2015 IEEE 15th International Conference on Data Mining Workshops

# Citation Prediction Using Diverse Features

Harish S. Bhat, Li-Hsuan Huang, Sebastian Rodriguez
*Applied Mathematics Unit*
*University of California, Merced*
*Merced, CA, USA*
*hbhat@ucmerced.edu*

Rick Dale, Evan Heit†
*Cognitive and Information Sciences*
*University of California, Merced*
*Merced, CA, USA*

***Abstract*—Using a large database of nearly $8$ million bibliographic entries spanning over $3$ million unique authors, we build predictive models to classify a paper based on its citation count. Our approach involves considering a diverse array of features including the interdisciplinarity of authors, which we quantify using Shannon entropy and Jensen-Shannon divergence. Rather than rely on subject codes, we model the disciplinary preferences of each author by estimating the author's journal distribution. We conduct an exploratory data analysis on the relationship between these interdisciplinarity variables and citation counts. In addition, we model the effects of (1) each author's influence in coauthorship graphs, and (2) words in the title of the paper. We then build classifiers for two- and three-class classification problems that correspond to predicting the interval in which a paper's citation count will lie. We use cross-validation and a true test set to tune model parameters and assess model performance. The best model we build, a classification tree, yields test set accuracies of $0.87$ and $0.66$, respectively. Using this model, we also provide rankings of attribute importance; for the three-class problem, these rankings indicate the importance of our interdisciplinarity metrics in predicting citation counts.**

## 1. Introduction

Funding agencies and researchers with limited time and resources increasingly seek metrics and models to quantify the potential impact of a collaboration, a proposal, or a paper [1]–[3]. One way of measuring the impact of a paper is through its citation count, the number of times the paper has been cited by other papers. In this work, we mine bibliographic data to build predictive models for a paper's citation count. We view the problem as a classification problem with each class corresponding to an interval of citation counts. As part of our approach, we extract a diverse set of features from bibliographic data, quantifying the effects of (i) author interdisciplinarity, (ii) author influence, and (iii) title words. We use these features in addition to features that are often used in citation prediction models. The best classifier that we build has a test set accuracy exceeding 87% and an area under the ROC curve greater than $0.87$.

A distinguishing feature of our work is the way in which we model the interdisciplinarity of authors. We do not attempt to tag authors by their discipline, nor do we make use of subject classifications for either papers or authors. Instead, we use the massive data set at our disposal to estimate each author's distribution of publication outlets, as a proxy for each author's scientific domain. In our work, we seek to quantify interdisciplinarity at the paper and journal levels, and then to explore statistical relationships between interdisciplinarity and measures of journal/paper quality such as impact factors and citation counts. We also identify an author's importance in the author network and the quality of the keywords in the title as important predictors for the paper's quality as reflected by its citation count.

After briefly reviewing related work in Section 2, we describe our data set in Section 3. In Section 4, we give a detailed account of how we extract features from the data. Next, Sections 5 and 6 describe, respectively, the exploratory analysis and predictive modeling we conducted. We discuss overall conclusions in Section 7.

## 2. Related Work

We briefly review selected past work on this problem, focusing on papers that apply machine learning methods to build predictive models.

A recent paper computes Author Rank, a ranking of authors based on their average citation counts in previous years [4]. From Author Rank, other features such as total past influence for authors, maximum past influence for authors and venue rank are computed. These features are then used to build predictive models with methods such as logistic regression, support vector machines, linear regression, support vector regression, and classification and regression trees. The goal is to predict citation counts for new papers from past author and venue impact. The main finding is that a measure based on the citation graph evolution works well for this prediction.

† *This study is based upon work peformed while Evan Heit was serving at the National Science Foundation (US). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.*

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.131

In [5], Ibáñez et al. tried to predict the number of citations of a paper within each of the four years after publications by comparing classification methods such as Bayesian networks, logistic regression, decision trees and $K$-nearest neighbors. The logistic regression and naive Bayes classification methods yield good accuracy in predicting citation counts. Similarly, in [6], regression models were pursued and Yan et al. noted that Author Rank, Venue Rank are most predictive in citation counts. Interestingly, by exploring the content of papers using word tokens in abstracts, they find that the content of a paper is not predictive of paper citation.

In an especially influential paper, Castillo et al. [1] built a predictive model for a paper's citation count based on features such as *a priori* author-based, *a priori* link-based, and *a posteriori* information using classification methods. They showed that prediction accuracy degrades using *a priori* information and noticed difficulty in improving baseline prediction using *a priori* attributes.

We do not directly compare the performance of these algorithms in the current study, which is based on a different dataset and with subtly different goals. These past studies do provide important context for the present work. We develop basic measures of interdisciplinarity for a published paper and explore whether these, too, predict citation count. There has been considerable discussion of the importance of interdisciplinary work, and its likely impact in future scientific endeavors [7]. Do interdisciplinarity features enhance our ability to predict citation counts? This is the question we seek to answer here.

## 3. Data

Using direct API access to the Thomson Reuters Web of Knowledge (WOK) database, we assembled a massive set of bibliographic data. Because the database was previously known as Web of Science (WOS), we will use WOK and WOS interchangeably. The data acquisition proceeded in two levels. In the first level, we formed two lists of journals: the top 250 journals by impact factor in both the science and social science sections of Thomson Reuters Journal Citation Reports (JCR). We then downloaded database entries for all papers published in the years 2005–2010 in 247 of the top 250 science journals, and 248 of the top 250 social science journals, because some of these top journals did not have complete issues yet in our target date range. In what follows, we refer to this set of journals as the "top 495" journals. In this way, we obtained entries for over 800,000 papers involving over 800,000 authors.

