# Project Report: CSE 519 Data Science Fundamentals
# Ranking of Academic Papers

## Objective:

Construct a ranking metric to evaluate academic papers and researchers. The ranking metric will be used to determine the top researchers in various disciplines. A 'reach function' will be constructed to determine the influence of a paper. Finally, devise a machine learning model that can predict which academic papers will gain popularity in near future.

## Goals achieved:

We are using DBLP-Citation-network dataset available on AMiner. The DBLP bibliography covers topics like Algorithms, Artificial Intelligence, Bioinformatics, Computer Architecture, Computational Biology, Data Mining, Data Structures etc. The fields available in the dataset consist of paper ID, paper title, paper authors, paper venue, published year, citation number and an abstract. Goals achieved:

1. Classify the publications into the respective domain of study.
2. Clean and preprocess the publication dataset.
3. Calculate the h-index of the authors in different domains.
4. Rank authors in different domains based on our evaluation metric based on Pagerank algorithm for citation network.
5. Rank publications using a modified time independent Page Rank algorithm.
6. Ranking of the domains of the publications.
7. Devise a reachability factor to determine the interdisciplinary impact of top papers in each field.
8. Time analysis of the paper citation dataset in terms of popularity of different domains across time periods.
9. Train a machine learning model that will predict which recent publications that will be popular in the future. We test this model recent papers obtained from the arxiv dataset.

In the following sections we discuss the details of implementation and results obtained:

## 1. Classification of publications into the respective domain of study

We have used REST APIs exposed by Microsoft Academic to classify a publication into a field of study. Microsoft Academic is a free public web search engine for academic publications and literature, developed by Microsoft Research. The REST call, when supplied with a paper title, returns a list of most likely domains for a paper in decreasing order of probability. We have captured the top three most probable fields for each paper. Sample Get REST request:

https://api.labs.cognitive.microsoft.com/academic/v1.0/evaluate?expr=Or(Ti='tools techniques for malware analysis and classification
')&attributes=Ti,AA.,F.FN

This returns a list of possible fields for the paper.

We added the top three fields obtained in the response as new columns for the domain of the publication.

The API has rate limiting, hence we needed to create multiple accounts to get multiple subscription keys and rotate the usage of keys.

To reduce the number of REST calls we are sending 10 titles in one request.

While searching for authors of a particular domain we will look up if the domain occurs in any of the top three columns and classify accordingly.

## 2. Cleaning and preprocess the publication dataset:

We are using the DBLP-Citation-network V10 as the dataset currently. The fields available in the dataset consist of paper ID, paper title, paper authors, paper venue, published year, citations and an abstract. We appended the field information as new columns as described above. For many rows author field, abstract field and citations

field are not available and were replaced by empty fields. For the papers for which Microsoft academic did not return domain details, the value “NA” was written.

## 3. H-index calculation for authors for multiple domains:

The H-index captures output based on the total number of publications and the total number of citations to those works, providing a focused snapshot of an individual’s research performance.
Example: If a researcher has 15 papers, each of which has at least 15 citations, their h-index is 15.
A scientist has index h if h of his/her N papers have at least h citations each, and the other N − h papers have no more than h citations each. For instance:
**Input:** citations = [3,0,6,1,5] **Output:** 3
**Python code snippet:**

```
def hIndex(citations):
    citations.sort()
    n = len(citations)
    for i in xrange(n):
        if citations[i] >= (n-i):
            return n-i
    return 0
```

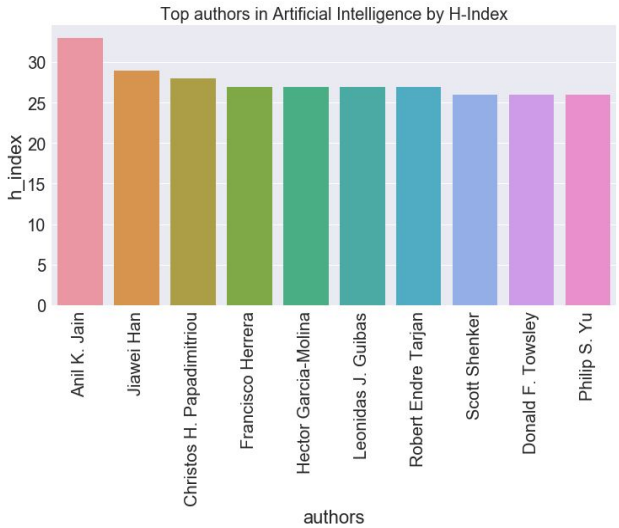
**Method to calculate h-index:**

- We construct a Paper citation graph which has paper ids as nodes and citations as edges. We have made use of NetworkX library to construct the complex network with ease.
- A map of authors and a list of published papers is created.
- Another map of authors and citation count of the published papers is created.
- h-index of each author is obtained by passing the citation count array of each author to **hIndex**() function.

