CLUSTERING FACULTY

Mohamed Gulaid, Gierad Laput, Jeremy Salfen
SI 650 - Information Retrieval
April 29, 2013

ABSTRACT

Higher education and academic research is increasingly interdisciplinary. Collaboration between faculty members across specializations has added significant value in scientific breakthroughs and the formation of new intersections of disciplines like bioinformatics and bioengineering. In a large academic setting like the University of Michigan, identifying potential collaboration between faculty members in different departments can be challenging. How can similar researchers be identified to support potential interdisciplinary collaboration? In this study, we applied information retrieval techniques to identify faculty members with similar research interests within University of Michigan. We used text data collected from the personal websites of faculty members in a number of different departments. Using cosine similarity, we clustered the faculty members hierarchically and then turned the resulting dendrogram into a network. We found communities of similar faculty members using graph partitioning techniques like modularity. Finally, we experimented with penalties for faculty members in the same department to encourage the formation of interdisciplinary clusters.

INTRODUCTION

Rapid advancement in digital technology and proliferation of software use in different disciplines has blurred the lines between different academic disciplines. Additionally, because of increasing complexity of scientific and societal problems, breakthrough in research often requires contributions from multiple fields. Many groundbreaking advances in science and technology are the direct result from collaboration from different scientific fields: e.g., the discovery of the structure of DNA, medical imaging techniques such as MRI, and satellite-based global positioning system (GPS).[1]in a large academic setting like the University of Michigan, identifying potential collaboration between faculty members in different departments can be challenging, especially for new faculty members and graduate students.

To infer potential interdisciplinary collaborations between different faculty members, we harvested text data from the personal websites of faculty members in a number of different departments. After cleaning the data, we applied information retrieval techniques like hierarchical clustering and community detection using centrality measures to find possible connections between faculty members.

[1] http://www.gradsch.osu.edu/Depo/PDF/LIFESCIENCESFINALREPORT092109(2)(2).pdf

One challenge we try to deal with is reducing the number of parameters in clustering. We do not know how many clusters we want, so k-means clustering is not a viable option. However, we also do not know where to cut the output of hierarchical clustering. So in this study we decided to try a hybrid approach, using the output of hierarchical clustering as a graph that can be partitioned using a method like modularity.

Our algorithm was successful in clustering faculty members by department. However, we want to infer potential collaboration across departments, so we implemented penalization in our algorithm to disassociate faculty members from the same departments. Our penalization algorithm returned mixed results. The diversity indexes of interdisciplinary clusters were low, indicating that our algorithm successfully clustered faculty members from different departments. However, we could not reach conclusive findings because other evaluation methods yielded mixed results. We conducted limited human evaluation wherein we asked faculty members to rate potential collaborators found by our algorithm. Feedback from surveyed researchers shows an above average rating for potential collaborators. Because of the lack of comprehensive data to evaluate against, we could not validate our findings in a comprehensive way.

RELATED WORK

The idea of constructing a scientific collaboration network from public profiles of faculty members is not new. Prof. Mark Newman at University of Michigan analyzed the collaboration networks of scientists from different fields, using the author attributions from papers over a five year period. Prof. Newman found a number of interesting properties of these networks. In all cases, scientific communities "seem to constitute a 'small world' properties in which the average distance between scientists via a line of intermediate collaborators scales logarithmically with the size of the relevant community."[2] Additionally, Prof. Newman algorithm found that the networks are highly clustered, meaning that "two scientists are much more likely to have collaborated if they have a third common collaborator than are two scientists chosen at random from the community."[3] This may indicate that scientists work in close-knit communities. And the clusters vary by scientific fields. For example, clusters in high-energy physics presented the largest clusters because of the magnitude of experiments that expand across universities and countries. The size of collaboration between faculty members fits a power-law distribution. The context of collaboration is important to understand size of and nature of collaboration.

Relatedly, in studies of the networks of citations between researchers, Robert Merton in 1968 and Barabasi et al. in 2001 showed that the number of links to papers (the number of citations) resembles a heavy-tailed distribution following a Pareto distribution or power law. Additionally, both studies found that node selection between authors' work is governed by preferential attachment. Preferential attachment in social networks is similar to cumulative advantage, whereby eminent researchers will compound prestige over time into an increasingly larger

---

[2] M.E.J.Newman. 2001. *The Structure of Scientific Collaboration Networks,* Proceedings of the National Academy of Sciences.

