Noname manuscript No.
(will be inserted by the editor)

Predicting Citation Patterns

Defining and Determining Influence

David Guy Brizan[1*] · Kevin Gallagher[2] ·
Arnab Jahangir[3] · Theodore Brown[1]

Received: date / Accepted: date

Abstract Definitions for *influence* in bibliometrics are surveyed and expanded upon in this work. On data composed of the union of DBLP and CiteSeer^x, approximately 6 million publications, a relatively small number of features are developed to describe the set, including *loyalty* and *community longevity*, two novel features. These features are successfully used to predict the influential set of papers in a series of machine learning experiments. The most predictive features are highlighted and discussed.

Keywords Citation Analysis · Bibliometrics · Big Data · Machine Learning

1 Introduction

Research work that revolutionizes a field are few and far between; most scientific advances are built upon earlier work by others. But even an entirely new idea creates a progression of further scientific literature.

This paper examines the predictability of the future "impact" or influence of a paper in a number of new ways. The question asked here is: are there factors that allow us to predict with some confidence that in the future a paper

This research was supported, in part, under National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113 and the City University of New York High Performance Computing Center at the College of Staten Island.
The authors also acknowledge the Office of Information Technology at The Graduate Center, CUNY for providing database and server resources that have contributed to the research results reported within this paper. URL: http://it.gc.cuny.edu/

[1] Department of Computer Science, CUNY Graduate Center, 365 Fifth Ave, New York, NY 10016
[2] Department of Computer Science, NYU Tandon School of Engineering, 6 MetroTech Center, Brooklyn, NY 11201
[3] Department of Computer Science, Hunter College CUNY, 695 Park Avenue, New York, NY 10065
* Corresponding author: dbrizan@gradcenter.cuny.edu

will be influential. One way of measuring the impact of a paper historically is to examine its citation history. There are many ways of doing this. For example, the length of longest number of years that a paper has been cited at least some number of times, x, the total number of papers that cite a paper, or fitting a curve to the count would be another way. Each has its merits.

Presented are three ways to predict the possible future impact of a paper and then machine learning techniques are used to tease out what factors (called features in machine language) allow us to predict the influence of a paper based on these characterizations of influence. The influence factors that are measured and defined more carefully below are: the count of the number of citations, the time length of the citations, and how many different communities (different areas of research), each as defined below. Each is measured within what is called here a "cohort," so as not to bias the results by field, since different fields have different citation characteristics.

The data for this research come from citation sources of primarily computer science publications, DBLP and CiteSeer in 2010–2011. Over 11 million citations are in the database used and over 6.0 million publications.

2 Related Work

Authors use citations to "establish new conceptual relationships between the current work and any earlier item cited" [23]; however, since the volume of publications is too vast for any one scholar to read [15], authors tend to focus their work on only a subset of the papers available to them. As far back as 1926, Alfred Lotka [14] discovered a Zipfian distribution of references to articles: many publications cited a few times and a few cited many times.

In the *Universalist* view of bibliometrics, the best articles–those with a greater contribution to science–are the ones which are cited. Others are naturally ignored. Since a citation is viewed as a small reward for good research work bestowed by peers who are also experts, it has been argued that authors who accumulate a number of influential papers ultimately receive larger awards such as invited lectures, Presidential Addresses [27] and, for a lucky few, a Nobel Prize [15]. In the Universalist view, a paper's contribution to scholarship can be measured by its citation count; the authors, journals and institutions associated with those papers likewise can be rewarded for their associations. This has led to related metrics, briefly reviewed below.

Squarely in the universalist view is [6]. The authors of this paper examine the impact of biomedical articles using a model to predict the number of citations within the next 10 years. Their corpus was under 4000 articles, compared to our much larger set. Their feature selection was much wider since they were dealing with several very different types of medical articles.

Impact Factor [23] has been used to compare journals. The Impact Factor is a measure of the average number of times papers are cited within two years of a particular date. For example, the impact factor for a journal in the year 2010 is x/y where x is the total number of citations the journal received in

2008 and 2009 and y is the total number of citable items published in the journal in 2008 and 2009.

Authors have been compared with the h-index [10], perhaps because its calculation is easy: an author has an h-index of h if she has published h papers, each of which has been cited at least h times. Hirsch went on to show [11] that h-index is more predictive of future research output than some other citation-based metrics, such as total number of citations, total number of papers, or mean citations per paper. We note that the h-index measures historical citations and that Hirsch himself made no claims about predicting the impact of any single publication by an author.

