Citation network Analysis

by: Jason Speedy

Abstract: the goal of this project was to built citation analysis tools for Blake Romero's thesis project. the thesis is proof of intellectual paradigms in anthropology could be mapped in a citation network. the tools needed where some means of generating and visualizing this citation network as well as tools that scaled the scope of the project from small scale proof of concept to the scale of the entire written body of anthropological literature. this was done with Gephi as the primary visualization, with some plug-in development for the analysis and scope scaling tools.

The primary research goal of my partner Blake, was extraordinarily large scale and with that came the problems of efficiently clustering and mapping a graph of potential size of 3000000 nodes and more edges within a reasonable time frame without the use of a supercomputer. Another problem that arose with this size of a network was collecting a dataset of proper size. these two problems were at the center of the project I worked on this semester. The latter of the two problems was actually the first concern of the project in that without data, testing would be impossible. this problem was solved by the Alaska Journal of Anthropology giving us PDFs of all the articles they have on hand. this gave us a medium sized dataset to test with and prove the basic concept of the research.

in the beginning the plan was to use a crawler to scrape anything related to anthropology and then generate a graph to trim, due to communication complications that wasn't possible and the plan was sent back to square one. the next plan was to get any citation network and see if the concept held across all disciplines since we couldn't find an anthropology one, this plan also was dropped part way through due to the AJA providing a base set for the original plan so once again we shifted gears to an anthropology focus. much of the planning stages happened in CSCE 394 where about twice a week Blake and myself would meet and update each other on current progress and direction of the project, professor Cenek was partially involved in this stage of the project as well. in totality due to the frequency of the meetings and the consistent attempts at communication between the involved parties the overall project development temple resembled a rather agile development process, but do to the primary planning changing partway into several stages much of the time was spent reorganizing priorities in the early to middle parts of the semester and much of the end was spent frantically programming.

Requirements: the formal requirements were not extraordinarily clear but in general they followed as such

- a visual way to interact with a graph
- an easy way to get citation data from a PDF
- a way to reduce the time needed to fully analyze a large scale citation network

Design: the overall design of the project took the form of several smaller parts.

- the PDF scraper was a standalone java utility that read a set of files and then outputs citation information to a designated save file.
- the graph weighting utilities took the form of a plug-in in gephi, since gephi is java based it followed a pretty simple object oriented design with nodes edges and some simple methods to work with them.
    - the SPC function and the Edge Sampling and Merging function
- the trimming functionality was also built into the same plug in

Analysis & Conclusions: As of this moment my project is incomplete, the PDF scraping tool works with a few minor errors and is sufficient for what it is going to be used for. the graph weighting and edge contracting tools are still not finished but will be shortly. I have learned a lot from this project in both computational and interpersonal arenas. the communication work was very insightful as well as building tools to add to functionality that I haven't worked with before was very enlightening. in total I feel this project did much to cement much of the different things about computer science into a more reliable and solid understanding, and with that I feel that I can call myself a competent programmer.

References:

Batagelj, V. (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, *5*(1), 187-203.

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, *57*(1), 27-57.

Naruchitparames, J., Gunes, M. H., & Louis, S. J. (2011, June). Friend recommendations in social networks using genetic algorithms and network topology. In *Evolutionary Computation (CEC), 2011 IEEE Congress on* (pp. 2207-2214). IEEE.

Nerur, S., Sikora, R., Mangalaraj, G., & Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Communications of the ACM*,*48*(11), 71-74.

Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O., & Helbing, D. (2014). Exploiting citation networks for large-scale author name disambiguation. *arXiv preprint arXiv:1401.6157*.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2011). Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological forecasting and social change*, *78*(2), 274-282.

Vallada, E., & Ruiz, R. (2010). Genetic algorithms with path relinking for the minimum tardiness permutation flowshop problem. *Omega*, *38*(1), 57-67.

Xhignesse, L. V., & Osgood, C. E. (1967). BIBLIOGRAPHICAL CITATION CHARACTERISTICS OF THE PSYCHOLOGICAL JOURNAL NETWORK IN 1950 AND IN 1960.

Zio, E., Golea, L. R., & Rocco S, C. M. (2012). Identifying groups of critical edges in a realistic electrical network by multi-objective genetic algorithms.*Reliability Engineering & System Safety*, *99*, 172-177.

Zio, E., Golea, L. R., & Rocco S, C. M. (2012). Identifying groups of critical edges in a realistic electrical network by multi-objective genetic algorithms.*Reliability Engineering & System Safety*, *99*, 172-177.