analysis for a better understanding of the scientific development[3]

Calero-Medina and Noyons combine bibliometric mapping and citation network analysis in order to investigate the development of scientific knowledge about Absorptive Capacity, a term coined in 1988 that has had widespread influence on the field of Organization.

For citation network analysis in particular, they utilize two different methods: 1) main path analysis, and 2) hubs and authorities analysis. Main path analysis identifies the nodes that are most frequently used in "walks" from the most recent citations to the oldest. By computing all such possible paths, we can discover the papers that are more frequently encountered throughout time, pointing towards their centrality in the development of an academic specialization. This technique is combined with information gained from using hubs and authorities analysis, which identifies

[1] "Impact Factor - Clarivate."

[2] Bergstrom, West, and Wiseman, "The Eigenfactor™ Metrics."

[3] Calero-Medina and Noyons, "Combining Mapping and Citation Network Analysis for a Better Understanding of the Scientific Development."

papers that are both cited by other prominent papers as well as cite important papers themselves.

By combining these different perspectives, Calero-Medina and Noyons successfully identify 15 papers that comprise the main path component of the Absorptive Capacity field. Thus, this paper provides inspiration for using main path analysis to identify foundational papers in combination with hubs and authorities which actually ranks and scores the papers.

An integrated approach for main path analysis: Development of the Hirsch index as an example[4]

Liu and Lu begin by critiquing the technique of main path analysis. The original main path analysis only identifies a single main path, which is not representative of larger scientific networks that often have multiple main paths. Furthermore, the original algorithm greedily constructs the main path by repeatedly selecting the link with the highest search path count (SPC). However, as with many greedy algorithms, this algorithm is not guaranteed to produce the path with the largest cumulative SPC or contain the link with the largest SPC.

Therefore, Liu and Lu propose new variations on main path analysis. For example, global main path analysis aims to find the path with the true overall largest SPC. Another is multiple main path analysis, which identifies multiple local main paths by relaxing the search constraints to reveal more detailed information. Finally, key-route main path analysis guarantees that the link with the highest SPC is included by beginning the search from both ends of the link instead of the source nodes. Importantly, all of these methods can be combined as well.

Thus, the authors next apply an integrated approach that utilizes a combination of main path analysis methods in order to examine the development of the Hirsch index. Ultimately, their results prove that the main path analyses developed by Liu and Lu enhance our capability to capture different types of information about the relationships between scientific articles.

[4]Liu and Lu, "An Integrated Approach for Main Path Analysis."

Knowledge diffusion path analysis of data quality literature: A main path analysis[5]

In this article, Xiao et al. integrate local, global, multiple-global, and key-route main path analyses to uncover knowledge diffusion paths of data quality literature. In particular, they demonstrate that each type of main path analysis reveals different yet complementary information about development trends.

For example, local and global main path analysis highlight the papers that have provided major contributions to the field. On the other hand, multiple global and the key-route main path analyses provide more complete pictures of development trends by identifying multiple paths, revealing the divergence-convergence of the citation network as it evolves throughout time.

Finally, and perhaps most importantly, Xiao et al. also provide intuitive graphical representations of main path analyses in order to both convey their nuances and allow the reader to view the interrelationships between papers. This method of presenting results in particular serves as an inspiration for our project.

Critique and Motivation for Improvement

As seen from the literature review above, citation network and main path analysis are often limited to characterizing the development structures of specific concepts and subfields such as Absorptive Capacity, the Hirsch Index, and data quality literature.

This reality proves less than ideal for researchers and academics, who are told to "stand on the shoulders of giants" but are not given any tools that they can use to efficiently peruse and explore the development of their field. For example, conducting a literature review requires the ability to determine what works constitute essential background reading for a given paper, as well as assessing works with large impact when attempting to create new innovational methods.

However, this is not an easy process, as making literature reviews is a time-consuming manual problem that consists of recursively searching through papers' citations to try to understand what actual authorities and ground truths underlie a research problem. Beyond even just academics, people with casual interest in a field should also be able to have easy access to a field's literature without having to manually search for its most foundational papers. Such tasks would benefit

[5]Xiao et al., "Knowledge Diffusion Path Analysis of Data Quality Literature."

from comprehensive knowledge of the development of the techniques and concepts under question.

Project Proposal

Problem Statement and Motivation

In this project, we seek to create a recommender system for scientific articles that are foundational to an article of interest to a researcher. By recommender system, we mean a system that takes a paper P as an input and returns a set of the k most foundational papers to P in a reasonable amount of time so researchers can in theory use the system interactively.

By foundational, we mean papers that contribute the most important ideas to the article in question. A foundational paper is one that contributes ideas that are *used* to create the new article in question, and are often found in lists of "papers that stood the test of time." Reframed, foundational papers are often called classics that lead to the creation of whole fields built on top of their results and would have lead to the greatest negative implications for the current paper if retracted or falsified.

We postulate that these foundational papers are perhaps more important to researchers and academics than the articles directly cited by a given paper when conducting literature reviews, and more generally, when attempting to gain a broad understanding the development of a concept or technique.

Our project can be broken up into three distinct phases, which we will cover in the following sections: dataset wrangling, algorithm implementation and ranking.

