Recognizing Multidisciplinary Work on arXiv

Varun Ramesh
Stanford University
vramesh2@stanford.edu

Abstract

For this project, I constructed a graph from a record of all papers on arXiv. This graph is then used to evaluate how multidisciplinary an author's work is. Instead of simply looking at the topics that an author has submitted papers in, we can also see if they have collaborated with researchers who typically submit papers in other fields.

1. Introduction

Recognition for researchers engaged in multidisciplinary work is poor. There are no metrics for evaluating multidisciplinary researchers, these researchers are rarely the first author, and many prestigious journals are highly specialized. By examining networks of published papers, we may be able to provide better recognition for these researchers who would otherwise not be known.

2. Methods

2.1. Dataset

For this project, I downloaded preprint metadata from arXiv. This data is available using the Open Archives Initiative (OAI) protocol, and can be harvested incrementally using a variety of clients. The final metadata download was 2.6 GB. The data consisted of 173 categories, 822,687 authors, and 1,335,498 records.

2.2. Graph Construction

I constructed a directed graph from the metadata. For each author, I create a node, and for each topic I create a node. For every author, I add an edge to a topic if they published a paper in that topic. I also add an edge to every coauthor they have published with. The resulting graph had 822,860 nodes and 69,247,290 edges. An analysis of connected components revealed 1 weakly connected component and 25,535 strongly connected components.

I also plotted the out degree distribution, as shown in Figure 1. The distribution shows that most researchers have under 10 coauthors, while a few have over a thousand.

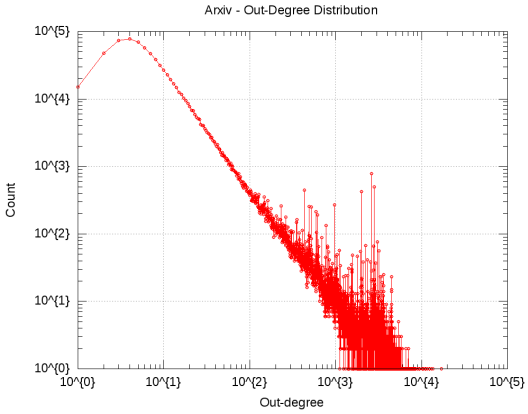

Figure 1. The out degree distribution for the constructed graph.

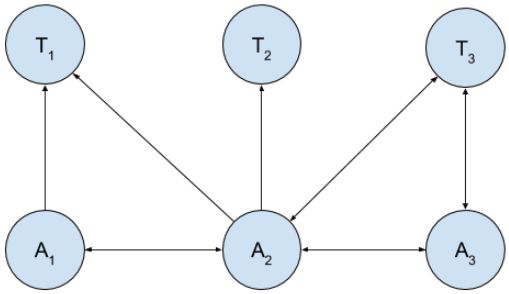

Figure 2. An example network.

2.3. Scoring Algorithm

In the simplest case, one can simply see how many topics an author has published, and report that as the score. We can generalize this to find, within n hops of the author node, how many unique topic nodes are reachable. I decided to pick $n = 3$, primarily for performance reasons. Due to the fact that each breadth-first search is independent, I was able to run on a 72-cpu EC2 instance.

As an example, the network in Figure 2 results in all three authors having a topic score of 3.

3. Results

The top ranked researchers, as calculated using the previously mentioned method, are shown in Table 1. A selection of rankings of Stanford professors are shown in Table 2.

Rank	Researcher	Topic Score
1	E. Canessa	83
2	Ou-Yang Zhong-can	76
3	Jean-Jacques E. Slotine	76
4	YuanJie Du	75
5	GuoJun Qiao	75
6	Georgios Pitsilis	75
7	I. Dan Melamed	74
8	Anthony Roberts	74
9	Shuanhong Wang	74
10	Zhigen Zhao	74

Table 1. The top ranked researchers, according to the topic score.

Rank	Researcher	Topic Score
10929	Jure Leskovec	20
22689	Percy Liang	16
47178	Andrew Ng	12
173047	David Mazieres	6

Table 2. A selection of Stanford professors and their rankings.

4. Future Ideas

There are several ways the scoring mechanism can be expanded. First, we can weight topics based on the distance from the author. As an example, one-hop topics would be weighted by 0.5, two-hop topics would be weighted by 0.25, and so on in a decaying fashion. Another possibility is to weight topics based off of how closely related the topics themselves are, thereby authors get more score for contributing to topics that are different. The difference between topics can be determined through the use of spectral embedding - thus the topic difference is simply the euclidean distance of the embedded vectors. Furthermore, we can use more detailed categorization schemes, and get more paper data from other databases.

Finally, we can follow citation links in addition to co-authorship links. This will help identify researchers who make contributions that are then applied by others to various fields.