Project proposal (CSE 519): Ranking arXiv papers

Amol Damare
adamare@cs.stonybrook.edu
SBU ID: 107914028

Punit Mehta
pmmehta@cs.stonybrook.edu
SBU ID: 111461860

October 23, 2017

1 Introduction

Researchers in various fields often present the findings of their research through either presenting a paper in a conference or publishing it in a journal. There are a lot of scientific papers published everyday. In the research community, 'arXiv.org' is a central repository of electronic pre-prints of such papers. There are many research areas available in arXiv such as physics, mathematics, computer science, nonlinear sciences, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, 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 an 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 the 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 three 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 propose our evaluation strategies. In the 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 the number of papers published and citations received by these papers. Similarly, there is a method based on g-index proposed in [2] that also uses citations to rank the 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] wherein they found that not every citation is equal and developed a ranking that will incorporate this key concept.

3 Problem Statement

The goal of this project is to analyze all the scientific papers available on arXiv.org and come up with a ranking/scoring system that will give an estimate on how good the paper is. Specifically,

we have identified the following three key components (sub-problems) that we will try to 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 sub-problem, 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 the second sub-problem, we will consider papers for which we have pre-prints on arXiv.org for a relatively longer 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 the last sub-problem, 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

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”, “http://polar.lsi.uned.es/revista/index.php/ia/ article/view/479”]
abstract	string	abstract	Our ability to generate…

Figure 1: Paper data model and its fields [6]

Following are the data sources that we have as of now and are planning to use for the project.

1. 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.
2. Specifically, we will be needing data about the paper such as names of authors, paper’s total citations, keywords, references and so on. We will also need data about its authors such as their number of papers, total number of citations, affiliated institutes etc. While exploring about the options to get this information, we found that Microsoft academic

A	B	C	D	E	F	G
id	***title***	***authors***	***n_citation***	***references***	***venue***	***year***
53e997	Fuzzy Sets	[{'name': 'Lotfi A. Zadeh'}]	91213	['53e99b26b	Information and Control	1965
53e998	Statistical Learning Theory	[{'name': 'Vladimir Vapnik', 'org': 'Bell Laborato	67786		Technometrics	1998
53e998	Cancer statistics, 2008.	[{'name': 'Jemal Ahmedin', 'org': 'Cancer Occur	49611	['53e9a9b7b	CA: a cancer journal for clinicians	2002
53e997	Introduction to algorithms	[{'name': 'thomas h cormen', 'org': 'massachuse	46246	['53e9979bb	Introduction to algorithms	1990
53e997	Phylogenetic Inference.		41468	['53e99a20b	Encyclopedia of Parallel Computing	2011
53e997	Fuzzy Sets	[{'name': 'James Buckley', 'org': 'Mathematics D	37484			
53e997	Convex Optimization	[{'name': 'Stephen Boyd'}, {'name': 'Lieven Van	35930		Convex Optimization	2004
53e998	DEVELOPMENT AS FREEDOM	[{'name': 'AMARTYA SEN'}, {'name': 'Amartya S	29763	['53e99acab	Journal of Public Health Policy	1999
53e997	Unsupervised learning	[{'name': 'Geoffrey Hinton'}, {'name': 'Terrence	29443		Unsupervised learning	1999
53e997	Random Forests	[{'name': 'Leo Breiman', 'org': 'Statistics Depart	29355	['53e9979fb7	Machine Learning	2001
53e997	Matrix analysis	[{'name': 'Roger A. Horn', 'org': 'The Johns Hopl	27168		Matrix analysis	1985
53e997	Working memory.	[{'name': 'Alan Baddeley', 'org': 'Department of	26361	['53e9acfeb7	Scholarpedia	1992
53e998	Support-Vector Networks	[{'name': 'Corinna Cortes', 'org': '<i>AT&T Bell L	25937	['53e99a67b	Machine Learning	1995
53e998	Subjective well-being.	[{'name': 'Ed Diener'}]	24149	['53e9b582b	Psychological bulletin	1984
53e997	Administrative behaviour	[{'name': 'Herbert A. Simon'}]	23502		Australian Journal of Public Administration	1950
53e997	Sampling Techniques	[{'name': 'William G. Cochran'}]	22976	['53e9a23eb	7602d9702b44cf3']	1963
53e998	Principal component analysis	[{'name': 'Erkki Oja'}]	21286		The handbook of brain theory and neural networks	1998

Figure 2: Top records with max citations

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 mainly going to use this data for the purpose of this project. We played with their json data files using panda data-frame and following are our initial observations and findings.

Figure 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.

Figure 2 shows some of the top records (having maximum citations) that we found after some data-frame processing.

Similarly, figure 3 lists the top venues with the total number of citations. Clearly, Nature and Science are the most popular in scientific community.

Figure 4 shows the number of papers publised in Nature in the last 100 years. The data we used to plot this is from one of the data files (and not the entire database as it's distributed across files) and it's not fair to conclude anything based on this graph at this moment as there is a lot of more data to be processed.

3. Scirate [8] is a platform that shows the top papers for varieties of fields. It's mentioned on their portal that the data is available under Creative Common License. So, we have requested them to give access to the data and we hope that we will get it. With the help of Scirate, we can even tune our model better according to our use-cases.