[3] Ibid

advantage. This means that new authors are more likely to co-author with someone who already has a large network of co-authors.[4]

METHODOLOGY

We scraped the homepages of faculty in 26 departments spanning mathematics, engineering, physical sciences, social sciences, and health sciences (see Table 1). This gave us a total of 1535 faculty homepages. We removed all html, javascript, and other formatting from the data. Then we made all words lowercase, removed stopwords, and filtered out commonly occurring bigrams. Finally, we lemmatized the words.

Aerospace Engineering	Health Behavior and Health Education (SPH)
Architecture and Urban Planning	Health Management and Policy (SPH)
Astronomy	Industrial and Operations Engineering
Atmospheric, Ocean, and Space Sciences	Linguistics
Biomedical Engineering	Materials Science and Engineering
Biophysics	Mathematics
Biostatistics (SPH)	Mechanical Engineering
Chemical Engineering	Physics
Civil and Environmental Engineering	Psychology
Complex Systems	School of Information
Electrical Engineering and Computer Science	School of Natural Resources and Environment
Environmental Health Sciences (SPH)	Sociology
Epidemiology (SPH)	Statistics

[4] A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, T Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications, Volume 311, Issues 3–4, 15 August 2002, Pages 590-614.

Table 1- Faculty members in these 26 departments were used in the study

Next we converted the text of each homepage into a vector representation of tf-idf scores for unigrams and bigrams. We set a minimum document frequency threshold of 3 to filter out potential noise. And we set a maximum inverse document frequency of 0.5 to filter out tokens that appear in a majority of homepages. Finally, we removed any faculty members from our data who had fewer than 40 tokens after these filtering steps because we found that faculty with very sparse vectors resulted in poor clustering.

Using this vector space model for representing faculty homepages, we computed the pairwise distance between each vector using cosine similarity. This gave us a similarity matrix which used to perform hierarchical clustering using average linkage. We did not try other clustering options -- e.g. single, complete, or Ward's linkage -- though they would be interesting to experiment with in future research. The resulting dendrogram is in Figure A.

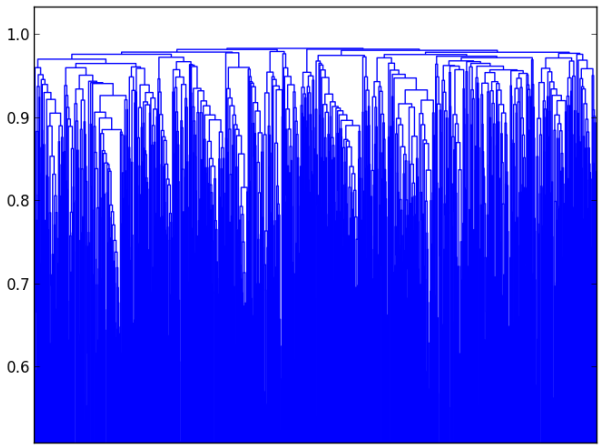

Figure A - Dendrogram from hierarchical clustering.
The y-axis represents cosine similarity.

Although the dendrogram looks like a graph with nodes and edges, the edges connect clusters rather than individual nodes. In order to convert it into a graph, therefore, we had to figure out how to connect the individual nodes. First we tried creating an edge between the two closest nodes for every two clusters that formed a connection in the dendrogram. This approach resulted in network with long chains (see Figure B).

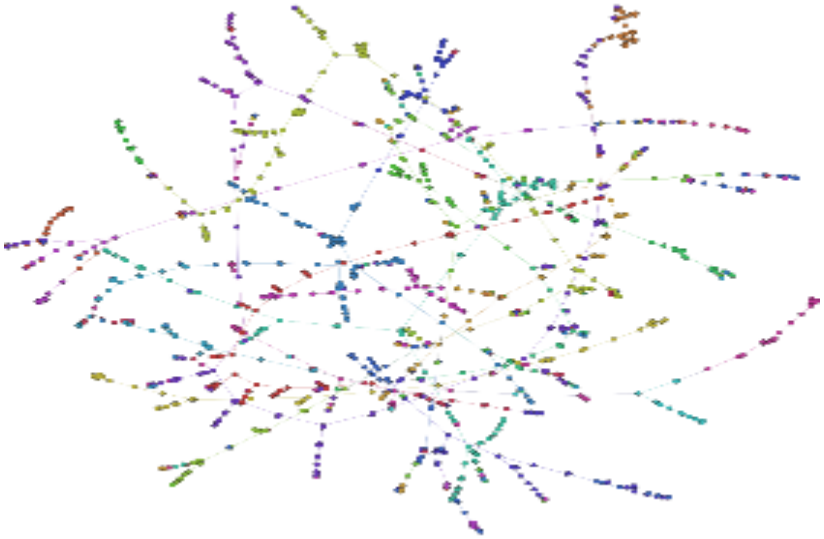

Figure B - Graph from approach connecting the two closest nodes for every connection between clusters. Colors represent departments.