Using the list of all authors that appeared in papers from this first level, we conducted a second level of data collection. In this second level, we downloaded database entries for all papers by these authors over the time period 2000–2006. The total number of bibliographic entries acquired in the second level exceeds 7 million.

The sum total of all data described in the two-level process above is contained in 220 GB of JSON files. Our first step was to import this data into mongoDB, a type of NoSQL database. Because of the way that we downloaded the data, it became clear that our raw files sometimes contained more than one entry for the same paper. Because the WOS ID for each paper is unique, we used this field as the mongoDB identification field. In this way, we imported into mongoDB only one entry for each distinct paper in our raw data set.

The resulting database consists of bibliographic entries for $N_P = 7,957,302$ unique papers. These entries contain a substantial amount of raw information on each paper: the title, abstract, author names and affiliations, year published, journal name, times cited, etc. Essentially, for each paper, we obtain at least as much information as one would obtain by querying the Web of Knowledge database through its web interface, familiar to many researchers.

We use the mongoDB data for model training, including cross-validation to determine model hyperparameters. To test the model, we use a test set that does not intersect at all with the training set. To form the test set, we use all papers from the top 250 journals in science and the top 250 journals in social science, published during the years 2011–2014.

A key problem in analysis of bibliometric data is the unique identification of authors. There are at least two problems: (i) the same name can sometimes be shared by multiple distinct authors, and (ii) a single author may use different forms of her/his name on different papers. Our solution to this problem is to ignore the actual name of each author and instead identify each author by his/her DAIS (Distinct Author Identification System) ID. DAIS, a proprietary system used by WOK, claims to solve problems (i) and (ii) above.

We have two observations regarding the DAIS ID. First, not all authors have been assigned a DAIS ID. Using distinct author names (specifically, WOS standard names represented as Unicode strings), we counted $3,415,707$ unique authors; counting DAIS IDs, we counted $N_A = 3,092,291$ unique authors. Second, all of the author-related features used in this paper were originally computed using WOS standard names. Upon recomputing the features using DAIS IDs, we found the features to be more informative.

## 4. Feature Extraction

Feature extraction occupies a central role in our approach. Starting with over 200 GB of raw data, we employ a diverse array of techniques to extract meaningful information. We place our features into four broad categories:

1) Author Interdisciplinarity. We compute two metrics (Jensen-Shannon Divergence and Mean Entropy) that quantify the interdisciplinarity of each paper.
2) Author Influence. We score each paper based on the authors' PageRank in two coauthorship graphs.
3) Title Words. For each paper, we find the journal (in the set of the top $495$ journals described above) whose titles are most similar to the paper's title.
4) Classical Features. In this category, we include features such as: number of references cited, age, length, number of authors, past citations of the

authors, and past number of papers published by the authors.

We now detail each category.

**Author Interdisciplinarity.** When one seeks to model interdisciplinarity, the first question is how to determine precisely the discipline or field of a single author. Each paper in our data set is associated with WOS subject codes. Our primary objection to these codes is that they overly constrain the vast space of possible domains to which a paper or an author may belong. For example, consider an author who publishes in mathematics, computer science, biology, and physics. While some of the intersections of these fields may be associated with standard subject codes, e.g., "biophysics" or "computational biology," one would find it difficult to assign one subject code that covers all four of the author's scientific domains. Consequently, we argue for a data-driven approach that captures—in a precise, quantitative way—the vast range of possible scientific domains to which an author may belong. Additionally, we seek to avoid an arbitrary or subjective assignment of either authors or papers to particular disciplines.

To model an author's disciplinary interests/preferences, we consider the author's journal distribution. That is, for each author, we estimate the probability that the author publishes in each of the $N_J = 11,829$ journals (represented by ISSN numbers) in our data set. In this way, we build a journal distribution—actually, a probability mass function (p.m.f.)—for each author. We represent this p.m.f. as a sparse vector $\mathbf{x}$ with length $N_J$. To estimate $\mathbf{x}$, we compute

$$x_j = \frac{\text{number of times author has published in journal } j}{\text{total number of publications for this author}}$$

This computation is carried out efficiently—on the training set—using mongoDB's native Map-Reduce framework.

Having computed a journal distribution for each author, we can now quantify the interdisciplinarity of a single author as well as a group of authors. We first compute the Shannon entropy of each author's journal distribution. Entropy measures intrinsic interdisciplinarity, i.e., the spread of the author's journal distribution:

$$H(\mathbf{x}) = -\sum_{i=1}^{N_J} x_i \log(x_i). \tag{1}$$

For each paper, we compute *avgent*, the average of the entropies of the paper's authors. This is our first metric for quantifying the interdisciplinarity of each paper.

The second metric for quantifying the interdisciplinarity of a paper builds on the entropy. For an $N$-author paper, suppose the authors have journal distributions $\mathbf{x}^1, \ldots, \mathbf{x}^N$. We compute the Jensen-Shannon divergence (JSD or *jsd*) of these distributions:

$$D_{JS} = H\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^i\right) - \frac{1}{N}\sum_{i=1}^{n} H(\mathbf{x}^i), \tag{2}$$

with $H$ defined as in (1). The JSD aims to quantify how mutually different the authors' journal distributions are from one another.