Several improvements have been suggested to the H-index. For example, the g-index [4] is determined by the top g articles which received a total of g^2 citations collectively. Compared to h-index, it allows a highly-cited publication to increase the author's reputation. In addition to improvements, the h-index has also been extended to institutions as, for example the h_2 index, which can describe an institution which as at least h_2 individuals, each with h-index of at least h_2 [16]. Metrics similar to h-index have also been extended to author networks [21].

However, many with a *Particularist* view have noted that, rather than the work, characteristics of the author is what draws citations to a paper. This has been described as the "Matthew Effect," which named for the biblical passage which describes the rich getting richer, often at the expense of the poor. In terms of academic rewards, it has been used to describe two manifestations of the same phenomena, namely unequal recognition for the same amount of work due to authors' social or demographic characteristics. The first manifestation of the Matthew Effect occurs when a larger share of credit for the ideas in a publication is given to a senior co-author whose contribution is relatively small in comparison to other authors. While the issue of credit for authorship may still abound, one convention [26] lists the first author as the person with the largest contribution, the last as the most senior researcher and others by alphabetic order or by order of contribution.

The more common manifestation occurs when citations become concentrated on one publication when the same work was performed and published elsewhere. This has been attributed to the quantity of research being too voluminous for any one author to consume [15] or to benign but important demographic factors such as language, country or research focus [27]. In an important critique of the "meritocracy" of science, the Matthew Effect has been shown to result in a gender bias (the "Matilda effect") [20], with women receiving less credit than their male counterparts for the same work.

In terms of goals and approaches, our work is closest to that of van Dalen and Henkens (2001) [27], Judge et al. (2007) [12], Haslam et al. (2008) [9] and Newman (2009) [18]; all extract features about funding, sponsoring institutions, articles, journals and authors to determine what has caused some articles to garner a large number of citations. Newman goes on to make predictions about which publications will enjoy this success in the future, which, as shown in later work [19], was largely accurate. While each study uses different

data, examines different features and come to slightly different conclusions, the majority conclude that the journal in which an article is published strongly correlates with the number of citations it will receive. Haslam et al. show that some other features (for example, author institution, prestigious funding and length of article) predict acceptance to the more prestigious journals, which in turn predicts higher citations.

Almost all found one or more universalistic features to be important, such as the presence of rhetorical devices – tables, figures, colons in the title – theoretical focus or methodological rigor as well as volume and recency of references [9]. Interestingly, Haslam et al. found a lack of gender bias in their study. Two of the studies (van Dalen and Hankens as well as Newman) show a *first mover* advantage: a correlation between high citations and being the first research in a sub-discipline.

Almost all these experiments resulted in findings that particularistic features are just as important as universalistic ones, including the order in the journal (first being better than last) [27] [12]. Haslam et al. demonstrated that the reputation of the first author should be high but should have another author with an even higher reputation in order to increase the likelihood of being cited.

Many of these studies use features derived from human-annotated sources to describe or predict the number of citations to a paper. Likewise, many of these studies use a small number (hundreds or thousands) of samples. Our objective is to make the same prediction about citation success with a completely automated system on a large scale.

3 Approach

In our review of the work in bibliometrics, we find that influence is generally derived from the count of references received. We expand on and refine this definition in this section. First, we define a quantifiable metric for comparing publications among different times and different disciplines.

For the purpose of this work, a "*publication*" is a single piece of work be it be a book, an article in a journal or an article in a conference. A "*community*" is a set of journal issues or conferences with the same title. And a "*cohort*" is a series of journal issues or conference titles from a single year. In this case, the cohort for a conference would most likely be represented by the single conference but a journal could have several issues within the same year.

3.1 Quantifying Influence

Although the raw counts of citations received by two publications may be compared directly, this technique is inadvisable. Because more recent publications have had fewer opportunities to be read, we expect recent publications to be referenced fewer times than older work. In the same vein, there are differing

numbers of active researches in the myriad of disciplines and sub-disciplines of research, leading to a range of opportunity for a publication to be referenced.

Because we expect a range of distributions of references due to the effects of time and discipline, the direct comparisons of the count of references among publications should only occur for similar cohort's publications. In comparing publications across cohorts, we apply a technique inspired by Shi et al., 2009 [24]: publications are ranked within each cohort to determine how influential each one is.

We say that a publication becomes *influential* when its influential value, defined to be the percentile of the influence value of its cohort is higher than or equal to others. However, regardless of the norms and variances within a cohort, an uncited publication is never considered influential. These are defined precisely in Section 3.2 with formulas.

