Data Science Fundamentals

(CSE 519 - Fall 2018)

PROJECT PROPOSAL

Objective:

The objective of the project is to devise a ranking metric to evaluate academic papers and researchers using various parameters from the dataset of academic papers.

Background Research:

Rank of Academic papers has always been an important for both researchers as well as those who are willing to fund researchers with their hard earned money. Having a good metric to rank the researchers as well as their publications is an important area of study.

Google has been working on gathering the database of academic publications over the past years and improving their metric to rank each scholar based on their relative importance in their discipline of study. As of now, they use h-index[1] as a metric to gauge the visibility and influence of recent articles published by a researcher.

The h-index although is a good metric but has following limitations.

- It does not take into account the number of authors on a paper. A scientist who is the sole author of a paper with 100 citations should be given more credit than one who is on a similarly cited paper with ten co-authors.
- It penalizes early career scientists. Outstanding scientists with only a small number of publications cannot have a high h-index, even if all of those publications and ground-breaking and highly cited.
- Review articles have a greater impact on the h-index than original papers since they are generally cited more often as they are more readable and have more practical analysis.This often skews the ranking.

Initial work done in this field used just the number of citations as a tool for analysing the journal importance. Later research propose the use of 'PageRank' algorithm along with the citation count to improve the ranking metric as described in literature [2][3] and [4]. Most recent studies have been focused on capturing the temporal nature of the citation network which were considered as static network till now[5]. The idea here is to not rely heavily on the citation count for determining the importance of a paper rather look at other factors like the time at which the paper was published. For example, an article which is new could be given a slightly larger weightage as the citation count is expected to be low for it, therefore compensating for the deficit. This sounds fair as well. There are couple of tools available for public which can be used to get data such as citation count and existing metrics for validating our model(Google Scholar Citation, Scopus and Web of Sciences). It will be somewhat challenging to extract the data from these sources though.

Some of the open challenges which are identified in this field are:

1. No distinguishment is made between a researcher who publishes an article all by himself and receiving say 'X' citations and a group of researchers (more than 1) who publish an equally cited article. There should be some extra weight given to the individual researcher while ranking him,
2. It is often the case that review articles tend to become more popular due their simplicity and practical analysis of a complex problem but it doesn't represent the original work, thereby it is quite possible for the original author to have a lower rank than author of the review article.We can although avoid this by removing the review article from the dataset but again that will be challenging in itself.

Dataset:

Firstly, we have planned to use Citation Network Dataset[6] which is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. This Dataset will be really helpful as it contains voluminous amount of data. It can be used to find the most influential paper by studying influence in the citation network. Our basic Dataset will consist of:

Data set	#Paper	#Citation Relationship
Citation-network V1	629,814	>632,752
Citation-network V2	1,397,240	>3,021,489
DBLP-Citation-network V3	1,632,442	>2,327,450
DBLP-Citation-network V4	1,511,035	2,084,019
DBLP-Citation-network V5	1,572,277	2,084,019
DBLP-Citation-network V6	2,084,055	2,244,018
DBLP-Citation-network V7	2,244,021	4,354,534
DBLP-Citation-network V8	3,272,991	8,466,859
ACM-Citation-network V8	2,381,688	10,476,564
ACM-Citation-network V8	2,385,022	9,671,893
DBLP-Citation-network V9	3,680,007	1,876,067
DBLP-Citation-network V10	3,079,007	25,166,994

Data Description: This dataset contains fields about paper title, authors, year, publication venue, index id of this paper and abstract of the paper.

Secondly, we have also planned to use Open Academic Graph (OAG)[7][8] which is generated by linking two large academic graphs: Microsoft Academic Graph (MAG) and AMiner. This dataset contains 64,639,608 linking (matching) relations between the two graphs. We will use this dataset to study citation network, paper content and others.

Thirdly, we have planned to use citeseer dataset[9][10] as well. It contains 384,413 publications in the form of vertices and 1,751,463 citations in the form of edges.

Fourthly, we are also planning to use CORA dataset[11] which consists of 2708 scientific publications classified into one of seven classes.The citation network consists of 5429 links.

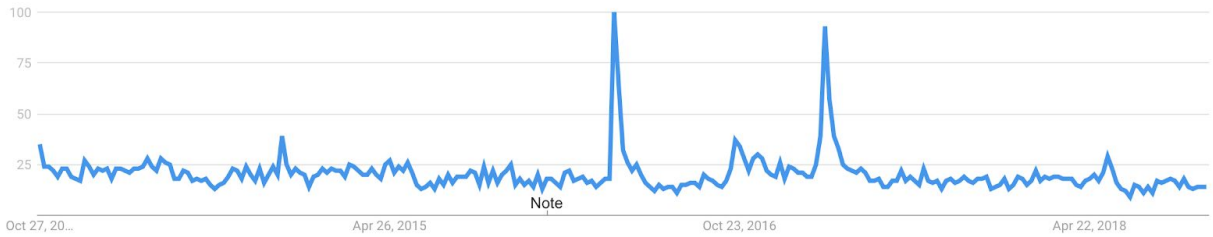
Approach:

The main challenge of the project is to consider various factors into account like author's impact, journal's impact, number of citations count of the paper, the publication date of the paper, the popularity of domain of the paper at the time of its publication etc. So we have planned to give a ranking metric to the paper based on 4 parameters as of now:

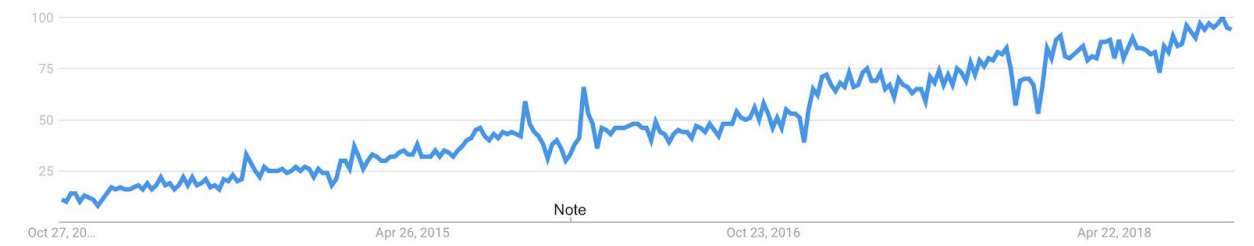
a. **Author's impact:** Authors with the higher reputation most likely to receive a higher scoring metric for their research papers.
b. **Number of citations count of the paper:** If a paper is cited by many other important publications, that means the paper is of good quality and should be given a higher score rating.
c. **Popularity of domain of the paper at the time of its publication:** The main idea of the research paper should be given considerable amount of weightage. If the research domain of the paper is more popular at the time of its publishing, that means it is most likely that the paper will be read by a much larger audience and it will be more popular and should be given a higher scoring metric.
d. **Journal's impact:** Good papers are most likely to be published in top journals and conferences. So they should be given a higher scoring metric.

Although we have a huge amount of data, the data available is in scattered form and each dataset contains information about author name, publication year, abstract, title and subject area which is present in a text file. We have to convert these text files into a csv file so that we can easily read them and perform manipulations on them. So, we are combining whole data and forming one dataset which is directly used for further data cleaning.

As a part of data cleaning, we are removing all the fields which are not useful and are not contributing in the prediction. After that, we had done some preliminary analysis of our 4 parameters and plotted some graphs to analyse how their impact varies over time. Below are our preliminary findings on some of these parameters:



The above graph is the trend of H.G. Wells with time and we used Google trends to plot this. We can take into account in our model that H.G. Wells was famous in January 2016 and Jan 2017. So the papers published by him around that time would most likely receive a higher rating.



The above graph shows the popularity of data science with time. This clearly shows that data science is a hot topic and any research paper published on data science now-a-days will most likely receive a higher rating.

To get the citation counts of the papers, we will be using page-ranking algorithm combined with the Google Scholar Metrics https://scholar.google.com. To build our ranking metric, we will be creating a network graph model in which we have edges from the paper nodes to author nodes, conference/journal nodes, number of citations node and domain popularity node. We will be exploring this step more and will be doing improvisation in the main model to evaluate how much weight should be given to each nodes and how to calculate these nodes score metric. So our final scoring metric will be calculated from the formula :

$$S(P_i) = \alpha\ \text{Author}(P_i) + \beta\ \text{Journal}(P_i) + \gamma\ \text{Citations}(P_i) + \delta\ \text{DomainPopularity}(P_i)$$

where $S(P_i)$ is the scoring metric of the i^{th} paper.

Validation:

We are planning to use various validation metrics to measure research impact. We have planned to use various criterias to validate Author and Journal impact separately.

Methods to measure Author Impact are below:

a. H-Index: Number of papers(h) with a citation number >= h.
b. G-Index: Given a set of articles ranked in decreasing order of the number of citations that they have received, the G-Index is the largest number such that the top g articles received at least g^2 citations.
c. I10 Index: Number of publications with at least 10 citations.

Methods to measure Journal Impact are below:

a. Journal Citation Reports: It is a product of ISI Web of Knowledge and is an authoritative resource to impact factor data.
b. Eigenfactor: It is measured as its importance to the scientific community.
c. Article Influence Score: It measures the average influence per article of the papers published in a journal.

d. SCImago Journal and Country Rank portal: It is a free online resource that uses citation data from Scopus to provide journal impact data.
e. Google Scholar Metric: It allows authors to view journal rankings and ratings by various h-indexes.

We are planning to upload the paper for which we have to predict our results on the website[12]. For each of the above metrics, we are going to obtain the ranking of the research paper. Once obtained, the ranking will be compared with the one obtained from our model. Hence, this will help us in validation.

Another way to validate our model results are permutation test, k-cross validation, chi-square test.

Next Steps:

1. Preprocess the data to create some useful features which will be used for better training of our model. Like we are thinking to add a column to give a rating to the popularity of the topic by analysing the Google Trends.
2. Perform EDA(Exploratory Data Analysis) and draw plots to find out the relationships between different parameters and to analyse how the popularity of a paper varies over time as well.
3. Build a scoring function from the features generated and train our model using those features.
4. Validate our model against some other metrics like H-Index, G-Index, Google Scholar Metric etc.
5. Identification of papers on arXiv which will become popular on internet.

References :

1. https://scholar.google.com/intl/en/scholar/metrics.html
2. Garfield, E. 1972. Citation analysis as a tool in journal evaluation. *Science* 178:471–479.
3. Pinski, G., and Narin, F. 1976. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics.
4. Sayyadi, H., and Getoor, L. 2009. Futurerank: Ranking scientific articles by predicting their future pagerank.
5. Yujing W., and YunHoi T., and Ming Z. 2013 Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information
6. https://aminer.org/citation
7. https://aminer.org/open-academic-graph
8. https://academicgraph.blob.core.windows.net/graph/index.html
9. https://citeseer.ist.psu.edu/myciteseer
10. http://konect.uni-koblenz.de/networks/citeseer
11. https://relational.fit.cvut.cz/dataset/CORA
12. http://guides.library.cornell.edu/impact
13. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD2008). pp.990-998.