Three points regarding entropy and JSD help to increase intuition about these quantities. First, for a single-author paper, the JSD is always zero. Such papers can still have a large mean entropy if the author's journal distribution is very spread out. This models an author with a broad background, disciplinary preferences, and/or interests.

Second, it is possible for the mean entropy of a paper to be zero while the JSD is positive and large. As a limiting case, this will happen for an $N$-author paper in which each author has published exclusively in one journal, and all of the journals are mutually distinct. This models a paper in which multiple monodisciplinary and complementary authors jointly produce a highly interdisciplinary paper.

Third, JSD and mean entropy are bounded below by zero. Neither has a theoretical upper bound that is independent of $N$, the number of p.m.f.'s being compared. However, the empirical distributions—calculated using the JSD and mean entropy for each paper in our training set—show that there are effective upper bounds for both quantities.

**Author Influence.** Our working hypothesis is that authors with a better position in their network of coauthors will tend to coauthor papers with higher citation counts. By a better position, we mean (i) knowing more authors interested in reading/citing their papers, or (ii) knowing more authors who themselves produce highly cited works.

We quantify these effects by computing the PageRank [8] of each author in two similar but slightly different coauthorship graphs. In the first such graph, each vertex is an author, and two authors are joined by an edge if they have coauthored at least one paper together. Identifying authors with their DAIS IDs, the graph has $N_A = 3,092,291$ vertices and $48,590,127$ edges. Using sparse numerical linear algebra, we form the graph's adjacency matrix $A$ and degree matrix $D$. We then consider the Markov chain in which at each step, with probability $d$, a walker follows a simple random walk on the graph with adjacency matrix $A$, while with probability $1-d$, the walker chooses the next vertex uniformly at random from all vertices. The transition matrix for this Markov chain is

$$M = dAD^{-1} + (1-d)(N_A)^{-1}\mathbf{1},$$

where $\mathbf{1}$ stands for an $N_A \times N_A$ matrix of all 1's, and $D$ stands for the degree matrix. The matrix $D$ is purely diagonal; the $i$-th entry on the diagonal is the sum of the entries in the $i$-th column (equivalently, row) of $A$.

We then compute the equilibrium or stationary probability distribution for the Markov chain: the vector $\boldsymbol{\pi}$ satisfying $M\boldsymbol{\pi} = \boldsymbol{\pi}$. We do this through power iteration, i.e., by computing $\boldsymbol{\pi}^{j+1} = M\boldsymbol{\pi}^j$ until $\|\boldsymbol{\pi}^{j+1} - \boldsymbol{\pi}^j\| < 10^{-8}$. We compute the matrix-vector product $M\boldsymbol{\pi}^j$ without storing the dense matrix $M$. As an initial guess, we let $\boldsymbol{\pi}^0$ be the uniform distribution on all $N_A$ vertices. In practice, we observe convergence in less than 60 iterations. For consistency with the literature, in the above Markov chain, we choose $d = 0.85$.

The second coauthorship graph is a weighted version of the first coauthorship graph. In the first graph, we set $A_{ij} = 1$ when authors $i$ and $j$ have coauthored at least

one paper together. In the second graph, we introduce an edge weight based on the decile of the citation count of the paper that authors $i$ and $j$ have coauthored. These deciles correspond to the following citation counts (rounded down): $\mathbf{d} = (0, 2, 7, 13.8, 22.5, 28.6, 40, 63.6, 102, 321)$. If the paper coauthored by authors $i$ and $j$ has $c$ citations, then we set $A_{ij} = \min_j \{c \geq d_j\}$. In this scheme, even a paper with 0 citations obtains a weight of 1. We therefore distinguish between pairs of coauthors who have authored a paper together versus those who have not, even if the paper in question has never been cited.

Using this adjacency matrix in place of the earlier one, we obtain a different Markov chain and a different equilibrium distribution $\widetilde{\boldsymbol{\pi}}$. The vectors $\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\pi}}$, contain the PageRank and the PageRank-w (short for PageRank on the weighted graph) for each author. For each paper, we compute *avgpr* and *maxpr*, the average and maximum PageRank for its authors, respectively. We do the same for PageRank-w, producing *avgprw* and *maxprw* for each paper.

**Title Words.** There are good reasons for why one might view a paper's title as predictive of the paper's success. Readers may decide whether to read a paper based on its title. In 2015, we rarely read printed, bound copies of journals—instead, we search for papers in databases. A paper's title strongly influences whether it appears in search results on a particular topic. Finally, a sudden burst of activity in a particular research area can sometimes yield a collection of highly cited papers that all have similar words in their titles, e.g., "nonnegative matrix factorization." In short, analyzing paper titles is a first step towards analyzing the actual content of the paper.

