CSE 519 - Project proposal

Amol Damare
adamare@cs.stonybrook.edu
Sbu Id : 107914028

Punit Mehta
punit.mehta@stonybrook.edu
Sbu Id: 111111111

October 22, 2017

1 Introduction

The main way researchers in various fields present findings of their research is through presenting a paper of their research either in a conference or publishing the paper in a journal. There are a lot of scientific papers published everyday. "arXiv.org" is a central repository of electronic pre-prints of such papers. Research areas include physics, mathematics, computer science, nonlinear sciences, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics etc. Everyday researchers publish their papers and upload pre-prints on arXiv.org. It is not a trivial task to judge these papers even for a expert in a particular research area. An automated way to evaluate these papers and ranking them would be very useful. Such a ranking can be used to determine not only the best papers but also it can be used in other applications such as recommendation of papers, determining how key characteristics of a good paper etc. We want to tackle this problem during our course project for CSE 519 Data science fundamentals. In this proposal we will identify 3 key components of the problem statements. We will also give description of the data that we are going to use and present preliminary findings from our exploratory analysis of the data. Finally we will describe the approach we will be using to solve each component of the problems and key challenges that we might face during each phase. In next section we will review some of the work that is already done in this field.

2 Related Works

There has been a lot of work done in the field of ranking. Most important from our perspective is h-index proposed by [1]. The h-index gives the ranking for researchers depending upon number of papers published and citations received by these papers. Similarly there is g-index proposed in [2] that also uses citations to rank work of a researcher. There are other ranking algorithms that rank objects other than researcher or researcher's work. Most famous example of such ranking would be PageRank algorithm [3]. Other significant works in this area include Citation analysis by E. Garfield [4], and work by Pinski and Narin [5] which they proposed not every citation is equal and developed a ranking that will incorporate this key concept.

3 Problem Statement

The goal of this project is analyze all scientific papers available on arxiv.org and come up with a ranking/scoring system that will give estimate of how good the paper is. Specifically we have identified 3 key components or subproblems that we will try and solve.

1. Ranking the papers which are already published and are available on arxiv
2. Ranking the papers which are not published anywhere but available on arxiv from quite some time
3. Ranking new papers which are recently uploaded on arxiv and are not published or accepted anywhere yet.

In the first subproblem we will consider the papers which have been accepted by a conference or have been published in a journal. Depending upon quality of the paper, these papers will have a good number of citations. In second subproblem we will consider papers for which we have pre-prints on arXiv.org for a relatively long time but they are not published or presented at any conference/journal due to reasons unknown. These kind of papers may or may not have citations. In last subproblem we will consider papers that are relatively new and are not published anywhere. For these papers we will most probably will not have citations. Next section we will describe the data in detail.

4 Data

All the pre-prints on arXiv.org are freely available. We can use the apis provided by arXiv.org to get all the data we need. Specifically we will be needing the data about paper such as authors, citations, keywords, references and so on. We will also need data about authors such as number of papers, total number of citations, affiliated institutes etc. Thankfully, Microsoft academic graph and open academic graph [6] already crawl this information for arXiv as well as other sources. A snapshot of this data is freely available. We are going to use this data for the purpose of this project. 1 shows the fields available for each of the paper. It includes abstract of the paper, link to author objects , links to the reference papers, keywords etc.

TODO: ADD ABOUT INFO SCIRATE. ADD SOME FIGURES/GRAPHS ABOUT DATA ???

5 Our Approach and Evaluation criteria

5.1 Approach

Judging a scientific paper is a hard task even for a expert. But there are some heuristics that can help in estimating the quality of a paper e.g. if we know a particular author generally writes good papers, probability of a new paper by same author being good is very high. Another heuristic is citations. If a paper is referenced in a lot of other papers then we can safely assume that it will be a good paper. [1] [**?**] have presented ideas based on citations. But we can not get a generalized scoring function from citations alone because number of citations depends upon the field. E.g. if field is difficult in general papers in that field will have less citations. Also if there are few number of people working in that field number of citations will be low. Hence number of citations alone can not be used to estimate the quality of the paper. Other factor that needs to be considered is rank of conferences or journals where the paper is published. In general if paper is published in a good conference it will be a good paper. When experts are reviewing scientific papers they usually have some criteria from which they can determine estimate the quality of the paper e.g. a good paper generally has about 8 pages(page limits for most of the conferences) with 1 page of introduction and 1 page of related section. It should have a figure and experiment section, a technical section with mathematical proof etc [7].

Field Name	Field Type	Description	Example
id	string	MAG or AMiner ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	department of computer science university of illinois at urbana champaign
venue	string	paper venue	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos	list of strings	fields of study	["relational database", "data model", "social network"]
n_citation	int	number of citation	29790
references	list of strings	citing papers' ID	["53e99ef4b7602d97027c2346", "53e9aa23b7602d970338fb5e", "53e99cf5b7602d97025aac75"]
page_stat	string	start of page	11
page_end	string	end of page	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/ 370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873",

Figure 1: Paper data model and fields for each paper [6]

To evaluate quality of each paper we will use impact of the author(may be h-index), number of citations, number of references, quality of references, field of study, affiliated institutes, conferences/journals published. In addition to these parameters we will also extract features from paper itself such as number of pages, number of figures/tables etc. We are also thinking of extracting features using nlp(may be creating embeddings for papers using nlp/deep learning) and using these features in our model. We hope to uncover more factors that will contribute to estimate of quality of the paper during the course of this project.

5.2 Evaluation Criteria

Keeping in mind above approaches in mind and our 3 part problem statement we have following evaluation criteria for each of the subproblem

1. For first subproblem we will have most amount of information since we are considering papers that accepted in a conference or published in a journal. We can calculate the score by getting the conference levels wherein the paper is published, citations of the paper, impact of the authors, affiliation details and other parameters which we will extract from the text of the paper

2. We have little less information in this subproblem than the first since we are considering papers which are not published anywhere. For unpublished papers which are available from quite some time, we assume that if they are good, they will have a good number of citations. We can take such papers, evaluate our model and if they are good, our model should give positive results.

3. In this case we have the least amount of information and evaluation is also hard. For new papers, we will not have citations or any other external data to score it. In this case, we

will mostly use the details available in the paper (that is authors' details, references and their credibility, number of pages, a proper format with relevant titles, figures etc.). We found one good resource [8] that already classifies good arxiv paper up to some extent and we can get the top papers from there (which will be very recent) and can evaluate our model by comparing it.

6 Conclusion

TODO

References

[1] J. E. Hirsch. An index to quantify an individual's scientific research output, 2005, Proc.Nat.Acad.Sci.46:16569,2005; arXiv:physics/0508025. DOI: 10.1073/pnas.0507655102.

[2] Egghe, L. Theory and practise of the g-index Scientometrics (2006) 69: 131. https://doi.org/10.1007/s11192-006-0144-7

[3] Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank citation ranking: Bringing order to the web (Technical Report). Stanford InfoLab.

[4] Garfield, E. Citation analysis as a tool in journal evaluation Science,178,471-479.

[5] Pinski, G., Narin, F. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics Information Processing and Management, 12(5), 297-312

[6] Microsoft Academic Graph https://www.openacademic.ai/oag/

[7] Andrej Karpathy blog http://karpathy.github.io/2016/09/07/phd/ "Writing paper" section

[8] SciRate.org https://scirate.com/arxiv/stat.ML

[9] arXiv.org https://arxiv.org/

List of Figures