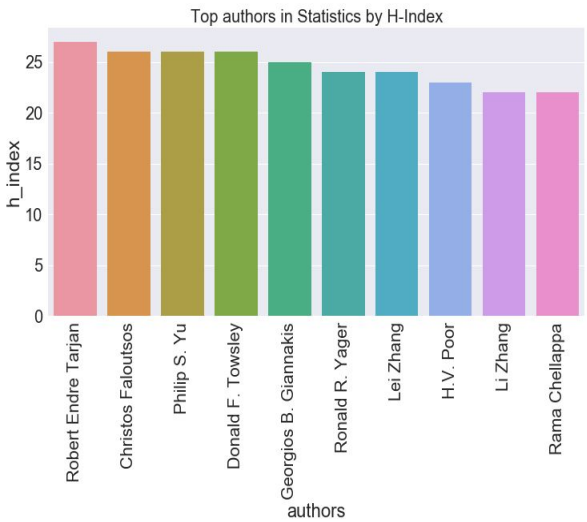
Based on the frequency of different domains of publications we have analyzed the following domains:

1. Artificial Intelligence
2. Statistics
3. Discrete mathematics
4. Biology

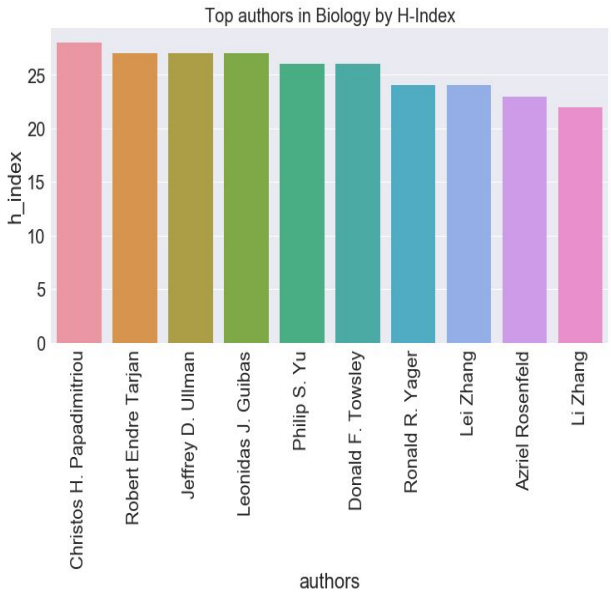
This can be extended to different other domains as well.
The dataset contains primarily similar computer science fields, choosing biology as a field to analyze will help us understand cross-domain applications when the domain is vastly different.
**Results for h-index:**



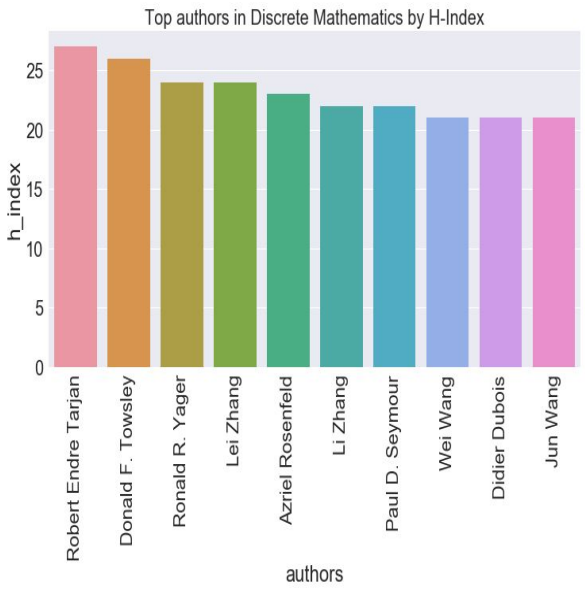

Author Ranking for Artificial Intelligence




Author Ranking for Statistics




Author Ranking for Biology




Author Ranking for Discrete Mathematics

**Observations**:

1. Since the DBLP dataset is primarily concerned with fields of Computer Science we see that authors in the field of Biology have relatively lower ratings.
2. There is quite an overlap between top authors in Biology and Statistics which shows the vast applications of statistical computer science in dealing high volume of biological data, especially genetic data.
3. There is a significant overlap between AI and statistics authors which resonates the fact that a lot of AI techniques have their origins in rudimentary statistical methods.

## 4. Page Rank calculation for authors for multiple domains and for papers in multiple domains:

The importance of a research paper is directly proportional to the number of research papers that cite it and this can be computed using PageRank on citation network. We might also be interested in knowing about the important conferences and authors. Using the PageRank score of the research papers we derive the conference scores. We rank the conference paper based on the quality of the research paper they publish. For the author, not only the quality of the research paper they publish matters but also the conferences where they publish their paper. So, to determine the author score we use both the research paper scores and conferences scores, which we can evaluate against h-index or i-10 index.

Some of the existing metrics like h-index just focus on the quantity of the citations but not on the quality of the citations, hence would not give correct results. Also, the number of citations also depend upon the area of research as some fields publish more research paper compared to other fields. We can nullify this effect by normalizing the score by the total number of papers published in that specific area.

In our citation graph, each node represents a research paper with various attributes like title, author, conference name, year etc and there is an edge from the citing node to the cited node.

In the basic PageRank we start with initializing all the candidates to a constant value, generally unity and then iteratively modifies each candidate's score depending on the score of the candidates that point towards it. It stops when all the candidate scores converge, i.e. become constant. The PageRank algorithm is based on the fact that the quality of a node is equivalent to the summation of the qualities of the nodes that point to it. In this case, quality refers to the score of the research paper

We are going to make some changes to adapt this algorithm for our use case.

1. When a research paper cites more than one research paper, its effect on the score of the paper it cites should diminish by a factor equal to the number of papers it cites. We divide the inlink score by the number outlinks of the inlink.