For each title, we convert all words to lower-case, remove all numbers and punctuation, and then remove stop words—we use a standard list of 127 stop words such as "an" and "of." Next, we apply the Lancaster stemmer [9] from Python's Natural Language Toolkit (NLTK). Aggregating the results across all publications, we arrive at a list of $N_T = 1,099,873$ unique terms. We let $u_j$ denote the number of publications' titles that contain the $j$-th term, for each $j$ such that $1 \leq j \leq N_T$. For each such $j$ and each $i$ such that $1 \leq i \leq N_P$, let $t_{ij}$ denote the number of times term $j$ occurs in the $i$-th document's title. Then we define the $N_P \times N_T$ term-document matrix $S$ using term frequency–inverse document frequency (or TF-IDF) [8]:

$$S_{ij} = t_{ij} \log(N_P/u_j).$$

Now that we have the matrix $S$, we form a second term-document matrix. Let $J$ be a journal from the set of top 495 journals. We take the rows of $T$ corresponding to publications from $J$. We then keep only those rows for which the citation count of the corresponding paper is in the top 90-th percentile of all citation counts for papers from $J$. Summing over these rows, we produce one row that represents the titles from the top papers from $J$. Performing this calculation for each $J$, we obtain a $495 \times N_T$ term-document matrix $R$.

Now let $\hat{R}$ and $\hat{S}$ denote the matrices obtained from $R$ and $S$, respectively, by dividing each row by its 2-norm. Let $\hat{S}^T$ denote the transpose of $\hat{S}$. Setting $C = \hat{R}\hat{S}^T$, we now observe that $C_{ij}$ contains the cosine similarity between the $i$-th row of $R$ and the $j$-th row of $S$. To put this another way, for each $j$ in $1 \leq j \leq N_P$, the $j$-th column of $C$ contains the cosine similarities between the $j$-th publication title and each of the 495 aggregated titles in the matrix $R$. Because $N_P/u_j \geq 1$, we see that $S_{ij} \geq 0$. Hence $\hat{R}$, $\hat{S}$, and $C$ are all nonnegative.

For each publication, we take the maximum value in the corresponding column of $C$ and call that *cossim*, the best cosine similarity score for that publication. The ISSN of the journal corresponding to the row in which the maximum value appears is another feature, *issn*, as is the 5-year impact factor of that journal, *issnboost*. The *cossim* feature models the effect of the words in the title. The computed value records how similar a publication's title words are to title words from one of the top 495 journals. The identity of the journal does not itself figure into *cossim*.

We do not include as a feature the identity of the actual journal in which a publication has appeared. We do this to maximize the potential for our model to be used on papers that have not yet been published. However, we are fully aware that papers that are published in journals from certain fields (e.g., biology and medicine) tend to have higher citation counts than papers published in journals from other fields (e.g., mathematics). The features *issn* and *issnboost* enable us to model this effect. In effect, we are using a paper's title to determine or predict the identity of the journal where, one might say, it ought to be published.

**Classical Features.** Here we include features shown to be predictive in prior work. For each paper, we extract

- *numauth*, the number of authors.
- *numref*, the number of papers cited in the paper's bibliography.
- *age*, the paper's age in years, obtained by subtracting the year the paper was published from 2014, the year the data were obtained.
- *length*, the length in pages.
- *avgcit* and *maxcit*, the average and maximum number of past citations garnered by the authors.
- *avgnp* and *maxnp*, the average and maximum number of prior publications of the authors.

In total, we obtain 2 attributes for author interdisciplinarity, 4 attributes for author influence, 3 attributes from the title words, and 8 classical attributes. Hence we have a total of 17 attributes or predictors for each of the papers in the database. We also extract *tc*, the number of times each paper was cited; this is used to construct the response variable.

## 5. Exploratory Modeling

Equipped with the 17 attributes described above, we first carry out an exploratory analysis. The purpose of this analysis is to gain insight into the attributes and how they might be related to the citation count of a paper.

We first compute the Pearson and Spearman correlation between each attribute and *tc*, the citation count for each

| Feature | $\rho_P$ | $\rho_S$ | Feature | $\rho_P$ | $\rho_S$ |
|---|---|---|---|---|---|
| avgcit | 0.4283 | 0.5775 | issn | 0.0034 | -0.0085 |
| maxcit | 0.3885 | 0.5314 | issnboost | 0.0294 | 0.0903 |
| avgnp | 0.0355 | 0.3847 | age | 0.0234 | 0.0455 |
| maxnp | 0.0353 | 0.3736 | numrefs | 0.2865 | 0.7285 |
| avgpr | 0.0312 | 0.3788 | length | 0.0135 | 0.5382 |
| maxpr | 0.0388 | 0.3825 | numauth | 0.0511 | 0.0955 |
| avgprw | 0.0427 | 0.4090 | jsd | 0.1593 | 0.4019 |
| maxprw | 0.0468 | 0.4012 | ent | 0.1691 | 0.4167 |
| cossim | 0.0788 | 0.0334 | | | |

TABLE 1. CORRELATION—IN BOTH PEARSON ($\rho_P$) AND SPEARMAN ($\rho_S$) SENSES—OF EACH ATTRIBUTE WITH CITATION COUNT.