3.2 Three Measures of Influence

Each publication π can be described by certain attributes, including a date of publication, $date(\pi)$, the cohort set in which it was published $cohort_\pi$, and a (possibly empty) set of references to that publication, R_π. Each reference to π is made on a date equal to the citing publication's date of publication.

Given our method of quantifying influence in Section 3.1, we see publications as influential in three ways. The most commonly used definition of influence is by volume of its citations. This is the basis behind metrics such as H-index and journal impact factor. Another measure of influence is the "staying power" of a publication, seminal work which has been referenced continually for an extended period of time. Alternately, "staying power" could be that the publication was rediscovered years after being published can be said to be influential. Together, these two define a influence by *l*ongevity. Finally, a third way that a publication can be considered influential, is if it has influence over a wide variety of fields or sub-disciplines.

We calculate the Volume-based degree of influence of a publication π as shown in (1).

$$Volume(\pi) = \frac{\mid R_\pi \mid}{\sum_{c \in cohort_\pi} \mid R_c \mid} \tag{1}$$

Similarly, we calculate the Longevity-based degree of influence of any publication π by (2).

$$Longevity(\pi) = \frac{maxdate(R_\pi)) - date(\pi)}{\sum_{c \in cohort_\pi)}(maxdate(Rc)) - date(c))}, \tag{2}$$

where maxdate in set x is $maxdate(x) = max_{y \in x}(date(y))$

Finally, we calculate the Diversity-based degree of influence of a publication π as shown in (4).

Let $U(z) = 1$, if $z = true$ and 0 if $z = false$.

$$d(y) = \sum_{x \in R_y} \mid U(community(x) \neq community(y)) \mid, \tag{3}$$

$$Diversity(\pi) = \frac{d(\pi)}{\sum_{c \in cohort_\pi} d(c)} \tag{4}$$

Note that all our variables are in the range [0,1].

4 Data

We acquired data from two sources: DBLP and CiteSeer^x. Using a MySQL database, we imported data for publications, authors and communities from these two sources, cleansed the data and performed checks on the data quality. Details of each can be found below. We also performed entity resolution on many records, matching and merging exact or non-exact duplicate records.

4.1 Data Sources, Import and Calculations

In November and December, 2010, we imported data from the DBLP computer science bibliography [13] (DBLP) which contained data on over 1.5 million publications, all in the field of computer science, along with their 882,254 authors and 6,500 communities. We accepted DBLP's entities with respect to publications and authors.

Into the imported DBLP data, we added records from CiteSeer^x [1,7] between January, 2011 and February, 2012. CiteSeer^x records are automatically collected from a number of sources which we find are mostly, but not exclusively, also on the topic of computer science. As with DBLP, we accepted CiteSeer^x' entity resolution for publications. In addition, we imported data on authors where it was present. From CiteSeer^x, we created 5.7 million additional author records and 5.1 million publication records. Importantly, we were able to glean 11.7 million citations among publications from this source. Note that, at the time of data import, DBLP had no records for citations among its publications.

We matched CiteSeer^x data into the existing DBLP data by publication title, publication year as well as authors last names and (where present) first initials. Using these import and merging criteria, we collected a total of 6.4 million authors and 6.0 million publications from the union of both sources, with 222,890 publications and 134,604 authors common to both sources.

In this acquired data, when we found that that two publication records refer to the same publication entity using the technique above, we merged those records, taking the earlier publication date of all records, and we linked the community data. This linking resulted in 133,540 communities for our data. Some publications – books and technical reports, for example – had no

community data. These were placed in "singleton communities" – a community of exactly one publication. There were 1.65 million (27.4%) publications in singleton communities.

On this set of data, we derived several metrics, including H-indices (Hirsh, 2005) for authors and impact factors for journals or conferences.

4.2 Data Profile

In our data set, we find some interesting trends. Firstly, as shown in Fig. 1, the majority of publications in our data set have been published since 1990.

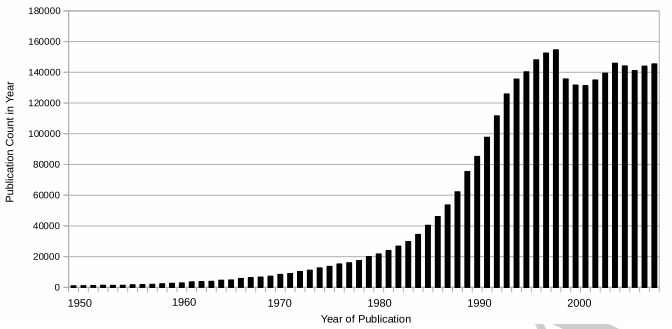

Fig. 1 Publication Count per Year, 1950 - 2008.