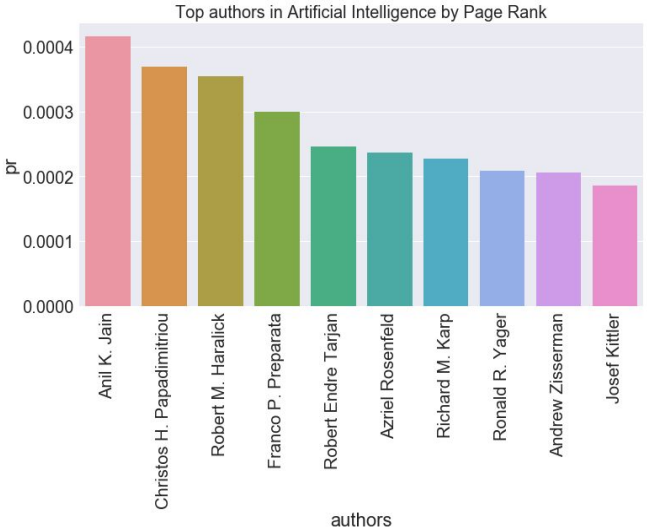
2. We use the damping factor, to prevent the scores of research papers that do not have any inlinks from falling to zero.

Pagerank Equation is as :

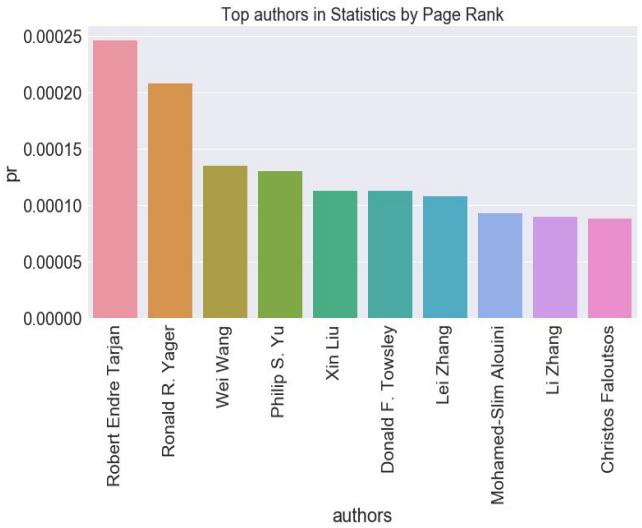
$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

$P_j$ = pages under consideration,
$M(p_i)$ = set of pages that link to $p_i$
$L(p_j)$ = number of outbound links on page $p_j$
N = total number of pages
D = residual probability

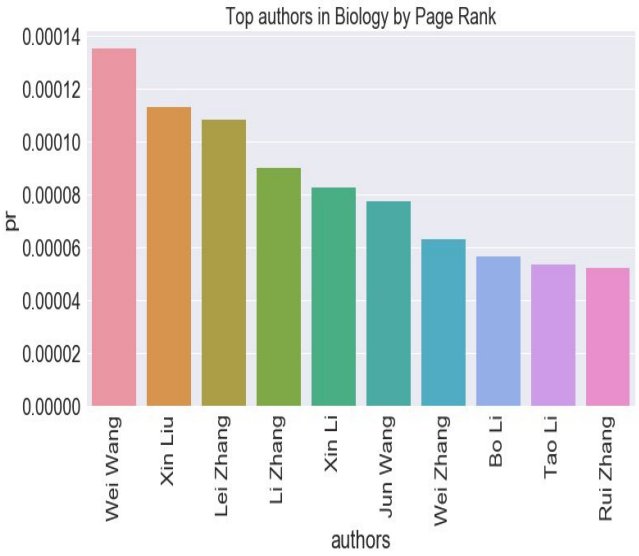
.



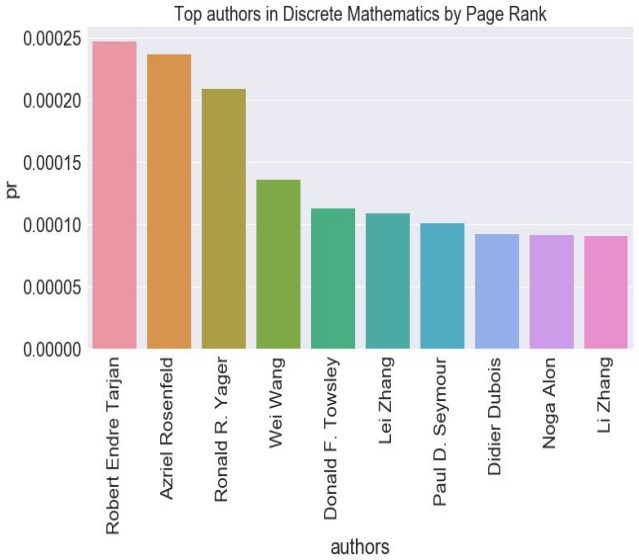

Author Ranking for Artificial Intelligence by Page Rank




Author Ranking for Statistics by Page Rank




Author Ranking for Biology by Page Rank




Author Ranking for Discrete Mathematics by Page Rank

**Observation:**

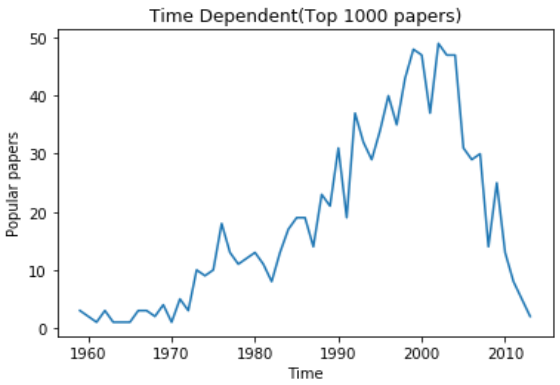
The h-index only takes into consideration the papers that are quoting it but Page Rank is also factoring in the papers the paper being rated is quoting i.e. both the indegree and outdegree of the citation graph is taken into account. Hence there is a considerable difference between the results of the two metrics. Page Rank is a much better metric than h-index.

