Data and Visual Analytics (CSE 6242)
Project Final Report
A System for Recommendation of Scholarly Articles to Newbies

Payam Siyari[*], Ari Kapusta[†], Francis Gallego[‡], Kangqi Ni[§], Samaneh Ebrahimi[¶]

April 24, 2016

1 Introduction - Motivation

The standard approach to finding important papers in a field is to manually search relevant keywords, to browse the citation network, and to read many papers. This process is tedious and takes much time. A focus on papers with high citation count can help reduce the number of background papers to read, but a more automated and systematic approach may be possible. The citation count may not necessarily reflect the importance of the paper. Researchers and Academia wishing to perform background reading on a new area of interest, particularly in multidisciplinary research would be particularly interested in a solution to this problem.

We propose a system for suggesting seminal or highly important papers to new users, using citation data. Our system is web-based and end-to-end, allowing non-experts in a field to quickly and easily find a short list of key papers within a field from key words. Our work uses centrality measures of the citations network and matching keywords for recognizing similar and influential papers. We include methods for selecting starting points of the search by keyword, visualizations of the resultant graph and its measures, and a summary of the most relevant important papers to give the user control and intuition of the results.

2 Problem Definition

We are trying to solve the problem of what papers should a person read to get a background in a new field. This problem is encountered particularly frequently in multidisciplinary fields

[*]payamsiyari@gatech.edu
[†]akapusta@gatech.edu
[‡]francis.b.gallego@gmail.com
[§]vincent.nkq@gatech.edu
[¶]samaneh.ebrahimi@gatech.edu

of research, or otherwise when wanting to expand one's knowledge into a new field. A solution to the problem would be a short list of papers that represent the most important, seminal papers in the field.

3 Survey

The specific feature that we are after lies at the intersection of many areas.

3.1 Sense Making

Various works have suggested methods for making sense of networks for a goal of user interpretation [11]. [4] introduced Apolo, a system that combines visualization and user interaction with a machine learning element (i.e. belief propagation) to allow a user to explore a citation network and create meaningful visual representations.

[1] presents OPAvion, a 3-module graph mining system: Pegasus to perform off-line computation on massive graphs, OddBall to find patterns and anomalies in graphs, and Apolo to allow users to interact with and explore the graph.

3.2 Recommender Systems

A huge body of the research in analyzing citation networks focus on recommendation of related papers to a set of input papers. [8] looks at the general issues and techniques for recommending scholarly articles, and concludes that scholarly recommendation is a unique problem since compared to other domains, scholarly articles are substantially denser and more subtle, and their usage patterns are more challenging.

[12] describes how text mining is used to retrieve results for similar papers on a certain topic. This paper also indicates that matching papers based on text similarities is a weak method to determine how close any two papers are on a conceptual level.

In [5], with the additional information from the author networks, analyzed by logistic regression, the authors can get improved results of link prediction in citation network

[14] introduces the notion of using a random walk to determine the probability of a research paper using a linked article in its citation as opposed to a completely new, unlinked paper.

[9] proposes a paper recommendation system in which, clustering and neighbor-based recommendation algorithm is used while they assume authors prefer papers similar to ones they published before.

[13] proposes a method which assembles a citation graph, ranks the nodes according to a modified form of PageRank, clusters the network hierarchically with MapEquation framework, and finally recommends for the given seed paper based on its location in the hierarchical tree.

[2] features semantic repositories, natural language-like query interpretation and P2P overlays for efficient computations in its recommendations.

[6] introduces a hybrid research paper recommender system, which combines keyword-based search with citation analysis, author analysis, source analysis, implicit ratings, explicit

ratings and the "Distance Similarity Index" (DSI) and the "In-text Impact Factor" (ItIF).

3.3 Graph Mining: Centrality Measures

In this project we will focus on using centrality measures for serving the purpose of finding seminal papers in a citation network. Many measures of centrality have been developed throughout the research in the field of graph mining [3].

In [15], the authors measure four centrality measures (closeness, betweenness, degree and PageRank) for authors in a co-authorship network and find that the four centrality measures are significantly correlated with citation counts.

In [10], betweenness centrality is shown to be an indicator of the interdisciplinarity of journals in the data from *Journal Citation Reports* of the *Science Citation Index* and the *Social Sciences Citation Index* 2004.