We note a steady rise in the amount of published work, especially for the 20-year period before 2000, along with the drop in publications at that year. One of the authors conjectures that industry incentives around the Y2K problem encouraged people out of research and into efforts to fix established systems. That said, we have not attempted to develop this into a hypothesis or subject it to scientific experimentation.

We also determined that over 70% of the authors in our data set have written one publication only. There are a few authors with several hundred publications. At the same time, authors are likely to collaborate on publications. While the plurality of publications have a single author, the majority have two or more authors, and a few publications have several dozen authors. Fig. 2 illustrates both these distributions. Note that the counts along the vertical axes are in the log scale.

Finally, we find that many publications (38.81%) have not been cited. Where a publication is cited, it has the distribution shown in Fig. 3, which is also in log scale along the vertical axis. Most publications are referenced once, but a few – books, for example – are referenced several hundred times.

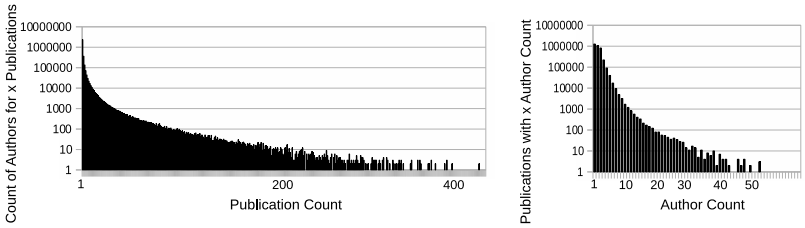

Fig. 2 Publication Count per Author and Author Count per Publication.

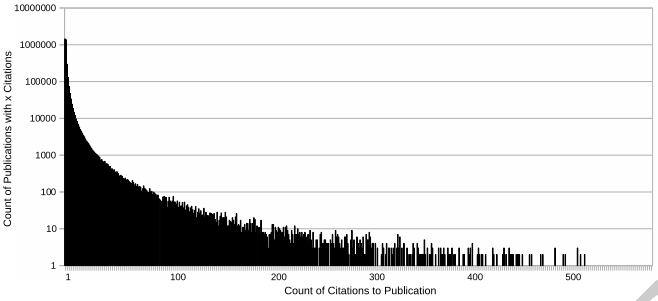

Fig. 3 Publication Count vs. Citation Count.

4.3 Data Quality

The data in our data set is not free from error. Some were introduced by errors or shortcomings our import and merging/linking processes, and some errors can be traced to the sources at DBLP or CiteSeer^x.

Some of the publications we acquired from our two sources have missing or implausible dates. In these cases, we removed the publication from our data quality measurements and from our experiments, described in the next section. We did, however, keep any community (journal or conference) data for other publications. We proceed only with this *plausible set* of data for quality checks and experiments. We estimate the correctness of our data in two ways: by the gap between publication and citation and by the performance in formation of communities.

Despite filtering for invalid dates, it is possible for a publication to be referenced before its publication date, but it should be an uncommon occurrence and should occur within a narrow window of time. We believe this could be a useful as a metric for data quality. On this premise, we examined the difference in years between when a publication appears and when it is first referenced. We find that 2.46% from the plausible set are cited before publication, the majority of which are within one year of publication. Bars below the x-axis show this on Figure 4. This could be caused by our criteria for merging publi-

cations, specifically our policy of accepting the earlier date of conflicting time data, or more likely due to errors in the source.

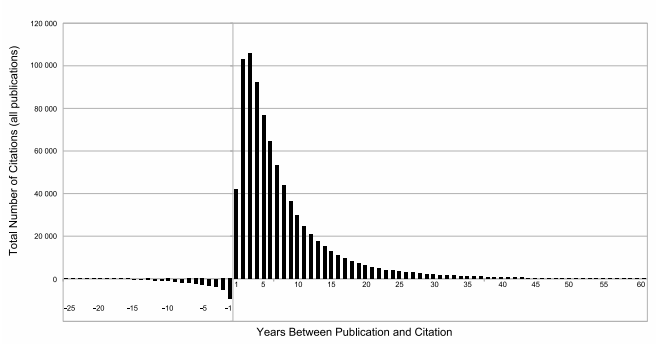

Fig. 4 Time Between Publication and First Citation.

We also sampled the community (journals and conferences) entries for correctness and completeness. One example in the Appendix shows all the aliases for the the Quantitative Evaluation of Systems (QEST) conference and how they are clustered in our database. We measure the "goodness" of clustering in two ways: by *purity* and *entropy*. As explained in [25], purity ranges from 0.0 to 1.0, with higher numbers being better. For entropy, on the other hand, lower numbers indicate better performance.