Then we hit upon the following idea: When a singleton joins a cluster, it connects to all nodes in the cluster. Otherwise, when non-singleton clusters connect, they form an edge between their two closest nodes. Intuitively, this makes sense because when a person joins a research group, he or she forms connections with everyone in that group, but when two research groups share collaborators, the connections are probably strongest only between those who are engaged in collaboration. This approach created a network that looks more like what we expect (see Figure C), with tight clusters mostly based on department.

Figure C - Graph from approach connecting singletons to all nodes in a cluster. Colors represent departments.

We were pleased to find that our approach to community detection identified clusters of similar faculty members without requiring the tuning of any parameters. Using visual inspection, we could see that our hybrid approach of transforming the output of hierarchical clustering into a graph for partitioning may be a useful technique for clustering faculty with similar interests.

Because we hoped to find potential interdisciplinary collaborations, we next attempted to create clusters of faculty members from different departments. Using the similarity matrix, we introduced a penalty for faculty members in the same department, and then proceeded with hierarchical clustering and graph transformation as before. We experimented with penalties of 0.25, 0.5, and 1.0, producing three different graphs for evaluation (see Figure D). Finally, we used modularity to detect communities within these graphs.

Figure D - Graph with penalty of 1.0 for faculty in the same department. The mix of colors show that department clusters have been disrupted.

Evaluation and Findings

Evaluation Methods

Department Clusters vs. Interdisciplinary Clusters

To test the accuracy of our clusters, we performed a mix of automated and human evaluations (clusters in our case refer to the different modularities from the resulting network). We portioned our evaluations to reflect the two main threads of our clustering algorithm: department clusters (no-penalty) and interdisciplinary clusters (with-penalty). For evaluating department clusters, our goal was to measure how “related” the different nodes are under each cluster. Evaluating the interdisciplinary clusters was more challenging because we wanted to measure how “distinct” but “related” the nodes are under each cluster.

Validation Sets: Department Listings and MCubed Interdisciplinary Projects

The dataset we used for evaluating the department clusters were based from the department listings that we crawled from the faculty web pages. We used this list to determine which "academic department" a particular node belongs to. For example, when we compare whether two nodes belong to the same department, we use their node identifiers (ID) to retrieve the department names from our dataset and compare their values.