We also mention one recent paper, [7], in which the authors propose a unifying definition of centrality that includes all path-counting based centrality definitions (e.g., stress, betweenness or paths centrality) in directed acyclic networks, which citation networks happen to be an instance of.

4 Proposed Method

4.1 Architecture

We use a Mode-View-Controller (MVC) architecture for our system, depicted in Figure 1. The View is the web front-end; the user experience lies fully with the View. All interaction with the database contained in the Model is through the Controller. The View sends the Controller input from the user and Controller sends updates to View from its queries of the Model.

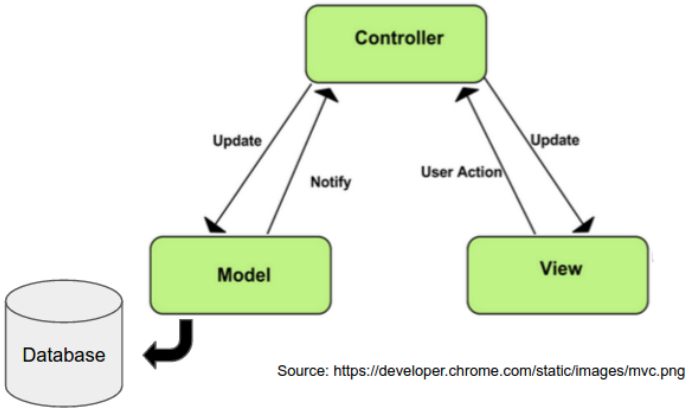

Figure 1: The MVC framework of the system.

4.2 Code Framework

We use Django (https://www.djangoproject.com/) for our code framework. Django is a high-level Python Web framework that allows easy integration of html and javascript front-ends with python back-ends. With it we have structure to create SQLite queries to the SQL database of the citation graph using python while we have a templated website with forms that update the site's D3 network visualization and text feedback based on user input and results from the python Controller back-end.

4.3 System Process

The layout of the web interface and visualization is shown in figure 2. The user begins by entering keywords (or a whole title) into text input fields in the top left. Some search parameters and options are available below (e.g. exact word matches or search result orderings). Search results, papers in the citation network that match the search criteria, are shown in the bottom left. The user can select papers from the search results to use as seeds for the network search or simply use the search keywords (i.e. use all search results as seeds). Keywords are both words within the title and the keyword meta-data of the paper.

Our Controller creates a graph, adding nodes up to a user-specified depth from the seed papers within the whole citation network. Several algorithms are then run on the network to determine various measures that may suggest importance of papers. These measures include degree centrality, in-degree centrality, betweenness centrality, load centrality, current-flow betweenness centrality, eigenvector centrality, centrality-with-removal, and keyword matches. The user can select which measures they prefer to visualize in the graph and to receive paper recommendations using or the user can accept certain pre-configured optimization functions. For the ease of the user, we have pre-configured functions that assign weights to each of the measures to create a single value for each node. Whatever the user's measure preference, the top papers by that measure or function of measures are displayed in the bottom of the webpage. The nodes in the graph visualization can be colored, sized, and weighted to visually represent measure values for visual interpretation.

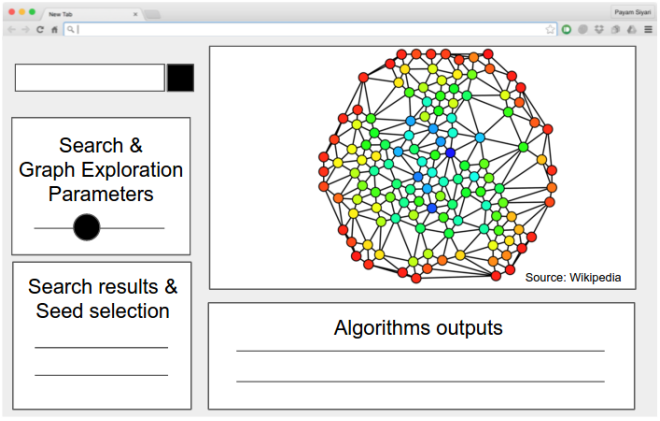

Figure 2: The layout and structure of the web interface and visualization.