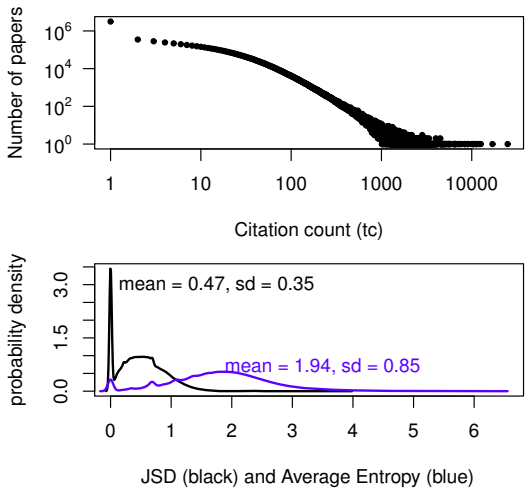


Figure 1. Top: empirical probability mass function of citation counts, plotted on a log-log scale. Bottom: kernel density estimates using the JSD and mean entropy for all papers in our database.

paper. The results are summarized in Table 1. The results show that many of the new features we have included are indeed related to the citation count. In particular, we find reasonable Spearman correlation for author interdisciplinarity as modeled by JSD and mean entropy, as well as author influence as encoded by *avgprw* and *maxprw*.

Next, we turn to the empirical probability mass function (p.m.f.) of *tc* itself. In the top panel of Figure 1, we plot this p.m.f. on a log-log scale; the resulting plot matches prior work on a different data set [6]. We conjecture that the heavy tail of this distribution will lead to occasional, massive errors in any regression model that seeks to predict a paper's exact citation count. Therefore, in this work, we consider classification problems in which we predict the interval in which the citation count will fall. This is described in more detail in Section 6.

The bottom panel of Figure 1 shows kernel density estimates for the probability density functions of JSD (black) and mean entropy (blue). These densities are computed using the JSD and mean entropy values for each paper, not each author. The spike at 0 for the *jsd* density is due to single-author papers.

As we saw above, though JSD and mean entropy are correlated with the citation count, the correlation is not extremely close to 1. Therefore, we attempt to visualize the relationship between these variables and citation count. Our first set of graphs are box-and-whisker plots, shown in Figure 2. We form 10 equispaced bins of mean entropy and JSD that cover the respective ranges of these variables. The horizontal axes of these plots list the centers of each bin. We then plot a box-and-whisker on a log scale for the citation counts that fall into each bin. The top, middle, and bottom horizontal lines in each box indicate the upper quartile, median, and lower quartile of the citation counts in the corresponding bin. The whiskers are located 10 times the interquartile range away from the nearest box.

There are three points to make regarding these plots. First, both mean entropy and JSD are related to citation count in a nonlinear way. Median citation counts are higher when mean entropy is near 3, while median citation counts are higher when JSD is near 2.54. However, citation counts tend to drop off when mean entropy or JSD is either too low or too high. We interpret this as saying that a moderate amount of interdisciplinarity (higher than the average as shown in the marginal densities plotted in Figure 1) seems to be related to a higher citation count. Second, we note the presence of a large number of outliers (points outside the whiskers) in almost every bin. This echoes the heavy-tailed distribution of citation counts described above. Finally, there are a huge number of 0 citation count papers not shown. The number of such papers is sufficiently high to push the mean citation count close to zero in each bin.

Next we explore the multivariate relationship between both measures of author interdisciplinarity and paper citation count. In the top panel of Figure 3, we plot one point for each journal in our top 495 subset. The mean entropy and JSD values for each journal are the averages of the corresponding values over all papers (in our database) that were published in that journal. In this plot, the color of the point indicates impact factor, with red and blue indicating high and low impact factors, respectively.

In the bottom panel of Figure 3, we produce a plot that summarizes the relationship between mean entropy, JSD, and citation count for all $N_P > 7.9$ million papers in our database. Here the color of the point indicates citation count, with red and blue signifying high and low citation counts, respectively. However, for visualization purposes, we have divided the overall (*jsd*,*ent*) space into a $100 \times 100$ grid of rectangles. We have averaged the citation count in each rectangle and plotted one point at each rectangle center.

Overall, the results indicate that journal and paper impact depend jointly on both interdisciplinarity metrics. For a journal, if either mean entropy or mean JSD is sufficiently high, the impact factor of the journal tends to be high. For a paper, we find two hot spots (patches of red in the bottom plot in Figure 3). The patch towards the left corresponds to papers with authors who each possess moderate intrinsic interdisciplinarity (mean entropy near 3) while being similar to one another (JSD near 0.25). The patch towards the right corresponds to papers with authors who are significantly different from one another ($1 \leq \text{JSD} \leq 2$). For these papers, we find it is probable that authors garner a relatively large

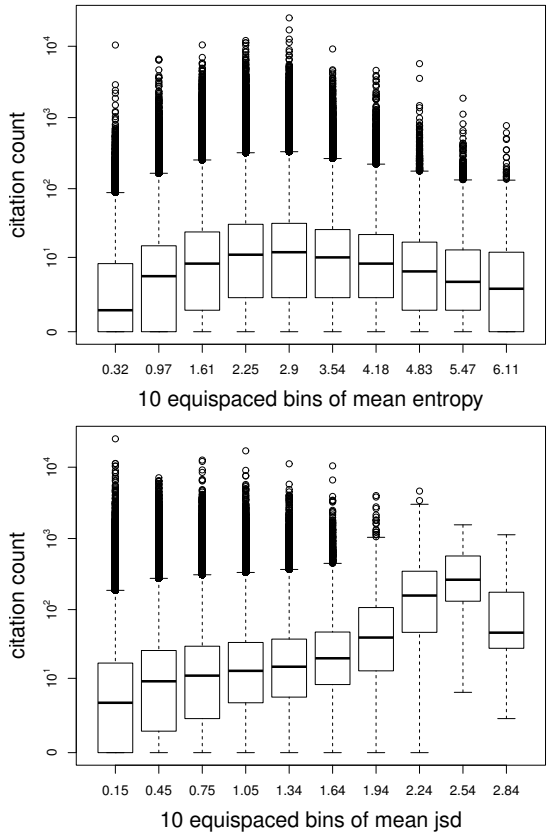


Figure 2. Box-and-whisker plots show a nonlinear dependence of citation count on mean entropy and JSD.