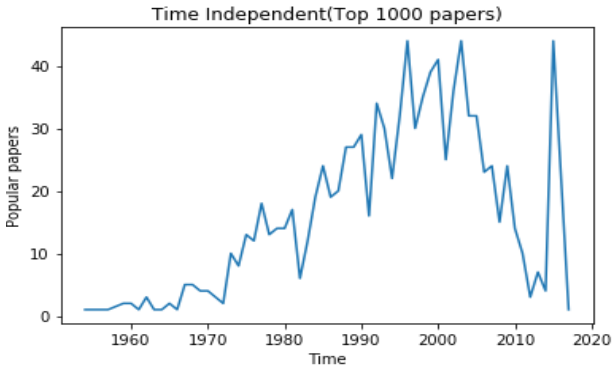
**Validation of Page Rank results for top authors:**
We have validated that our rankings of top authors for a field are indeed valid by researching about the author on Google Scholar. For example for the field of Artificial Intelligence:

| Author | Page Rank | H-Index From Google Scholar | Citation count from Google scholar |
|---|---|---|---|
| Anil K. Jain | 0.000434 | 179 | 185972 |
| Christos H Papadim | 0.000385 | 123 | 76609 |
| Robert M. Haralick | 0.000377 | 76 | 62622 |
| Robert Tarjan | 0.000264 | 113 | 74379 |
| Ronald R Yager | 0.000208 | 116 | 68038 |

## 5. Modified Page Rank:

But still, one problem still remains is that of the time. Time is a very important factor when it comes to ranking the research papers. The older research papers have more time to be studied by the researchers all over the world and be cited in various research papers. The newer ones did not get the same chance to get cited by different papers. We need to do some normalization of our scores based on time.
The metric we use is Average Year Citations Count, AYCC. We observe that this metric captures the time bias against the newer papers well and has high values for older papers and low values for newer ones.
$AYCC(Y) = \sum(PI(PY))/N(PY)$ .
AYCC(Y) is the metric score for year Y.
PI(PY) is the inlink count for papers published in year Y
N(PY) is the total number of papers published in the year Y.
We pre-compute the total number of citations for each year and the number of research papers published in each year. Using them, the average number of citations per paper for each year is determined.
We will use the Average Number of Citations per Paper in a year, i.e. the average of the total number of citations of the research papers published in each year, as out metric. This metric is suitable for normalization with respect to time as it captures the fact that an older paper has more time to be cited by researchers in comparison to the recent papers.
Now as each paper has the published year associated with it, we divide the PageRank score by the above metric for that year.
**Observations**:
Newer papers related to Machine learning, Big Data, Neural Networks topics which gathered strength lately but have lower citation count are appearing in top 1000 papers, which is not the case with time-dependent page rank algorithm. *In the time Independent Page Rank, we see that the count of better-rated papers is from recent years is higher.*




Comparison of top ten results from time-dependent and time-independent approaches:

**Time-Dependent Page Rank results**

| Title | Year | Score |
|---|---|---|
| A simple transmit diversity technique for wireless communications | 1998 | 0.000122122 |
| Fuzzy identification of systems and its applications to modeling and control | 1985 | 0.000114111 |
| Cognitive radio: brain-empowered wireless communications | 2005 | 0.000113806 |
| Snakes: Active Contour Models | 1988 | 0.000108725 |
| Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography | 1981 | 9.15E-05 |
| ANFIS: adaptive-network-based fuzzy inference system | 1993 | 7.89E-05 |
| Energy-efficient communication protocol for wireless microsensor networks | 2000 | 7.42E-05 |
| Ad-hoc on-demand distance vector routing | 1999 | 7.40E-05 |
| Textural Features for Image Classification | 1973 | 7.29E-05 |
| The anatomy of a large-scale hypertextual Web search engine | 1998 | 6.98E-05 |

**Time-Independent Page Rank results**

| Title | Year | Score |
|---|---|---|
| Ad-hoc on-demand distance vector routing | 1999 | 7.40E-05 |
| Textural Features for Image Classification | 1973 | 7.29E-05 |
| The anatomy of a large-scale hypertextual Web search engine | 1998 | 6.98E-05 |
| Fuzzy identification of systems and its applications to modeling and control | 1985 | 5.45E-05 |
| A simple transmit diversity technique for wireless communications | 1998 | 3.97E-05 |
| Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography | 1981 | 3.77E-05 |
| Snakes: Active Contour Models | 1988 | 3.63E-05 |
| ANFIS: adaptive-network-based fuzzy inference system | 1993 | 3.27E-05 |
| Cognitive radio: brain-empowered wireless communications | 2005 | 2.16E-05 |
| Energy-efficient communication protocol for wireless microsensor networks | 2000 | 2.03E-05 |

## 6. Domain importance

For a researcher looking for new topics for research, it becomes a very cumbersome task to go through the entire list of topics and decide upon which topics are important.