Data

While there are many options for citation networks, for this project we will use the 2017 snapshot of the Microsoft Academic Graph (MAG). We chose this dataset because it is available for free online when used for research purposes, and it has also been described as "the most comprehensive publicly available dataset of its kind" in a review article.[6]

However, because the dataset is very large and comprehensive relative to other citation networks, we must transform it into a workable form that we can analyze with SNAP. This will likely prove to be a significant task as the data comes in the form of 104 GB of files of JSON documents for each paper in the graph. While 104 GB (just barely) fits on a large flash drive, this dataset is too big to conveniently analyze ad-hoc on a laptop in unprocessed form. Though, we can reduce this significantly simply by excluding any metadata that does not directly represent an edge. This means reference ids and not abstracts or author names. In a brief feasibility assessment, we reduced 1.84 GB of raw entries to 475 MB of entries (a ~75% reduction in size!) by filtering non-edge metadata.

Once we get the data in a workable form, i.e. in SNAP on a laptop, we plan to do a high-level analysis of the network in a similar manner to the analysis by Broder et. al. of the web.[7] This will involve finding strongly connected components, estimating their size and getting a general idea of how the MAG compares to the web in the year 2000. If we find that the structure is similar, we will quickly gain a lot of intuition that will be useful for working on the MAG, if not, we'll still learn a lot about the general properties of the graph and how it differs from graphs we've seen in the past.

Algorithms

This step will primarily consist of evaluating and implementing various flavors of main path analysis and determining which performs the best for finding foundational papers. There are two major algorithmic decisions to make when implementing main path analysis: choosing a method of finding traversal counts and choosing a path search mechanism.

There are many different methods for computing edge traversal counts, but most yield similar results in practice. However, we found through our literature review that search path count (SPC) provided additional properties and therefore was the most preferred and widely used method.[8] For this reason, we plan to adopt SPC and focus our energy on the path search portion of the analysis.

Unlike for traversal counts, different path search methods can yield significantly different results. There are three common techniques, local search, global search and key-route search; we plan to evaluate these different approaches on our dataset and select the one that seems to work the best for our specific case of identifying foundational papers.[9]

[6] Herrmannova and Knoth, "An Analysis of the Microsoft Academic Graph."

[7] Broder et al., "Graph Structure in the Web."

[8] Xiao et al., "Knowledge Diffusion Path Analysis of Data Quality Literature."

[9] Xiao et al.

Ranking

Once we have a method we are happy with to compute main paths, we are left with the task of sifting through these paths and determining which k articles we return to the user and in what order. If main path analysis determines which city Waldo is in, our ranking step must actually find Waldo within that city.

We expect that this step will involve reaching for secondary metrics for each of these papers, like hubs and authorities[10] or PageRank,[11] to give us some context for each paper. PageRank could help us bias our ranking to important papers and hubs and authorities could bias toward authoritative papers and to or from review or hub papers, depending on which is more useful to the researcher. Because this step is a final filter that helps to ensure the researcher sees only the most relavent results, it will be somewhat qualitative in nature though extremely important. Most people only look at the first few results, so we need to make those count and thus being smart about how we select articles from the main paths is critical.

Evaluation

Because our project is a recommender system and thus tries to solve problems for researchers, we must include researchers in our evaluation. While we plan to report some rudimentary quantitative results around runtime, our project's success is based on whether or not researchers find the results of our algorithm useful. We are currently tentatively planning on running our recommender system on the papers of faculty and grad students that we know and asking them how well our recommendations match their own intuitions of where the foundational knowledge in their work came from. If our system consistently finds articles that the researchers were inspired by, that says something positive about our method and if not, we might have missed the mark. Our method also might uncover articles that are foundational to a work, but are unknown to the author of that work because they might be hidden two or three layers deep in citations from their article.

Deliverables

We intend to deliver a working recommender system and evaluation of that system as discussed in the section above. Ideally our system should be easy to run for demos. The MAG buys us some flexibility here as it is pretty comprehensive so in theory if our method works, it should work for just about any paper published before 2017.

[10] Kleinberg, "Authoritative Sources in a Hyperlinked Environment."

[11] Page et al., "The PageRank Citation Ranking."

References

Bergstrom, Carl T., Jevin D. West, and Marc A. Wiseman. "The Eigenfactor™ Metrics." *Journal of Neuroscience* 28, no. 45 (November 5, 2008): 11433–4. https://doi.org/10.1523/JNEUROSCI.0003-08.2008.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. "Graph Structure in the Web." *Computer Networks*, 2000, 12.

Calero-Medina, Clara, and Ed C. M. Noyons. "Combining Mapping and Citation Network Analysis for a Better Understanding of the Scientific Development: The Case of the Absorptive Capacity Field." *Journal of Informetrics* 2, no. 4 (October 1, 2008): 272–79. https://doi.org/10.1016/j.joi.2008.09.005.

Herrmannova, Drahomira, and Petr Knoth. "An Analysis of the Microsoft Academic Graph." *D-Lib Magazine* 22, no. 9/10 (September 2016). https://doi.org/10.1045/september2016-herrmannova.

"Impact Factor - Clarivate." Accessed October 18, 2018. https://clarivate.com/essays/impact-factor/.

Kleinberg, Jon M. "Authoritative Sources in a Hyperlinked Environment," n.d., 29.

Liu, John S., and Louis Y. Y. Lu. "An Integrated Approach for Main Path Analysis: Development of the Hirsch Index as an Example." *Journal of the American Society for Information Science and Technology* 63, no. 3 (March 1, 2012): 528–42. https://doi.org/10.1002/asi.21692.

Page, Larry, Sergey Brin, R. Motwani, and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web," 1998.

Xiao, Yu, Louis Y. Y. Lu, John S. Liu, and Zhili Zhou. "Knowledge Diffusion Path Analysis of Data Quality Literature: A Main Path Analysis." *Journal of Informetrics* 8, no. 3 (July 1, 2014): 594–605. https://doi.org/10.1016/j.joi.2014.05.001.