4.4 Innovations

We have a few key points that differentiate our system from others:

- End-to-end paper recommendation system
- Uses many algorithms to find and rate papers for recommendation
- Uses paper data (keywords, title), not only network structure data

5 Experiments/Evaluation

Objective evaluation of our system is difficult, because it is difficult to find ground-truth for paper importance and there is no baseline system for comparison. We will perform time-response evaluations for the website queries and measures. In addition, each group member (5 total) will use the system to find seminal papers in a field in which he or she is unfamiliar. We will report the time to perform the search and a qualitative evaluation of the results of the search by an expert in the field.

6 Conclusions and Discussion

None yet.

7 Distribution of Team Effort

The work is split. Nominally, the graph visualization and page layout (the View) are being done by Vincent (Kangqi) and Francis. The back-end computation (the Controller and Model) are being done by Payam, Ari, and Samaneh. These work assignments are nominal; excessive work-loads or bottlenecks will be redistributed.

A Appendix

A.1 Plan of Activities

A.1.1 Original Plan

Dividing 3 months into:

1. Data collection and cleaning, Coding data protocols and the skeleton of the system: 15-20 days
2. Improvements to UI and Data analysis, algorithms implementation and experiment design: 1.5 month
3. Finalizing UI and Analytics integration, Addition of extra features if time permits: 1 month

As a midterm progress, a demo interface and 1-2 algorithms should be implemented with their experiments.

A.1.2 New Plan

18 days remain before the final presentation and report are due. We have completed:

- Demo form in Django that takes user input and loads information generate by python code
- Basic algorithms in place
  - Python back-end takes keywords and returns paper titles with those keywords
  - Python back-end builds a graph from seed papers. Returns graph and three centrality measures for each paper.

The plan for the remaining time is:

1. Complete System: 10 days
   - Display Controller's results in the View
   - Implement more node importance measures and a weighing function
   - Implement search, graph, and measure user selected parameters
   - Touch up and improve UI
2. Evaluation of non-expert user searches evaluated by experts: 5 days
3. Prepare final report and presentation: 3 days

References

[1] Leman Akoglu, Duen Horng Chau, U Kang, Danai Koutra, and Christos Faloutsos. Opavion: Mining and visualization in large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 717–720. ACM, 2012.

[2] Cristian Bancu, Monica Dagadita, Mihai Dascalu, Ciprian Dobre, Stefan Trausan-Matu, and Adina Magda Florea. Arsys – article recommender system. In *Proceedings of the 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, SYNASC '12, pages 349–355, Washington, DC, USA, 2012. IEEE Computer Society.

[3] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), June 2006.

[4] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 167–176. ACM, 2011.

[5] Jingyu Cui, Fan Wang, and Jinjian Zhai. Citation networks as a multi-layer graph: Link prediction and importance ranking. 2010.

[6] Bela Gipp, Jöran Beel, and Christian Hentschel. Scienstein: A research paper recommender system. In *Proceedings of the International Conference on Emerging Trends in Computing*, pages 309–315, January 2009.

[7] V. Ishakian, D. Erdös, E. Terzi, and A. Bestavros. A framework for the evaluation and management of network centrality. In *SIAM International Symposium on Data Mining*, 2012.

[8] Michael J. Kurtz and Edwin A. Henneken. Finding and recommending scholarly articles. *Bibliometrics and Beyond: Metrics-Based Evaluation of Scholarly Research*, 2012.

[9] Joonseok Lee, Kisung Lee, and Jennifer G. Kim. Personalized academic research paper recommendation system. *CoRR*, abs/1304.5457, 2013.

[10] Loet Leydesdorff. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9):1303–1319, 2007.

[11] Robert Pienta, James Abello, Minsuk Kahng, and Duen Horng Chau. Scalable graph exploration and visualization: Sensemaking challenges and opportunities. In *Big Data and Smart Computing (BigComp), 2015 International Conference on*, pages 271–278. IEEE, 2015.

[12] Trevor Strohman, W. Bruce Croft, and David Jensen. Recommending citations for academic papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 705–706, New York, NY, USA, 2007. ACM.

[13] Ian Wesley-Smith, Carl T. Bergstrom, and Jevin D. West. A recommendation system based on hierarchical clustering of an article-level citation network. *In Preparation*, 2016.

[14] Ian Wesley-Smith, Carl T. Bergstrom, and Jevin D. West. Static ranking of scholarly papers using article-level eigenfactor (alef). *In Preparation*, 2016.

[15] Erjia Yan and Ying Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *CoRR*, abs/1012.4862, 2010.