In our sampling of community entries, we calculated our purity as 1.0, and our entropy as 0.0231. We believe the nature of our clustering algorithm generated very pure community clusters. The dearth of data caused us to create too few bridges as would have been necessary to unify the many community aliases. Still, we are satisfied with the relatively good results here.

5 Experiments

We conducted three series of experiments using machine learning techniques, each series corresponding to our three measures of *i*nfluence: Volume, Longevity and Diversity. For each series, we used a binary threshold of influence varying the threshold between [5%-95%] at increments of 5%. That is, the set of publications was divided into two sets, the top x% as the influential set and (1-x)% not influential. We randomly divided the plausible set of publications – publications with valid dates in their communities (journals or conferences) into two subsets, training and test, with the training set comprising 90% of the data.

We extracted 48 features from each publication (see Section 5.1) and used those features to train models in Weka [8]. The resulting trained models were used to predict the influential publications in the test subset. We experimented with several classification algorithms as implemented in Weka, specifically:

AdaBoost, J48, Naive Bayes, Random Forest and SVM (SMO). We ultimately chose J48 for its balance of performance and speed. We kept values at their default parameter settings.

We make no distinction among different types of references. For example, equal weight is given to citations which use the innovations of previous work, citations which oppose some published research – i.e. "negative citations," which Catalini et al. [2] discusses a technique for finding – or "self-citations" – citations to previous work by the same authors.

5.1 Features

As stated above we derived a total of 48 features from each publication in the data. The features are based on what is known about the publications, authors and venues at the time of publication. All features are listed in Appendix A. These features were used to train models in Weka and to predict the influential subset from the test portion.

Among the features we extracted are *loyalty*, representing an author's tendency to publish in one community and *community longevity*. The author's loyalty to a community calculated as the ratio of the count of all the author's former publications which appear in the community in question compared to the count of author's total publications. Since each publication may have more than one author, we use only the minimum and maximum of all the authors of a publication. An author's community longevity is relative to the time a paper is published. Specifically, it is the length of time between the date of publication and the date of publication of the first paper in the same community. Since we have not found the use of these features in predicting bibliometric influence, we believe loyalty and community longevity are novel features.

5.2 Volume Experiments

In this series, we conducted 19 experiments, one at each 5% increment in the interval [5%-95%]. In each cohort, we ranked the publications in order of total citations received. At each increment, x, a publication was determined to be influential if it was in the top x% according to this ranking and had at least 1 citation.

The purpose of the volume experiments is to expose the publications which the earliest work on bibliometrics ([3], for example) have defined as influential. We note that it is possible for a publication to be considered influential after having very few citations if a number of other publications in the same cohort had even fewer citations. We believe this is still a reasonable strategy for finding the most influential set of publications from a cohort.

5.3 Longevity Experiments

In this series, we conducted 19 experiments, one at each 5% increment in the interval [5%-95%]. In each cohort, we ranked the publications in order of total time between publication and latest citation. At each increment, x, a publication was determined to be influential if it was in the top x% according to this ranking and had at least 1 citation.

This series of experiments finds two heterogeneous sets of publications: those which have been continually cited throughout time and "sleeping beauties," [28] those which have been rediscovered after a period of dormancy.

5.4 Diversity Experiments

In this series, we conducted 19 experiments, one at each 5% increment in the interval [5%-95%]. In each cohort, we ranked the publications in order of total number of communities originating a citation to the publication. At each increment, x, a publication was determined to be influential if it was in the top x% according to this ranking and had at least 1 citation. This series of experiments is designed to find publications which are useful across many different areas.

6 Results

For each series of experiments and for each increment, we determine a label of *influential* or *not influential* for each publication. Our process with Weka produces a confusion matrix such as is shown in Figure 5, which illustrates the difference between the system's predictions and the actual label of ground truth.

Two metrics can be used to indicate the effectiveness of the model: prediction and recall. As shown in Figure 5, in a two-class prediction, such as the ones we propose, there are four possible outcomes when measured against the ground truth – i.e. what is known about the publication's performance:

- True Positive(TP): "Influential" for both ground truth and system prediction
- False Positive (FP): "Not influential" for ground truth with "not influential" system prediction
- False Negative (FN): "Influential" for ground truth with "not influential" system prediction
- True Negative (TN): "Not influential" for both ground truth and system prediction

At the 40% influence as shown in Figure 5, the precision, the probability that a publication is influential, is 85%. Likewise, the recall, the probability that a paper is not influential is at 65%. Both precision and recall performance increase as the levels of influence increase.