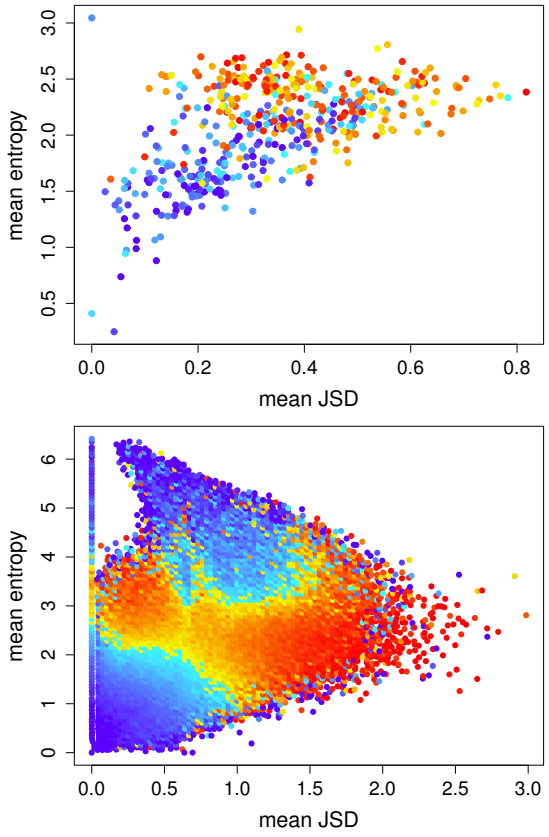


Figure 3. Top: Relationship between average entropy, JSD, and impact factor for the top 495 journals. Red indicates high impact factor (max of 153), while blue indicates low impact factor (min of 2.6). Bottom: Relationship between average entropy, JSD, and citation count for all papers in the training set. Red indicates high citation count (max of 4147.5) while blue indicates low citation count (min of 0).

number of citations even when they possess an average level of intrinsic interdisciplinarity (mean entropy near 2).

## 6. Predictive Modeling

Based on the findings from Section 5, we converted the problem to a classification problem. We considered a 2-class problem, in which class "0" consists of papers with zero citations and class "1" consists of the complement.

Additionally, we considered a 3-class problem. Here we use the approximate 33-rd and 66-th percentiles of the citation distribution, 0 and 12, to form three classes: papers with zero citations (label "0"), papers with citation count $c \in [1, 12]$ (label "1"), and papers with more than 12 citations (label "2").

For the 2-class problem, the training data is imbalanced, with approximately $2/3$ of the instances in class "1." We have deliberately constructed the 3-class problem so that the training data features an approximately equal number of instances of each class.

For each problem, we first used the entire training set to fit a variety of models. To build our models, we used Apache Spark [10] running on HP workstations with 12–24 cores and 16 GB of RAM. The following table summarizes the best training accuracy we achieved with each of the following models/algorithms:

| Model | 2-class Accuracy | 3-class Accuracy |
|---|---|---|
| Naïve Bayes | 0.645 | 0.499 |
| Logistic Regression | 0.851 | 0.665 |
| Tree | 0.884 | 0.723 |
| Support Vector Machine (SVM) | 0.677 | X |
| Random Forest | 0.874 | 0.704 |
| Boosted Trees | 0.859 | X |

TABLE 2. RESULTS FROM TRAINING VARIOUS MODELS ON THE ENTIRE TRAINING SET. X'S INDICATE UNAVAILABILITY OF THE CORRESPONDING ALGORITHM FOR MULTICLASS PROBLEMS IN APACHE SPARK V1.3.0.

We have ordered the results by how long it took to fit each model, with Naïve Bayes requiring only a few seconds and the ensemble methods (forests and boosting) requiring over an hour depending on particular hyperparameter values.

While we did try different hyperparameter values (number of boosting iterations, number of trees in the forest, etc.) for the different models, we did not delve into the Apache Spark code to change more fundamental details in

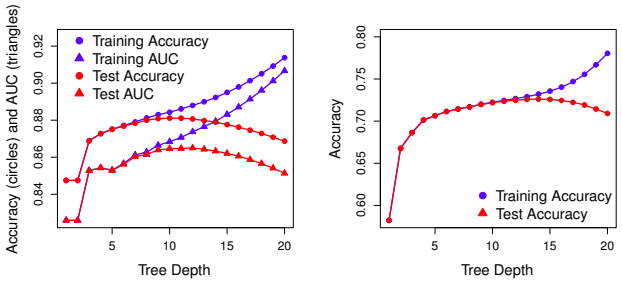


Figure 4. Results from 10-fold cross-validation reveal the values of tree depth beyond which the model overfits the data.

these algorithms. Boosting, in particular, refers to a family of algorithms, and it is entirely possible that a different flavor of boosting may yield a significant improvement over the boosting and single tree results presented here. However, advantages of the tree model are that its construction is relatively fast, the final model is easily interpretable, and the algorithm used to build the tree is completely standard. Additionally, using a single tree has not resulted in any significant reduction in accuracy as compared to an ensembles of trees (whether in a forest or a boosted model). Based on all the factors, we decided to proceed using single tree models.

