The project plan of BIA660

2020 Spring

group 1

Fubin Mo

Jichen Dai

Yuan Yang

Qihao Zhang

1. **Data (Jichen Dai）**
2. **Data resources:** The resources mainly come from the Google Scholar website, which is a freely accessible [web search engine](https://en.wikipedia.org/wiki/Web_search_engine) that indexes the full text or metadata of [scholarly literature](https://en.wikipedia.org/wiki/Academic_publishing) across an array of publishing formats and disciplines.
3. **Data collection:** Design a web crawler, first choose several topics for the crawler to start with. Then let the crawler crawl on Google Scholar, collecting papers and information we need.
4. **Data storage**: The content of the paper will be stored in separate .txt files, while the publication information such as publisher and the author will be collected together into a .cvs file.
5. **Evaluating the papers using criteria that we chose to complement the database. (h-index and g-index)**

**2.1 H-index: (Fubin Mo)**

It is designed to measure the influence of a certain author, that is, the number of times a paper written by the author is cited.

Step 1: According to the author, find out how many articles he/she has published.

Step 2: For each article how many times it has been cited.

Step 3: Sort by the number of times the author's article is cited.

Step 4: Find out the order is greater than the cited number which is the h index of the author.

**2.1 G-index: (Qihao Zhang)**

The g-index is an alternative for the older [h-index](https://en.wikipedia.org/wiki/H-index), which does not average the number of citations. It aims to improve on the h-index by giving more weight to highly-cited articles. G-index accounts for the performance of the author's top articles, as well ashelps to make more apparent the difference between authors' respective impacts.

Step: [Given a set of articles] ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g2 citations.

1. **Saving and showing the results (Yuan Yang)**

After getting the H index and G index of all the papers in the given database, we need to observe and test the two new attributes to see if most of the papers are obviously separate by the new attributes. If so, the attributes we get are effective to use in the database. If not, we should try to combine the two indexes to one with mathematical methods to distinguish more papers. Last, we need to save and show the results as new columns in the database in order of the papers.