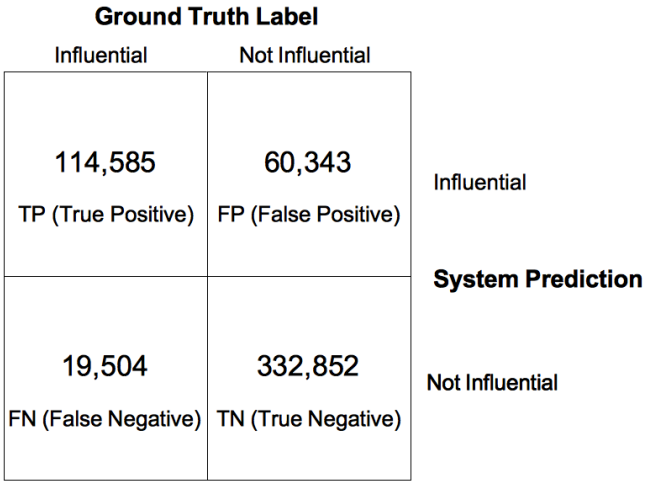

Fig. 5 Confusion Matrix for Volume Experiment at 40% Influence.

We use the confusion matrix to calculate the performance of our system. For each series of experiments, we report baseline accuracy and the system's prediction accuracy at each 5% increment in the interval [5%-95%]. We calculate the baseline (chance) accuracy as shown in (5) and the system's prediction accuracy as shown in (6).

$$Baseline\ Accuracy = \frac{TN + FP}{TP + TN + FP + FN} \tag{5}$$

$$System\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Because we are most interested in identifying the set of influential publications, we also report the baseline F1 score and the system's F1 score. These scores are derived from Precision (7) and Recall (8).

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Given the calculations for Precision and Recall, we derive the baseline (chance) score for the influential set as shown in (9), and we derive the F1 score for this set as shown in (10). We report these for each 5% in the interval [5%-95%] for each series of experiments.

$$Baseline\ F1 = \frac{TP + FP}{TP + TN + FP + FN} \tag{9}$$

$$System\ F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

6.1 Volume

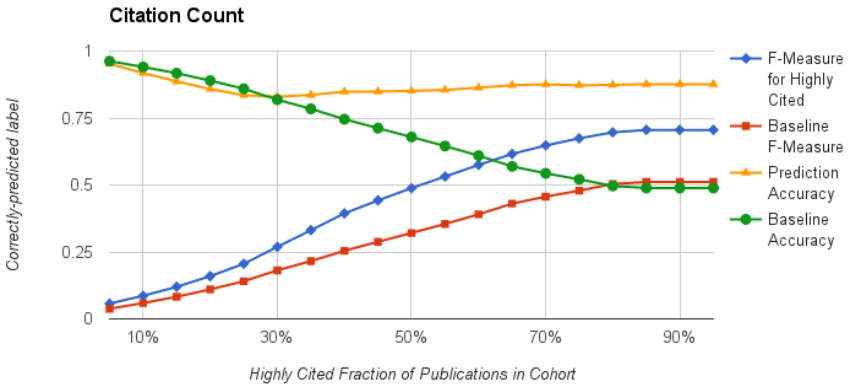

Fig. 6 Volume Experiment Results.

For the Volume experiments, we derived baseline (chance) and system predictions for each 5% increment for degree of influence in the interval [5%-95%]. As can be seen in the plot of these results in Figure 6, baseline accuracy falls steadily in the interval [5%-80%] and remains relatively constant in the interval [80%-95%]. The system's prediction accuracy shows a corresponding fall in a smaller range [5%-25%] and remains constant above that range. Baseline and system accuracy match at a point around 30% influence.

Focusing on the subset of influential publications, we see an expected constant rise in the baseline in the range [5%-80%], above which we see relative stability, mirroring the stability in baseline accuracy. This corresponds to the set of plausible publications which are not referenced.

We believe the performance of the model is high, especially given the relatively small number of features. Importantly, the system prediction of these influential publications remains above this baseline for the entire series of the experiment set, suggesting that our features are effective at identifying this target set.

