Citation Network Analysis for Science Surveyor

1. Introduction

1.1 Problem

It is very important to make the public better understand scientific research. Science journalists is the bridge between the two groups, but there is a problem in contextualization of journal articles in time. Science journalists have limitations that it is nearly impossible to achieve both the thorough understanding of field and the timeliness of the news report, as there are so much to read and research, including the past researches, whether a certain finding is significant and how a finding stands in the fields' consensus.

Such limitations would lead to dreadful consequences. The public might have a bad understanding of science nature and researchers if the journalist could not present the progress and findings of scientific research clearly.

1.2 Hypothesis

We believe that the dissemination of science would strongly result from the understanding of science journalists. If some system could make our findings more accessible to journalists, science researches would also become closer to the public. We want to prove that a system that clarify where a science finds stands and the impact of it in a field would be valuable to solve such problem.

1.3 Goal

We propose designing the Science Surveyor system, a tool to help science journalists contextualize the scientific literatures in a timely manner. This tool would allow journalists to submit a scientific study that they want to cover.

It would generate a map with three central dimensions, or layers: a consensus layer that would show whether the new finding is consistent with scientific consensus, a temporal layer that would show the pattern of publishing on this topic across time, a funding layer that would characterize the funding in that field. There might also be a general popularity calculated from the weighted average of the results in all three layers.

2. Relevant Research

There are a large amount of hidden information in scientific literature. First, there might be unexpected links between literatures which may generate useful hypothesis and even discovery, or we call it Swanson linking, according to Grohmann and Stegmann[1]. A famous application could be found in biomedical hypothesis from MEDLINE records[2].

Another application is generating research fronts and trends with different kinds of citing relationship, which is close to our project. According to the CiteSpace paper[3], the citation relationship evolved a lot since 1965, from co-citation clusters, article citing the same literature to fixed, time-invariant articles and co-citation network. Innovative methods like bibliographic coupling was also introduced. [4]

3. Method

3.1 Approach Overview

Science Surveyor is a web-based application to find out where scientific finds stands in its research area. We proposed to make an analysis tool covering important academic databases. We will start with open access repositories like arXiv and PLoS One, but later we would broaden our sources using API access to databases like Thomson Reuters, JSTOR, and Elsevier.

In this project, we plan to measure a published finding in four dimensions:

- The first and most important metric would be the centrality and connectivity in a citation network which is a strong indicator of the impact. Such analysis would be a great tool to find out central studies with reliable sources and reputation in similar researches.

- Another source would be information about funding. The private corporate funding could indicate biased results about the popularity of a research.

- The third source is the idea network. It would use n-gram algorithm or bag of words model to address the topic similarity and relationship.

- The last dimension is how it changes over time. Using the history information of citation and funding network by years or decades, we might find some patterns about popular areas.

We could generate a 3-dimension vector for each period of time we analysis. If we compare the literature area or a specific finding during different time, we may apply machine learning and regression analysis to find out some interesting trends from existing literatures. It would be a powerful tool for science journalist to understand the status of a certain area or work.

3.2 Compare to previous research

There are many kinds of citation relationship analysis we knew in the previous work. In the project, as we should try to keep it clear enough for journalist to understand, we decide to choose a non-direction citation relationship.

On account of the similar reason, we analyzed different centrality measures including current flow and load model, but only use metrics with pure topology relationship and score assigning, which is easier to understand, instead of some physics models.

The major advantage of our approach is the high dimension level of information. Unlike traditional research doing co-citation clustering and observe how groups evolve by time, we are observing a three-dimension vectors covered all three parts of a literature including citation, funding and ideas.

In order to keep such complexity easier enough for science journalist to understand, we are also developing language generation system transforming the network and machine learning data to daily language.

3.3 Relationship to solutions presented in this class

This is part of the project in experimental methods in the humanities by professor Dennis Tenen. Although this is a project for future journalism, which we discussed a lot in our previous classes, it is more close to the MEDLINE database query system. Journalist would enters a command about a findings or keywords and we return the popularity vector calculated using the network analysis result from funding and citation network and natural language processing result from the idea layer network, which would use bag of words learning and n-gram model mentioned in the lecture by professor Brian Roark.

3.4 Evaluation Method

The performance of such system includes several aspects. Generally whether it is user-friendly and the quality of results define wether this project would success. Traditional research about citation clustering would focus on a certain database like PubMed and compare the cluster view and time zone view of the result. However such result is a quite subjective evaluation which need human testers and evaluators.

There is also automatic measurement for efficiency and accuracy. First, the run time is a vital evaluation parameter which would define the working efficiency towards journalists. It is necessary to test how it works with different scales of the database. Secondly, the prediction accuracy about where a certain finding stands and how it might develop in the future also deserves carefully scrutinization. Basically we could use cross-validation when building a model while learning the research trends. We could also get some labeled finding/region match-ups, and try to do a validation in the research classification.

4. Experiment

4.1 Data

We use the ACL Anthology Network dataset from the University of Michigan's CLAIR Group. It has literature data from 2008 to 2013. In this experiment we mainly use the non-self author citation network data to generate the author centrality which contributes to the overall centrality evaluation of a finding.

4.2 Library

We use python as our project language. We compared between a few python tool-kit for network analysis including snap.py, graph-tool, NetworkX and igraph. According to the performance result, the python-core library like NetworkX and snap.py is 10 times slower than the C++ core library like igraph and graph-tool. As graph-tool works better at centrality measures like page rank and graph works better at degeneracy methods like k-core, considering that our core function is based on centrality measurements, we decided to choose graph-tool over igraph library.