Note that in our tree model, we treat *issn* as a categorical variable. We also choose the entropy splitting criterion rather than the Gini index.

**Cross-Validation.** Our next step was to select the optimal value of the depth of the tree, a proxy for the overall complexity of the classifier. To do this, we carried out 10-fold cross-validation. We first divided the training set into 10 random subsets or folds. In turn, we used each subset of 9 folds for training and the held out fold for testing. Aggregating the results over all 10 folds, and carrying this out for both 2- and 3-class problems, we obtain the results shown in Figure 4. Please note that in Figure 4, when we write "test," we mean the held out folds from cross-validation, not the true test set described in Section 3. Also note that AUC is short for "area under the ROC curve," a metric of classifier performance that we only compute for the 2-class problem.

As expected, the results show that training accuracy and training AUC increase monotonically as a function of tree depth: a more complex model fits the training data better. However, the test error starts decreasing when the model begins to overfit. For the 2-class problem, test accuracy (respectively, AUC) is maximized at a tree depth of 10 (respectively, 12). For the 3-class problem, test accuracy is maximized at a tree depth of 14.

**Test Set Results.** Based on the cross-validation results and our preference for parsimony, we choose a tree depth of 10 for the two-class problem. With this parameter set, we train the classifier on the entire training set. The resulting tree has 1731 nodes. We make predictions on the test set—the true test set described in Section 3—and compare these predictions against the true class labels. Before we give the results, we must explain two points.

First, suppose we encounter an author in the test set who does not appear in the entire training set. In this case, we have no way of estimating the author's journal distribution, PageRank, or past citation/publication record. Therefore, we ignore this author for the purposes of calculating a number of features: *avgcit*, *maxcit*, *avgnp*, *maxnp*, *avgpr*, *maxpr*. These features are computed using information on the authors who appeared in the training set. If a paper in the test set contains no such authors, we assign a value of 0 to the features just mentioned.

Second, suppose we encounter a title word in a test set paper that does not appear in the entire training set. Again, we ignore the word for the purposes of scoring the title's similarity to our reference set of papers from the top 495 journals. We have not encountered a situation where a paper has zero title words in common with title words found in our training set.

With these points in mind, the test set confusion matrices for the two-class tree model (left) and the logistic regression model (right) are

$$\begin{bmatrix} 228637 & 47511 \\ 27868 & 317015 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 227323 & 48825 \\ 37225 & 307658 \end{bmatrix}, \qquad (3)$$

with respective accuracies of 0.878 and 0.861. We have included the logistic regression results because, in our first models, the logistic regression model performed almost as well as the tree model (see Table 2).

On the test set, the tree and logistic regression models yield AUC values of 0.874 and 0.858, respectively. The areas under the precision-recall curves are 0.917 and 0.908 for the tree and logistic regression models, respectively. Taken together, the results indicate that the tree model is slightly but consistently superior to the logistic regression model for the 2-class problem.

Moving to the 3-class problem, based on the cross-validation results, we choose a tree depth of 14. We fit a tree of this depth using the entire training set; the resulting tree has 23321 nodes. When we make predictions on the test set, this tree model has the following confusion matrix:

$$\begin{bmatrix} 229690 & 31836 & 14622 \\ 27312 & 66881 & 119940 \\ 1433 & 14435 & 114882 \end{bmatrix}. \qquad (4)$$

The accuracy is 0.663. In comparison, the logistic regression model for this problem yields an accuracy of 0.611. For the 3-class problem, we conclude that the tree model is superior.

We do note from (4) that our model tends to misclassify papers with a moderate citation count (class 1, the second row) as highly cited papers (class 2, the third column). In other words, if the classifier says that the paper is in class 2, there is almost a 50/50 chance that the true class of the paper is 1 or 2. However, we note that the model succeeds in identifying papers that have zero citations (the upper-left corner of the matrix, corresponding to class 0).

**Attribute Importance.** Starting with the 2-class classifier, we knock out one attribute at a time, refit the model,

| | Attribute | Acc. | Attribute | AUC | Attribute | Acc. |
|---|---|---|---|---|---|---|
| 1 | numrefs | .8287 | numrefs | .8229 | numrefs | .6347 |
| 2 | numauth | .8698 | numauth | .8628 | jsd | .6601 |
| 3 | age | .8731 | maxcit | .8680 | numauth | .6603 |
| 4 | maxcit | .8747 | avgcit | .8680 | length | .6627 |
| 5 | avgprw | .8748 | avgprw | .8681 | ent | .6627 |
| 6 | avgcit | .8749 | issn | .8682 | issnboost | .6639 |
| 7 | issn | .8750 | avgpr | .8684 | age | .6710 |
| 8 | avgpr | .8752 | avgnp | .8687 | avgnp | .6739 |
| 9 | avgnp | .8753 | maxprw | .8689 | maxpr | .6740 |
| 10 | maxprw | .8754 | maxnp | .8693 | issn | .6762 |
| 11 | maxnp | .8759 | cossim | .8694 | avgcit | .6763 |
| 12 | cossim | .8760 | maxpr | .8696 | avgprw | .6770 |
| 13 | maxpr | .8760 | age | .8700 | maxcit | .6778 |
| 14 | issnboost | .8774 | issnboost | .8723 | maxprw | .6780 |
| 15 | length | .8786 | length | .8734 | maxnp | .6789 |
| 16 | ent | .8787 | jsd | .8736 | avgpr | .6798 |
| 17 | jsd | .8788 | ent | .8736 | cossim | .6819 |