To determine the importance of a topic, we use an approach which is based on the intuition that the topic's importance should be determined by not only its frequency in the dataset *but also the quality of papers in which the topic lies and quality of citations those papers have*. To achieve this we have used the mean time-independent Page Ranks of all the papers published in a domain.

The ranking of different domains obtained is captured below.

| | domain | score | count |
|---|---|---|---|
| 0 | fuzzy logic | 5.267538e-07 | 1365 |
| 1 | artificial neural network | 5.252845e-07 | 2197 |
| 2 | combinatorics | 5.119102e-07 | 37637 |
| 3 | database | 5.083643e-07 | 23846 |
| 4 | cloud computing | 5.082579e-07 | 844 |
| 5 | statistics | 5.064040e-07 | 20237 |
| 6 | mathematical optimization | 5.054782e-07 | 108926 |
| 7 | machine learning | 5.042490e-07 | 32500 |
| 8 | biology | 5.029652e-07 | 11287 |
| 9 | distributed computing | 5.027616e-07 | 86339 |
| 10 | information retrieval | 5.014698e-07 | 14074 |
| 11 | real time computing | 5.010005e-07 | 102996 |
| 12 | artificial intelligence | 4.996918e-07 | 221813 |
| 13 | data mining | 4.996474e-07 | 62157 |
| 14 | electronic engineering | 4.991917e-07 | 40165 |
| 15 | psychology | 4.981610e-07 | 10784 |
| 16 | human computer interaction | 4.981516e-07 | 22160 |
| 17 | computer vision | 4.969523e-07 | 56235 |
| 18 | computer security | 4.876061e-07 | 18457 |
| 19 | embedded system | 4.863892e-07 | 8086 |
| 20 | internet privacy | 4.802950e-07 | 6182 |

**Ranking of the different domains using mean time independent Page Rank as the scoring criteria**

## Key Observations:

- We see that the quality of a topic is dependent on both the quality of individual papers as well as the number of papers in the cluster. A topic with few but good quality papers can have a high ranking like *fuzzy logic* and *cloud computing*. Though artificial intelligence has the highest number of publications the quality of all the publications is not very high leading to its mediocre rank.
- Also, the topics that appear at the bottom of the ranking are the ones which have not been/could not be researched much. Some of such bottom-ranked topics are *internet privacy* and *psychology*. These topics can be of special interest to the new researchers looking for new dimensions of research.

## 7. Reachability of a paper across domains:

An interesting analysis is that of the 'Reach' of a paper, meaning, in very simple terms, how influential it is across domains other than its own.

We define a very simple method of calculating the reach of a paper:

- We assign higher importance to a paper citing the current one if the fields of the two papers are sufficiently different.
- Seeing as how we obtain a list of values from the API we are using, we compare the fields of the two papers and determine them to be sufficiently different, if the number of common fields in the intersection of the sets of the two fields is lower than a threshold.
- If there exist too many common fields, we can simply surmise that while it could be an important paper in its own domain, it is not as influential across different domains, and thus we assign a lower priority to it.
- *reachability_paper = 1*(num_papers_in_different_fields) + 1*(if_num_papers_in_same_field > 1)*

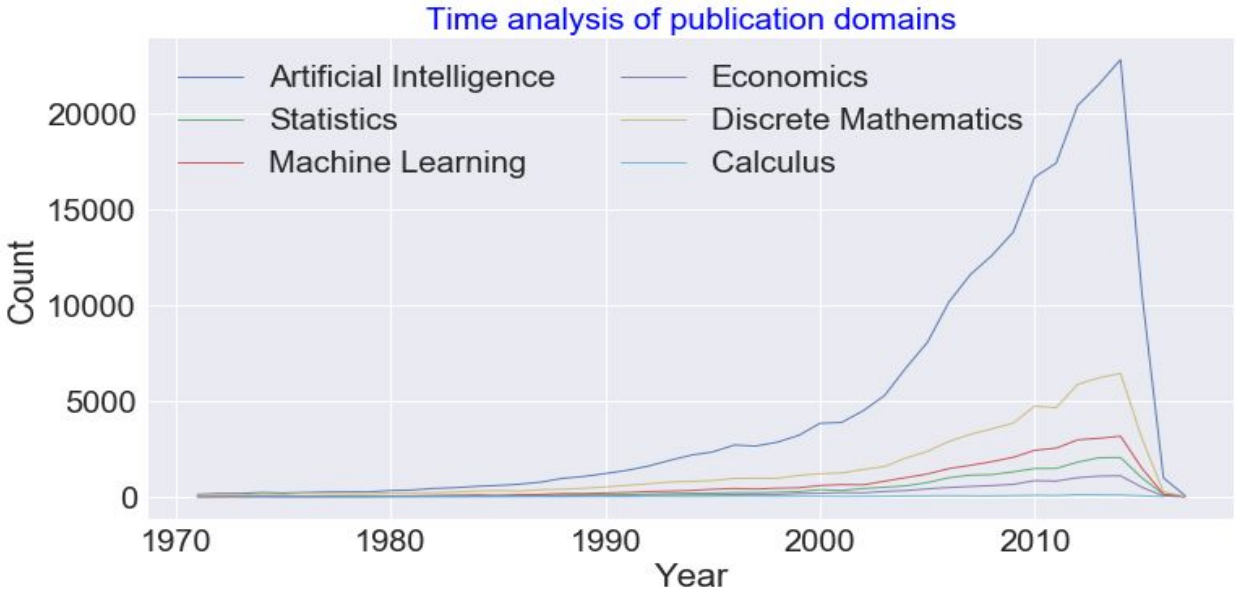
| title | reach_fields | page_rank | reach |
|---|---|---|---|
| ANFIS: adaptive-network-based fuzzy inference system | [multimedia, remote sensing, control theory, econometrics, inference, com... | 0.000039 | 94 |
| Wrappers for feature subset selection | [remote sensing, econometrics, control theory, systems engineering, geogr... | 0.000013 | 50 |
| Characterization of signals from multiscale edges | [control theory, systems engineering, computer network, computer science,... | 0.000011 | 23 |
| A learning algorithm for continually running fully recurrent neural networks | [computer graphics images, control theory, speech processing, financial e... | 0.000010 | 34 |
| Floating search methods in feature selection | [multimedia, remote sensing, control theory, geography, cancer, informati... | 0.000009 | 29 |
| A Bayesian Method for the Induction of Probabilistic Networks from Data | [computer security, computer network, information retrieval, computer sci... | 0.000009 | 32 |
| The mathematics of statistical machine translation: parameter estimation | [multimedia, financial economics, computer network, information retrieval... | 0.000008 | 54 |
| A survey of approaches to automatic schema matching | [ontology, systems engineering, data integrity, computer network, informa... | 0.000007 | 36 |
| Data clustering: 50 years beyond K-means | [combinatorics, cognitive psychology, computer security, control theory, ... | 0.000007 | 29 |
| A volumetric method for building complex models from range images | [combinatorics, missing data, least squares, computer security, delaunay ... | 0.000007 | 31 |