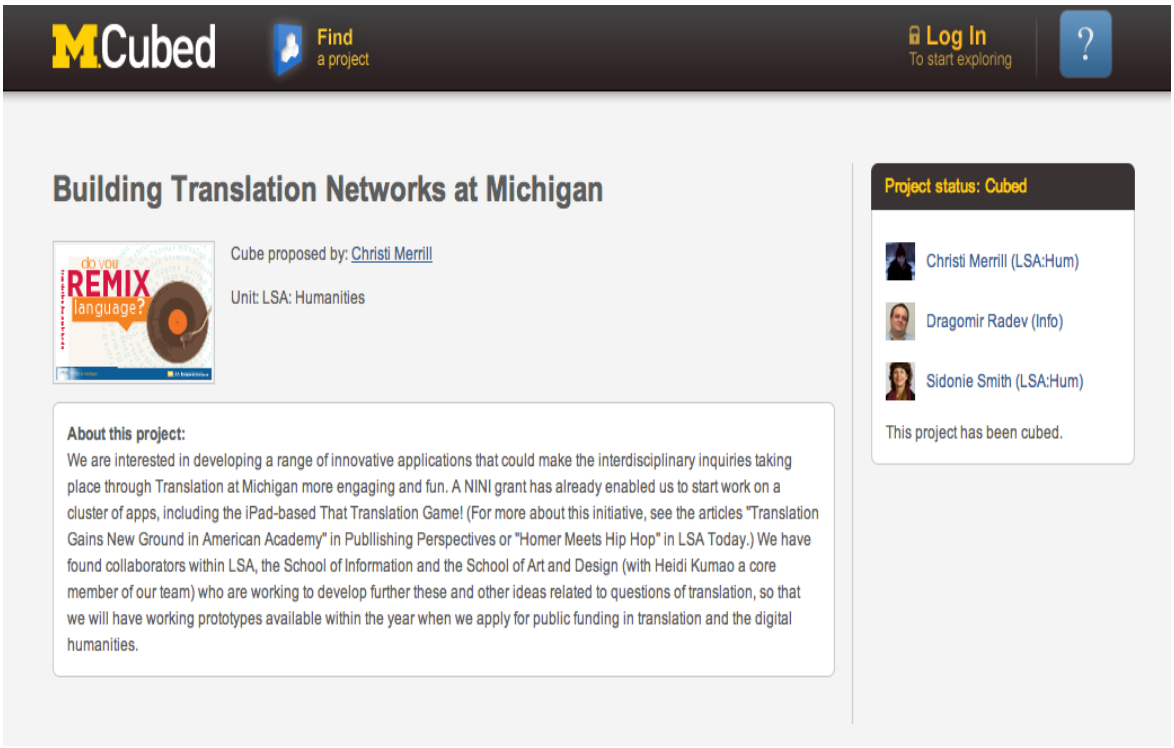

Figure E - Here is an example MCubed project showing collaborators from the School of Information and the LSA Humanities department.

For interdisciplinary clusters, we used the MCubed Project as our validation dataset. This decision was based on our assumption that MCubed projects were formed by faculty members from different departments across the University of Michigan (Figure E). We crawled the MCubed website and obtained 180 interdisciplinary projects, with each project having at least three interdisciplinary faculty members.

Diversity

In our first evaluation, we calculated the "diversity" of each cluster using Shannon's diversity index[5]. We go through each cluster (i.e., network modularity group in our case), and calculate diversity based on the department where each node belongs (Table 2). For example, a cluster with nodes that belong to the same department will have a diversity index of zero, while a cluster having a mixture of nodes from different department will have a diversity index greater than zero.

Penalty	**Diversity Index**
No Penalty (department clusters)	0.87
Penalty 0.25 (interdisciplinary cluster)	2.34
Penalty 0.50 (interdisciplinary cluster)	2.74
Penalty 1.00 (interdisciplinary cluster)	2.83

Table 2 - Calculations of Diversity Index in different clusters

Our numerical results indicate that our clustering algorithm yielded intuitively valid results. For department clusters, we obtained an average diversity index of 0.8, while interdisciplinary clusters had an average diversity of 2.8. This result is impressive given that our clustering algorithm was trained purely on text data that was crawled from faculty websites.

Number of Hops

In our second evaluation, we computed the "distance" (i.e., using the number of hops) among faculty members under each MCubed project (Table 3). For each faculty member, we searched for the representative node in our network graph, and compute the distance between other faculty members within the same MCubed project. Distance, in this case, was calculated as the length of the "shortest path" for two nodes in the network. We average the numerical value of these "hops" across all MCubed projects to get our final tally.

Penalty	**Average Hop Distance**
No Penalty (department clusters)	24.11 (SD=19.50)

[5] M. O. Hill, Diversity and Evenness: A Unifying Notation and Its Consequences, Ecology , Vol. 54, No. 2 (Mar., 1973), pp. 427-432 Published by: Ecological Society of America.

Penalty 0.25 (interdisciplinary cluster)	23.78 (SD = 12.88)
Penalty 0.50 (interdisciplinary cluster)	27.05 (SD: 13.01)
Penalty 1.00 (interdisciplinary cluster)	26.59 (SD = 10.19)

Table 3 - Calculation of Shortest path in different clusters

Our results showed that "department clusters" had the shortest average distance based on the projects from our MCubed dataset. This was contrary to what we originally thought, since we anticipated that our "interdisciplinary clusters" would mimic the distribution of cross-department collaborations from MCubed. Through visual inspection, however, we felt quite confident that our algorithm formed the interdisciplinary clusters we intended (see Table 3, and see the results for "Human Evaluation").

Membership

In our third evaluation, we identified whether faculty members in each MCubed project belonged to the same modularity groups in our clusters (Table 4). We used this technique as a “proxy” for determining how our clusters closely matched with “real-world” interdisciplinary groups. Our results for this evaluation were not particularly stellar. Overall, the modularity groups from our network graph matched the MCubed project memberships at a maximum of 8 out of 37 valid projects. However, upon closer inspection, we learned that most of the MCubed projects were not as “interdisciplinary” as we originally thought, given that a majority of the projects had two faculty members who were from the same department. The results of our membership evaluations are given in Table 4.