TABLE 3. WE PRESENT ATTRIBUTES SORTED BY IMPORTANCE FOR THE 2-CLASS (RESP., 3-CLASS) MODEL TO THE LEFT (RESP., RIGHT) OF THE DOUBLE VERTICAL LINE. IMPORTANCE IS MEASURED BY THE DROP IN MODEL PERFORMANCE WHEN A GIVEN ATTRIBUTE IS OMITTED.

and examine its performance on the test set. We then sort the results in order of increasing accuracy/AUC; the idea is that the most important attribute should, when omitted, cause the greatest decrease in test set accuracy/AUC.

The results of this procedure are given to the left of the double vertical line in Table 3. As in many prior studies in this area, the number of references cited by the paper is the most important attribute to predict its citation class. Among the non-classical features introduced in this paper, the most important ones appear to be *avgprw*, *issn*, and *avgpr*. These correspond to our models of author influence and the most similar journal based on title words. Note, however, that JSD and mean entropy rank last.

We repeat the procedure for the 3-class tree model and display the results to the right of the double vertical line in Table 3. There are two interesting findings here. First, JSD and mean entropy now occupy much higher positions in the ranking. This is in concert with our earlier findings in Section 5 that JSD and mean entropy are both related to the citation count of a paper. Second, note that omitting *cossim* actually causes the accuracy of the model to increase beyond the accuracy we reported when using all 17 attributes. The increase in accuracy from $0.663$ to $0.682$ is a relative increase of roughly $2.8\%$. Overall, we believe this indicates that further work on the tree model, in terms of pruning and variable selection, should improve performance.

## 7. Discussion

We find that our measures of entropy and JSD contribute to predictive models of citation count, our measure of a paper's impact. Importantly, we find this despite the many variables we have included in the predictive models. It should be noted that the 2-class models do not turn up this predictive contribution of our interdisciplinarity measures, though the 3-class models indeed show the reverse–interdisciplinarity measures are near the top of contributing variables. It seems likely to us that the nonlinear relationship we see above between entropy and JSD with citation count may relate to these basic patterns. Given the nonlinearity of the relationship, the interdisciplinarity features may better carve out the space of citations when the citation count is rendered as a 3-class response rather than a 2-class response.

There is a parallel phenomenon in studies of how individual people reason and make decisions with diverse evidence [11], [12]. In general, arguments based on diverse sources of evidence are considered more compelling than arguments based on homogenous evidence. However, the overall relationship between diversity and argument strength is nonlinear: arguments with extremely diverse evidence are considered less convincing. More broadly, efforts to improve understanding of team science and to create successful teams may be informed by studies of how to put together multiple sources of evidence in a compelling way. New scientific teams may be assessed for their potential to combine research with the right diversity of past work. The result, we think, may bridge two domains: what draws the attention of other researchers who evaluate and cite the work of others, and the citation patterns that emerge at scale.

## References


[1] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *String Processing and Information Retrieval*, ser. Lecture Notes in Computer Science, N. Ziviani and R. Baeza-Yates, Eds. Springer Berlin Heidelberg, 2007, vol. 4726, pp. 107–117.

[2] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *Proceedings of the Twelfth SIAM International Conference on Data Mining*, 2012, pp. 1119–1130.

[3] T. Yu, G. Yu, P. Li, and L. Wang, "Citation impact prediction for scientific papers using stepwise regression analysis," *Scientometrics*, vol. 101, no. 2, pp. 1233–1252, 2014.

[4] N. Pobiedina and R. Ichise, "Predicting citation counts for academic literature using graph pattern mining," in *Modern Advances in Applied Intelligence—27th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2014, Proceedings, Part II*, 2014, pp. 109–119.

[5] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting citation count of *Bioinformatics* papers within four years of publication," *Bioinformatics*, vol. 25, no. 24, pp. 3303–3309, 2009.

[6] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011*, 2011, pp. 1247–1252.

[7] S. M. Fiore, "Interdisciplinarity as Teamwork: How the Science of Teams Can Inform Team Science," *Small Group Research*, vol. 39, no. 3, pp. 251–277, 2008.

[8] L. Eldén, *Matrix Methods in Data Mining and Pattern Recognition*. SIAM, 2007.

[9] C. D. Paice, "Another stemmer," *ACM SIGIR Forum*, vol. 24, no. 3, pp. 56–61, 1990.

[10] S. Ryza, U. Laserson, S. Owen, and J. Wills, *Advanced Analytics with Spark*. O'Reilly, 2015.

[11] E. Heit, "Properties of inductive reasoning," *Psychonomic Bulletin & Review*, vol. 7, pp. 569–592, 2000.

[12] E. Heit, U. Hahn, and A. Feeney, "Defending diversity," in *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin*, W. Ahn, R. Goldstone, B. Love, A. Markman, and P. Wolff, Eds. Washington, D.C.: APA, 2005, pp. 87–99.