6.2 Longevity

Figure 7 contains the results of the Longevity series baseline ad system accuracy as well as baseline and system F-scores for the influential set. As with the volume experiments, we see an overall steady decline in baseline accuracy until is levels off at around 80% degree of influence. The baseline F-score exhibits the opposite behaviour – an overall steady rise to the 80% degree of influence point. Overall system accuracy also follows a similar trajectory in

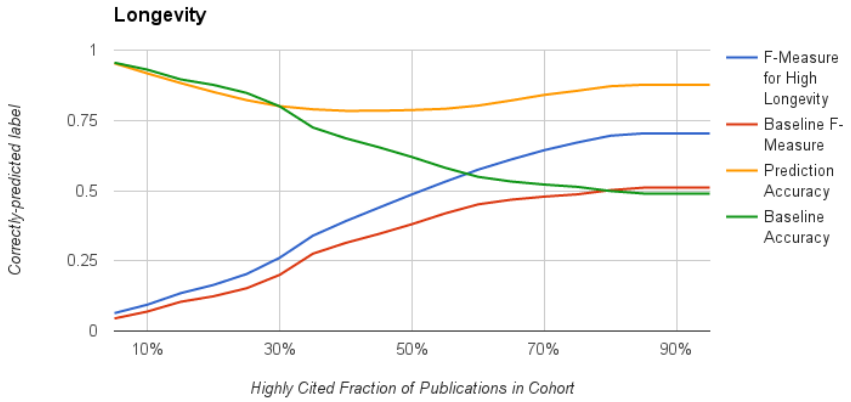

Fig. 7 Longevity Experiment Results.

comparison to the Volume series experiment results: a period of decline in the 5%-25% degree of influence range, followed by relative stability thereafter. Finally, the F-score for the influential set rises steadily through the series and is consistently above baseline.

6.3 Diversity

Figure 8 shows the baseline and system accuracy for the Diversity series of experiments as well as their baseline and system F-scores. Again, we see a pattern closely resembling the Volume experiment results, with high overall system accuracy and a prediction of the highly diverse set consistently above baseline.

6.4 Discussion

As measured by the overall system accuracy and F1-measure for the influential group, our system performance remains high across the three series of experiments. All of our features contributed to this performance, but the top seven highest contributions at the 50% increment are listed on Table 1, along with their relative contributions as measured by Weka's implementation of infogain, with larger numbers indicating a stronger contribution.

We suspected that the relatively similar performance for the three sets of experiments belied similarity in the underlying data sets and their labels. A side-by-side examination of the precision and recall performances could help this. A number of aspects of the recall and precision values shown in Figure 9 are interesting. We note the consistently high performance of the precision values as well as the visual similarity of those values, suggesting that the factors

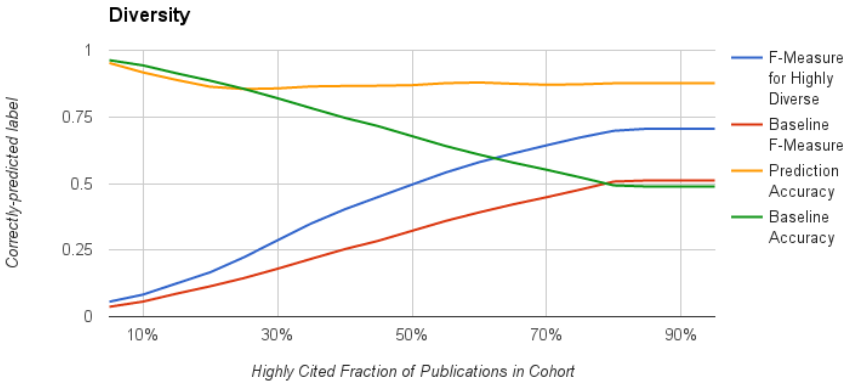

Fig. 8 Diversity Experiment Results.

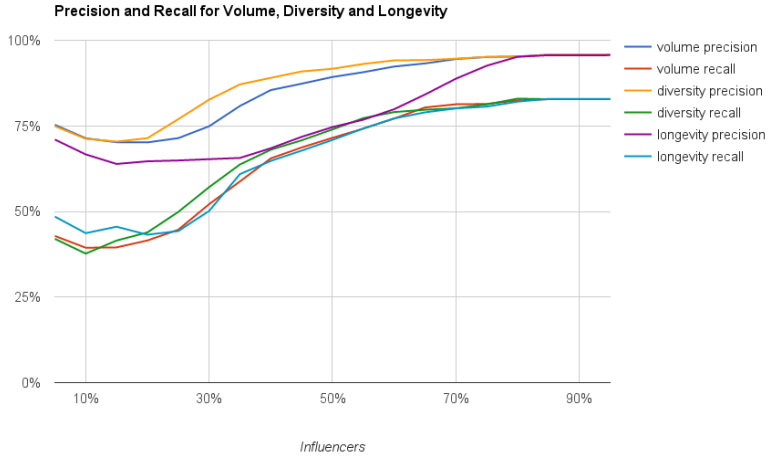

Fig. 9 Precision and Recall for All Experiments.

influencing any of them are the same. This is borne out below. The recall values, although lower, also exhibit similarity across experiments, indicating a possible correlation among publications which are influential because of their volume of citations, because of their longevity and because of their diversity.