4.3 Centrality Measures

There are a great variety of centrality measures we could choose from. There are some basic centrality measures. Degree centrality simply count the links. Closeness shows the distance to other nodes, making it a good measurement while the research are close. Betweenness measures the role a node plays as a bridge, which seems to be a good metric for connectivity. Eigenvector centrality like page rank would might be effective in the evaluation.

There are also some new approaches. Some researchers tried to avoid simply using shortest paths like betweenness centrality [5]. Percolation centrality assumes there is a infection source, which is great in some social network analysis [6]. Total communicability centrality describes the row sum of adjacency matrix, is great to describe the overall connectivity [7].

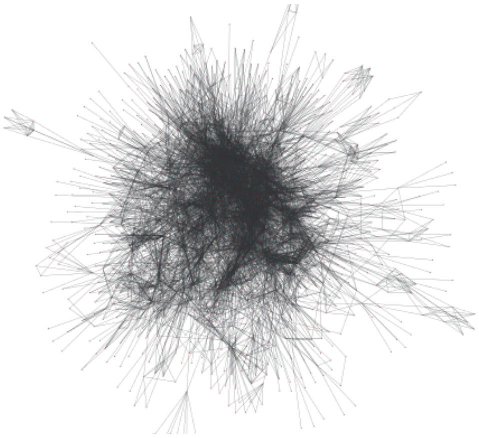

Figure.1 AAN Author Citation Network 2008 Example, SFDP layout

Considering the model compatibility to our database, with the general graph in Figure 1, degree centrality would be too simple to use. Closeness centrality would probably suffer while there are many cluster centers, so it is less efficient to calculate all the distance. The shortest path betweenness would already work and percolation is not necessary as it is not a transmission network. Total communicability might require too much math background for science journalists

In account of the accessibility to science reporters and the efficiency to work with other vector layers, we decided to choose betweenness and page rank as our experiment measurements.

4.4 Case Study

Table.1 Centrality of Michael John Collins

Centrality	2008	2009	2010
Betweenness	0.020279099	0.020908112	0.016284481
PageRank	0.003676429	0.003630617	0.003218470

Table.2 Centrality of Susan E. Brennan

Centrality	2008	2009	2010
Betweenness	0.000755152	0.000655850	0.000232821
PageRank	0.000335703	0.000292816	0.000209016

We applied both measurements to our network to see how it works with different types of nodes. For example, we choose one of the most cited author, Michael John Collins in AAN to see how the measure works. Both his betweenness centrality and page rank stayed in the first year, but dropped in the second year. We tried to compare it with a less famous author Susan E. Brennan, and find some interesting phenomenon which would help us improve the citation centrality measurement.

- Betweenness centrality have much higher value than page rank centrality in the central nodes, but they are at the similar magnitude in the outer nodes. Betweenness have more significant changes for outer nodes, shall we give it more weights?

- The influence of Collins is 80 times larger than Brennan in Betweenness criterion, while it is only around 10 times in PageRank. Which is more reasonable?

To answer this questions and find the better criterion to describe the influence, we would consult some professors in this area to improve our system.

5. Future Directions

5.1 Collaboration

As the objective of this project is to build a platform for scientist journalist and address the problems in the profession, we could contact and collaborate with a variety of scholars and experts to improve our system including scholarly publisher, science journalist, network analysts, the press and the public, and data visualization designers. We have already partnered with the Public Knowledge Project at Stanford to get a visible platform for the project.

5.2 Road Map

After we separately calculated the three-layer information, we would find a weighted average of these vectors which could best describe the popularity of a specific finding. It takes probably a few months to make all three layer works well, provided we find a efficient way to form the idea network. Otherwise it would take longer time.

Then we could add some practical functions for journalists like new item tracking and social media results to give a thorough context regarding a specific subject. This part is quite straightforward, probably a few weeks will do.

If it works well with the ANN data set, we would scale it up to larger open database like arXiv and PLoS One, and finally utilize the api of databases like Thomson Reuters, JSTOR, and Elsevier. This part would be a long term project and it is difficult to estimate the time as there might be a lot databases will different structure and api added.

6. Bibliography

[1] Grohmann, Guenter, and Johannes Stegmann. "C-MLink: a web-based tool for transitive text mining." In Proceedings of ISSI. 2005.

[2] Chen, Ran; Hongfei Lin & Zhihao Yang (2011). "Passage retrieval based hidden knowledge discovery from biomedical literature." Expert Systems with Applications: An International Journal (August, 2011), vol. 38, no. 8, pp. 9958–9964.

[3] Chen, Chaomei. “CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature.” Journal of the American Society for Information Science and Technology 57, no. 3 (February 1, 2006): 359–77. doi:10.1002/asi.20317.

[4] Nicolaisen, Jeppe, and Tove Faber Frandsen. "Consensus formation in science modeled by aggregated bibliographic coupling." Journal of Informetrics 6, no. 2 (2012): 276-284.

[5] Stephenson, Karen, and Marvin Zelen. “Rethinking Centrality: Methods and Examples.” Social Networks 11, no. 1 (March 1989): 1–37. doi:10.1016/0378-8733(89)90016-6.

[6] Piraveenan, Mahendra, Mikhail Prokopenko, and Liaquat Hossain. “Percolation Centrality: Quantifying Graph-Theoretic Impact of Nodes during Percolation in Networks.” PLoS ONE 8, no. 1 (January 22, 2013): e53095. doi:10.1371/journal.pone.0053095.

[7] Benzi, Michele, and Christine Klymko. “Total Communicability as a Centrality Measure.” Journal of Complex Networks, May 17, 2013, cnt007. doi:10.1093/comnet/cnt007.