**Top papers by their reachability across different domains**

As can be seen, several papers that deal with more generic topics tend to be more influential across different fields. *Work done in Data Clustering is useful to work done in the fields of Information Retrieval and Parallel Computing.*

## 8. Time analysis of the paper citation dataset in terms of popularity of different domains across time periods

To get an insight into the popularity of a particular topic of research during a time period, we have classified each of the papers by their domain and grouped them by year of publication. A plot of topic and year wise number of publications gives us an idea of the variance of interest in a particular topic with time.




**Key Observations:**

1. In the DBLP dataset AI has seen a rapid growth with the total number of citations going from five thousand in the early 2000s to excess of twenty thousand in 2012.
2. The number of publications in biology and statistics has increased with an increase in AI publications.
3. Discrete Mathematics has had far fewer publications than AI. A peak of ~6400 as compared to AI's peak of ~21000 in 2012. But this was not the case in the 1990s where the number of publications is higher for Discrete Mathematics.
4. The number of publications on Calculus is relatively low. Count of publications on Calculus has been on the decline since 2010.
5. Neural networks and Machine Learning publications were around during the 1990s but in 2010s it has seen an unprecedented growth of interest.

## 9 Prediction:

To decide how important a paper can turn out to be is a challenging task, even after reading the paper. But implicitly, one way to filter out some papers would be to only consider papers from top authors, where we simply assume that the most important papers would be published in top venues by prominent researchers in their fields. Another challenge here is the fact that different experts would have different opinions on evaluating the potential impact a paper might have.

The data from arxiv that we have contains information on the title, domain and authors of each paper. We have two metrics to rate the authors namely the Page Rank and the h-index. We identified the following features to be of importance to predict the popularity of a paper in the future.

1. The Page Rank average of the authors
2. The h-index average of the authors
3. The domain of the paper published
4. Paper Venue
5. Year of publication

Since the test dataset from arxiv which are very recent and do not necessarily have a paper venue or any citation count. The year was also not an important characteristic as the training data is prior to 2018 while test data is from 2017-2018.
After considerable research, we decided to use the Page Rank and the h-index average of the authors to predict the Page Rank of paper in the dataset. The problem is defined now as the prediction of Page Rank of a paper in arxiv dataset using the Page Rank average and the h-index average of the authors.
We used Linear Regression and Random Forest classifier to achieve this. The results from both these classifiers gave similar results.