Penalty	**Membership Accuracy**
No Penalty (department clusters)	0.2162 (8 projects with correct memberships / 37 total projects)
Penalty 0.25 (interdisciplinary cluster)	0.0811 (3/37)
Penalty 0.50 (interdisciplinary cluster)	0.0541 (2/37)
Penalty 1.00 (interdisciplinary cluster)	0.0270 (1/37)

Table 4 - Proportion of same modularity membership in different clusters

Human Evaluators

In our final evaluation, we asked human subjects to judge the relevance of our results. We emailed a handful of faculty members and provided them with seven faculty suggestions whom our algorithm thought would be a good match for interdisciplinary collaboration. We asked them to rate the "relevance" of each suggestion on a scale of 1 to 5 (1=poor match, 5=good match). Our human evaluators gave us satisfactory results (Table5). Our human evaluators rated our suggestions with an average of 3 (number of suggestions=14, number of human evaluators=2). Evaluations from faculty are relatively higher than evaluation through automation. Additionally, we conducted evaluation on two other faculty members (Table 6). We found that the possible average rating is 2.43 (lower than evaluation conducted by faculty members). Nevertheless, our findings are within reasonable range: close to average.

Prof. Qiaozhu Mei	**Prof. Emily Mower Provost**
Raskin, Lutgarde Civil & Environmental Engineering, 2 Megginson, Robert, Mathematics, 3 Tiffany C.E. Veinot, Health Behavior and Health Education, 5 Lydia Soo, Architecture and Urban Planning, 1 Xuming He Statistics, 5 Carl F. Marrs, Epidemiology, 3 Allison Earl , Psychology, 3	Phoebe Ellsworth Psychology, 3 Jill Horwitz, Health Management and Policy, 1 Sandra R. Levitsky Sociology, 3 Jose Fernando Caetano, Architecture and Urban Planning, 2 Peter D. Jacobson, Health Management and Policy, 1 Arthur Oleinick, Environmental Health Sciences, 2 Rachel K. Best, Sociology, 3 Phoebe Ellsworth Psychology, 5
Average: 3.14	**Average: 2.86**

Table 5: Results from Faculty members evaluation

Our Evaluation:

Prof. Mark Newman	**Prof. Eytan Adar**
Sandeep pradhan EECS, 3 Ziff Robert m chemical engineering, 2 Mingyan Liu (EESC), 3 E. Margaret Evans, psychology, 4	Booth, Victoria, Mathematics, 3 Mingyan Liu, EECS 2 Kazu Saitou, Mechanical Engineering, 2 Michal Zochowski, Physics, 4

Mark Newman, EECS & Complex systems, 3 jack D. Kalbflisch Biostats, 1 Christopher J. Sonnenday, M.D., M.H.S., 1	Sanford Sillman, Atmospheric, Ocean, and Space Sciences, 2 Vijay Nair, Statistics, 3 Rosina M. Bierbaum, Natural Resources and Environment, 1
Average Rating: 2.43	**Average Rating: 2.43**

Table 6: Our evaluation of two faculty members from School of Information

LIMITATIONS AND FUTURE RESEARCH

Our research was limited by a few factors. We only used text data collected from personal websites of faculty members. This data may not provide enough semantic information to establish meaningful relationships between faculty members. Some faculty members have little information in their homepages, and others have information not relevant to our clustering project. Secondly, evaluating potential interdisciplinary connections between researchers is subjective. It is difficult to assess possible intersections of interest between faculty members. Is it defined by faculty members from departments coauthoring the same paper, or is it the different methodological approaches applied in research?

In order to remedy these limitations in future research, we would like to mine larger sets of data that include citations and abstracts of publications, which would provide us with deeper and more varied information about each faculty member. Additionally, we would like to experiment with different topic modeling techniques to reduce the dimensions in our data.

CONCLUSION

Using raw text data collected from the personal websites of different faculty members from University of Michigan, we used techniques in information retrieval and text-based network analysis to infer potential interdisciplinary collaboration between faculty members from different departments. Our methods were successful in clustering faculty members from the same departments. We got mixed results, however, when we introduced penalties to identify clusters of faculty members from different departments. We have shown that this approach has promise, and future experiments may find more successful techniques for identifying potential interdisciplinary collaboration.

REFERENCES

1. A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, T Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications, Volume 311, Issues 3–4, 15 August 2002, Pages 590-614. http://www.sciencedirect.com/science/article/pii/S0378437102007367

2. M. E. J. Newman, *The structure of scientific collaboration networks*, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM

3. Text Data Graph Approach, International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006), October, 25-28th 2006, Mérida, Spain, Proceedings, vol. II, p. p. 586-592. http://hal.archives-ouvertes.fr/docs/00/16/59/64/PDF/XP_ESJ_Text_Data_Graph_Approach.pdf

4. M. O. Hill, Diversity and Evenness: A Unifying Notation and Its Consequences, Ecology , Vol. 54, No. 2 (Mar., 1973), pp. 427-432 Published by: Ecological Society of America.