Lasso: A meta search process to find collaborators

Ryan Mork[1,2] & Genevieve Shattow[3,4]

[1]University of Chicago
[2]University of Amsterdam
[3]Swinburne University of Technology
[4]University of Edinburgh

Collaborators in interdisciplinary fields should be as easy to find as a Google search. A quick search of Google Scholar using keywords of interest will yield a large number of papers with an even larger number of authors that can be potential collaborators. We propose a tool that searches Google Scholar with inputted keywords and looks for potential collaborators based on several parameters including topic grouping, number of collaborator, funding levels, and citations. In this document, we briefly describe the motivation for creating this tool, the algorithms and data required to implement it, a summary of its implementation, and future directions of development.

I. Motivation

While a keyword search of Google Scholar, PubMed, ArXiv, or any similar database is trivial, they are only the first step of how to find a potential collaborator. One instance where additional information might be needed is if a researcher develops a new data science tool and wants to use it for cross-disciplinary research. They might know what type of data they want to use or might be looking to work in an entirely new field. Alternately, a scientist might have an extensive data set but not know the person with the tools to analyse it. This tool will facilitate data-intensive and cross-disciplinary discoveries.

II. Lasso

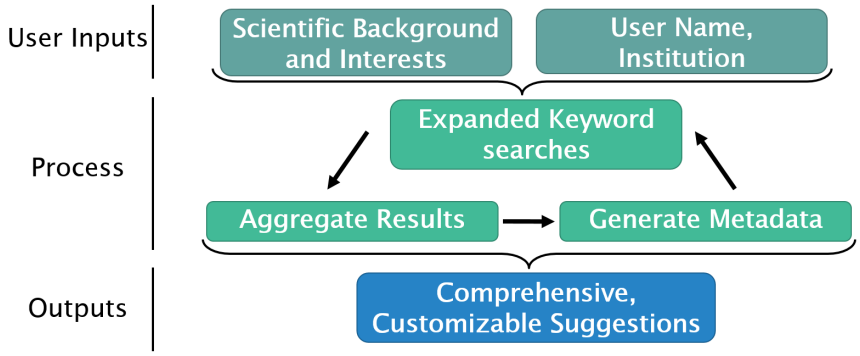
In its current set up, Lasso makes use of well known algorithms and publicly available data, which are described in this section. The required input is a keyword or set of keywords, and the user's name. We take the keyword(s), and mine Google Scholar for occurrences in the titles and abstracts. From these results, we utilise both the abstract content and the author lists. On the abstracts, we use a previously run topic grouping model (e.g. Blei & Lafferty, 2007) to determine a broad category for each result. With the names in the author lists, Lasso outputs a list of potential collaborators grouped by topic. Included are three scores for each potential collaborator, which are described below. We also limit the output to scientists who have co-authored at least a certain number of papers, allowing the user to find collaborators who are post docs or more senior.

Algorithms

Topic modelling: At its most basic, topic modelling is a method of grouping words that are co-occuring within documents. The algorithm requires a large number of documents and a set number of topics to find. In this case, we set a high upper limit to the number of topics, due to the array of subject covered in Google Scholar. The algorithm iterates over the abstracts to group key words into topics.

Collaborator Score: We determine a collaborator score which accesses each author's metadata. It takes into account the number of collaborators each author has, the number of repeat collaborations, and the network strength of these connections.

Funding Ranking: This is a score based on the amount of funding each potential collabo-



rator has, normalised by the year the funding was granted. It is incomplete because it only includes data from publicly available funding agencies.

Maturity Ranking: To determine this score, we use the number of refereed publications as well as other easily calculable information such as the author's h-index.

Data

For the initial implementation, all of the information is already contained in Google Scholar metadata. Specifically, we use the following fields:

- Paper titles and abstracts
- Author lists
- Number of publications
- Number of citations

We are not limited to Google Scholar. This algorithm can be applied to other similar databases such as the ArXiv, PubMed, ADS, among others. In the future, we will extend our results to include information scraped from publicly available data from funding reports listed by the NIH, NIAID, and NSF, etc.

III. Implimentation

The github repo can be found here: `https://github.com/Skemes/Lasso/tree/gh-pages`

A mock user interface can be found here: `http://skemes.github.io/Lasso/formInput.html`

IV. Future Directions

In its current form, Lasso is more a proof-of-concept than a fully developed algorithm. There are many directions it can move in the future. We plan on further developing the algorithm for idea generation and connection by expanding the set of metadata considered.

Outside of adding interpretations of metadata, we plan to develop topic modelling for smaller data sets, allowing the user to access smaller subtopics which do not have the required 100's-plus of papers that are typically required by topic modelling algorithms.

Finally, as with any product, the user is the consumer and their opinion is important. We intend to run usability studies to figure out what researchers want and will advertise and connect with research institutions. There is also obvious potential to include commercial partners such as LinkedIn as well as institutional partners such as the Data Commons.

V. Potential Issues

Google Scholar does not always distinguish between different authors of the same name. It also does not always recognise an author if first initials are used rather than the author's full name. This can lead to an author being identified more than once or not at all if they have too few papers.