**Top 20 papers predicted by Linear Regression:**

| Title | year | Auth_PR | authHIndex | authors | field | Score |
|---|---|---|---|---|---|---|
| Recommendations as Treatments: Debiasing Learning and Evaluation | 2016 | 0.000217 | 5.6666667 | ['Tobias Schna | artificial intellige | 0.004279 |
| Latent Skill Embedding for Personalized Lesson Sequence Recommendation | 2016 | 0.000217 | 7.5 | ['Siddharth Rec | artificial intellige | 0.004247 |
| Active Fairness in Algorithmic Decision Making | 2018 | 0.00014 | 19 | ['Alejandro No | artificial intellige | 0.002648 |
| An Experimental Study of Cryptocurrency Market Dynamics | 2018 | 0.00014 | 19 | ['Peter M Kraff | human compute | 0.002648 |
| Bots as Virtual Confederates: Design and Ethics | 2016 | 0.00014 | 19 | ['Peter M Kraff | human compute | 0.002648 |
| Learning Preferences for Manipulation Tasks from Online Coactive Feedback | 2016 | 0.000133 | 12.5 | ['Ashesh Jain', | artificial intellige | 0.002635 |
| Your Activities of Daily Living (YADL): An Image-based Survey Technique for Patients with Arthritis | 2016 | 0.000133 | 27 | ['Longqi Yang', | human compute | 0.002383 |
| Unsupervised Domain Adaptation Using Approximate Label Matching | 2016 | 0.000114 | 9 | ['Jordan T. Ash | artificial intellige | 0.002336 |
| Improving Recommender Systems Beyond the Algorithm | 2018 | 0.000112 | 7.6666667 | ['Tobias Schna | human compute | 0.002327 |
| Unbounded Human Learning: Optimal Scheduling for Spaced Repetition | 2016 | 0.000109 | 6 | ['Siddharth Rec | artificial intellige | 0.002305 |
| Deep Learning for Encrypted Traffic Classification: An Overview | 2018 | 0.000113 | 14 | ['Shahbaz Reza | computer netwo | 0.002241 |
| Similar Elements and Metric Labeling on Complete Graphs | 2018 | 9.89E-05 | 10 | ['Pedro F. Felze | artificial intellige | 0.002052 |
| Scene Grammars, Factor Graphs, and Belief Propagation | 2016 | 9.89E-05 | 10 | ['Jeroen Chua', | artificial intellige | 0.002052 |
| Diffusion Methods for Classification with Pairwise Relationships | 2015 | 9.89E-05 | 10 | ['Pedro F. Felze | artificial intellige | 0.002052 |
| Private PAC learning implies finite Littlestone dimension | 2018 | 9.11E-05 | 5.6666667 | ['Noga Alon', 'I | artificial intellige | 0.001986 |
| Taskonomy: Disentangling Task Transfer Learning | 2018 | 9.93E-05 | 14.5 | ['Amir Zamir', ' | artificial intellige | 0.001981 |
| TCPlp: System Design and Analysis of Full-Scale TCP in Low-Power Networks | 2018 | 9.65E-05 | 13 | ['Sam Kumar', ' | computer netwo | 0.001957 |
| Learning to Optimize Neural Nets | 2017 | 9.55E-05 | 12.5 | ['Ke Li', 'Jitendr | artificial intellige | 0.001947 |
| Fast k-Nearest Neighbour Search via Prioritized DCI | 2017 | 9.55E-05 | 12.5 | ['Ke Li', 'Jitendr | artificial intellige | 0.001947 |
| Learning to Optimize | 2016 | 9.55E-05 | 12.5 | ['Ke Li', 'Jitendr | artificial intellige | 0.001947 |

**Top 20 Papers predicted by Random Forest Regressor :**

| Title | Year | Auth_PR | authHIndex | authors | field | Score |
|---|---|---|---|---|---|---|
| Private PAC learning implies finite Littlestone dimension | 2018 | 9.11E-05 | 5.6666667 | ['Noga Alon', 'R | artificial intellige | 0.018599 |
| Unbounded Human Learning: Optimal Scheduling for Spaced Repetition | 2016 | 0.000109 | 6 | ['Siddharth Redc | artificial intellige | 0.015289 |
| Etymo: A New Discovery Engine for AI Research | 2018 | 7.14E-05 | 4.3333333 | ['Weijian Zhang' | artificial intellige | 0.012965 |
| Recommendations as Treatments: Debiasing Learning and Evaluation | 2016 | 0.000217 | 5.6666667 | ['Tobias Schnabe | artificial intellige | 0.011294 |
| Incorporating Knowledge into Structural Equation Models using Auxiliary Variables | 2015 | 7.55E-05 | 4.6666667 | ['Bryant Chen', ' | artificial intellige | 0.01083 |
| Unsupervised Domain Adaptation Using Approximate Label Matching | 2016 | 0.000114 | 9 | ['Jordan T. Ash', | artificial intellige | 0.010798 |
| Improving Recommender Systems Beyond the Algorithm | 2018 | 0.000112 | 7.6666667 | ['Tobias Schnabe | human computer | 0.009898 |
| A Predictive Model for Music Based on Learned Interval Representations | 2018 | 4.21E-05 | 4 | ['Stefan Lattner' | artificial intellige | 0.007511 |
| Learning Musical Relations using Gated Autoencoders | 2017 | 4.21E-05 | 4 | ['Stefan Lattner' | artificial intellige | 0.007511 |
| Imposing higher-level Structure in Polyphonic Music Generation using Convolutional Restricted Boltzmann Machines and Constraints | 2016 | 4.21E-05 | 4 | ['Stefan Lattner' | artificial intellige | 0.007511 |
| On the Security of Warning Message Dissemination in Vehicular Ad Hoc Networks | 2017 | 5.29E-05 | 3.5 | ['Jieqiong Chen', | computer netwo | 0.006438 |
| Optimal and Scalable Caching for 5G Using Reinforcement Learning of Space-time Popularities | 2017 | 7.39E-05 | 8.6666667 | ['Alireza Sadeghi | computer netwo | 0.005677 |
| Interference Minimization in 5G Heterogeneous Networks | 2017 | 2.14E-05 | 6.2 | ['Tao Han', 'Guo | computer netwo | 0.005658 |
| Efficient Reciprocal Collision Avoidance between Heterogeneous Agents Using CTMAT | 2018 | 4.69E-05 | 16.5 | ['Yuexin Ma', 'Di | artificial intellige | 0.00506 |
| Automatic Differentiation Variational Inference | 2016 | 4.53E-05 | 3.75 | ['Alp Kucukelbir' | statistics | 0.004754 |
| TCPlp: System Design and Analysis of Full-Scale TCP in Low-Power Networks | 2018 | 9.65E-05 | 13 | ['Sam Kumar', 'M | computer netwo | 0.004728 |
| Approximate Probabilistic Inference via Word-Level Counting | 2015 | 3.51E-05 | 5.6666667 | ['Supratik Chakr | artificial intellige | 0.004619 |
| Large-scale Validation of Counterfactual Learning Methods: A Test-Bed | 2016 | 7.70E-05 | 6.5 | ['Damien Lefort | artificial intellige | 0.004535 |
| Using Shortlists to Support Decision Making and Improve Recommender System Performance | 2015 | 8.86E-05 | 10.75 | ['Tobias Schnabe | human computer | 0.004196 |
| Symbolic LTLf Synthesis | 2017 | 2.46E-05 | 6.6666667 | ['Shufang Zhu', 'I | artificial intellige | 0.004097 |