5 Our Approach and Evaluation criteria

5.1 Approach

Judging a scientific paper is a hard task even for an expert. But there are some heuristics that can help in estimating the quality of a paper. For example, we know that a particular author

A	B
venue	n_citation
Nature	271589
science	105297
Information and Control	97492
Commun. ACM	86856
Machine Learning	84034
Technometrics	71293
Science (New York, N.Y.)	70089
Physics Today	63836
Scholarpedia	53162
Encyclopedia of Parallel Computing	52998
Ssrn Electronic Journal	52216
CA: a cancer journal for clinicians	51617
Encyclopedia of Machine Learning	49717
Ca-a Cancer Journal for Clinicians	47718
Artif. Intell.	47562
Introduction to algorithms	46825
SIGGRAPH	46558
Proceedings of the National Academy of Sciences of the United States of America	43239
Physical Review Letters	42210
Annual Review of Psychology	37059

Figure 3: Top venues with max citations

generally writes good papers and hence, probability of a new paper by the 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] [5] have presented ideas based on the citations. But we can not get a generalized scoring function from citations alone because number of citations depends upon the field that is, if the research in the field is difficult in general, papers in that field will have less citations. Additionally, if there are few number of people working in that field, number of total citations per paper will be low. Hence, number of citations alone can not be used to estimate the quality of a paper. Other factor that needs to be considered is a rank of the conferences or journals where the paper was 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 the quality of a paper. For example, a good paper generally has about eight pages (or page limits for most of the conferences) with one page of introduction and one page of related section. It should have some figures or some kind of experiment section, a technical section with mathematical proof etc [7].

To evaluate the 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 (possibly by creating embeddings for papers) and using these features in our model. We hope to uncover more factors that will contribute to the estimate of the quality of a paper during the course of this project.

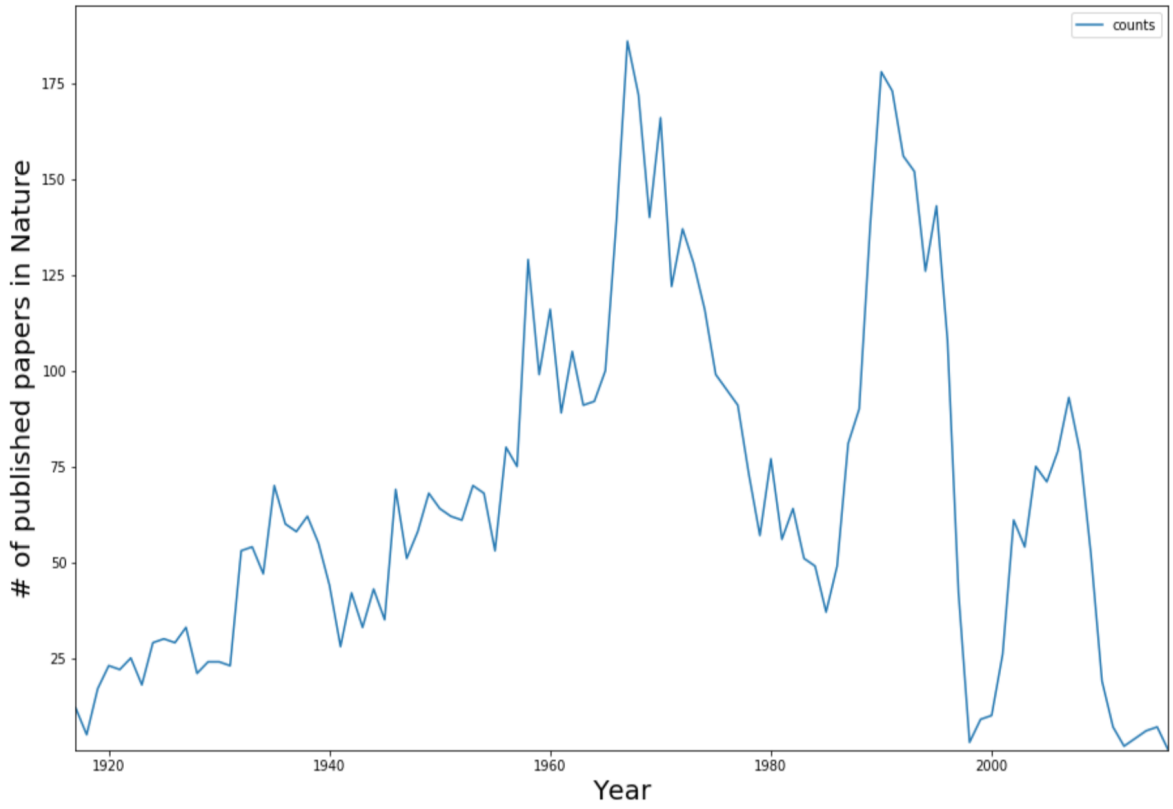

Figure 4: Papers published in Nature in last century

5.2 Evaluation Criteria

Keeping in mind the above approaches and our three sub-parts of the main problem statement, we have the following evaluation criteria for each of the sub-problem.

1. For first sub-problem, we will have most amount of information since we are considering the papers that are 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 the second sub-problem 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