**The papers highlighted in blue are predicted in top 20 results by both the classifiers.*

**Validation of the Predictions by Machine Learning models:**

To validate that the publications are indeed going to be popular in the near future, we have made use of the chatter about these publications on Twitter. The number of tweets about a paper is a good measure of the popularity of a publication. If the number of tweets about a publication is high it is bound to get many citations in future and hence will rate highly in any ranking criteria. We have made use of Twitter's JSON based REST API to get the tweet count of the top papers predicted by the Random Forest model.

**Validation of top 10 results of Random Forest Regressor using Tweet Count of the paper in the past 90 days:**

| Title | Auth_PR | authHInde | field | Score | Tweet Count |
|---|---|---|---|---|---|
| Private PAC learning implies finite Littlestone dimens | 9.11E-05 | 5.666667 | artificial in | 0.018599 | 26 |
| Unbounded Human Learning: Optimal Scheduling for | 0.000109 | 6 | artificial in | 0.015289 | 4 |
| Etymo: A New Discovery Engine for AI Research | 7.14E-05 | 4.333333 | artificial in | 0.012965 | 83 |
| Recommendations as Treatments: Debiasing Learnin | 0.000217 | 5.666667 | artificial in | 0.011294 | 13 |
| Incorporating Knowledge into Structural Equation M | 7.55E-05 | 4.666667 | artificial in | 0.01083 | 6 |
| Unsupervised Domain Adaptation Using Approximate | 0.000114 | 9 | artificial in | 0.010798 | 0 |
| Improving Recommender Systems Beyond the Algori | 0.000112 | 7.666667 | human cor | 0.009898 | 6 |
| A Predictive Model for Music Based on Learned Inter | 4.21E-05 | 4 | artificial in | 0.007511 | 11 |
| Learning Musical Relations using Gated Autoencoder | 4.21E-05 | 4 | artificial in | 0.007511 | 27 |
| Imposing higher-level Structure in Polyphonic Music ( | 4.21E-05 | 4 | artificial in | 0.007511 | 31 |

**Key Observations:**

1. Most of the papers predicted to be popular have high tweet count and are either from the fields of Artificial Intelligence or Human Computer Intelligence. The only paper with low tweet count in the above table is from 2016 and hence has not been tweeted about in the last 90 days.

2. The publication titled **Etymo: A New Discovery Engine for AI Research** is of particular interest. Etymo is a system that aims to help readers navigate a large number of AI-related papers published every week by using a novel form of search that finds relevant papers and displays related papers in a graphical interface. According to our model as well as Twitter, this paper is going to be particularly important *though it has zero citations at the moment.*

**Link to datasets and Jupyter notebooks:**

1. The training set from dblp: https://drive.google.com/open?id=1u4ji3ww-zPbgbPGNfiY8gite4YoeA0Qy
2. The test set from arxiv: https://drive.google.com/open?id=1qyHzoHzLHHLjQ5quX0Kr5xQJGkZ7hiW
3. Microsoft academic notebook for domain classification: https://drive.google.com/file/d/1elQMZfFxej341BWW5L70vO0cbvmhwezo/view?usp=sharing
4. Primary notebook: https://drive.google.com/file/d/16MO8i7JHycPn6uJHzfDGsFYk5HNCUkak/view?usp=sharing
5. Arxiv client notebook: https://drive.google.com/file/d/17rLnEGNULJQUbFmUi6hmzHRHCq0_Dof_/view?usp=sharing
6. Twitter Client for validating prediction results: https://drive.google.com/file/d/1fa9-iusunmsqLKDWtqnSW_isjS3zdTs1/view?usp=sharing

## 10 References:


1. Journal Ranking Wikipedia- https://en.wikipedia.org/wiki/Journal_ranking
2. An efficient algorithm for ranking research papers based on citation network (https://ieeexplore.ieee.org/document/5976510)
3. Nan Maa Jiancheng Guanb Yi Zhao: Bringing PageRank to the citation analysis () (https://www.sciencedirect.com/science/article/pii/S0306457307001203)
4. Aminer Dataset: https://aminer.org/citation
5. NetworkX Python API: https://networkx.github.io/
6. Kumar Shubhankar Aditya Pratap Singh: An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution(https://link.springer.com/chapter/10.1007/978-3-642-23088-2_23)
7. https://developer.twitter.com/en/docs/tweets/search/api-reference/premium-search