We observe that our system performance was high despite the relatively low number of features. Publication ID is consistently the top feature employed by our system across all series and all increments. This is not surprising since it can be used to determine both the source of the data (DBLP vs. CiteSeer^x) and a publication's source community. Community ID and Year Published are also

Table 1 Most Useful Features: Information Gain with Respect to Class.

Rank	Volume	Longevity	Diversity
1	Publication ID	Publication ID	Publication ID
	0.2033	0.1918	0.2024
2	Community ID	Community ID	Community ID
	0.0996	0.1000	0.0993
3	Year Published	Year Published	Year Published
	0.0747	0.0751	0.0744
4	Min Loyalty	Max Longevity	Min Loyalty
	0.0295	0.0264	0.0283
5	Max H-index	Min Loyalty	Max H-index
	0.0290	0.0247	0.0275
6	Max Loyalty	Max Loyalty	Max Loyalty
	0.0272	0.0230	0.0261
7	Max Longevity	Min Community Longevity	Max Longevity
	0.0229	0.0124	0.0234

always in the top three features for our models. We believe this indicates our system essentially builds unique statistical models for each grouping of venue and year. In other words, each cohort has its own set of features and unique weights for those features used in making a prediction of which publications will be influential.

The next tier of features, those frequently ranked in positions 4–7, are therefore even more informative in understanding what positively contributed to an accurate prediction. Not only do we know that our system's predictions would be less accurate in their absence, but we expect that the combination of features yields more fine-grained and accurate predictions as well as better explanatory power. In this secondary tier, we frequently find high loyalty and high longevity of the authors as well as relatively short titles compared to their accompanying longer abstracts all play important roles. This second tier of features is less consistent than the top, so varies depending on the experiment series and the increment.

Many other features have small, positive contributions. H-index of the authors, community longevity, the number of co-authors and number of references all play relatively minor roles compared to the second tier of features, many becoming important when more publications are defined as influential, at the 50% or higher increments. Impact factor of the community and keywords in the title and abstract have almost no contribution.

It is possible that these results indicate that an author develops name recognition after having published frequently, but this recognition need not be accompanied by success of previous work as measured through citations. Pithy titles may entice a reader, and long abstracts may provide a sufficiently adequate summary of a publication so that it need not be read in depth before being cited.

We notice the similarity in the results for the different experiments. We suspected a possible cause may be that the three influence measures served to describe the similar phenomena: a publication which is influential because

it has a large volume of citations is likely to be cited over a longer period of time (longevity) and by different research communities (diversity) as well.

7 Conclusion and Future Work

Despite the good system results, we would like to increase our system's performance by adding features. For example, other research has indicated that the closer a publication's position is to the front cover of its journal, the more likely it is to be cited. Other rhetorical features such as the use of punctuation in the title, the use of tables and figures and other items have been seen to have a positive effect in predicting a publication's influence. Although we measured time to first citation, we did not include it as a feature. All of these features and more should be brought to bear on the next generation of models.

One shortcoming of the model we have built is that its predictions are static, and its errors are permanent. Van Dalen ([27]) indicates that many publications receive the bulk of their citations within the first couple of years after publication and exponentially fewer after some peak value. We have confirmed this in our data. For citations based on volume, our system may be useful in generating an initial hypothesis, and a separate fitness model may be useful in propagating that hypothesis over time. Another shortcoming is our measure of time of last citation. Since we stopped collecting data at a fixed date for this work, some of the features and labels for the publications have undoubtedly changed – time to last citation, for example – and these could have biased our results.

Most importantly, we believe that influence is not a homogeneous phenomenon; it is composed at least of the three measures of influence as we have defined them. It is possible that there are more. Sleeping beauties, for example, may be substantially different in character from publications which are cited continually over time. If so, these heterogeneous phenomena should be discussed and modeled separately.

Likewise, we have treated all citations as equal, but Catalini et al. [2] show that there is value in treating negative citations differently since these affect the relative influence of the publication. Self-citations, sometimes discounted by researchers in the bibliometric communities, may also be treated differently.

References

1. Kurt D Bollacker, Steve Lawrence, and C Lee Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. Proceedings of the second international conference on Autonomous agents, pages 116–123 (1998).
2. Catalini, Christian, Nicola Lacetera, and Alexander Oettl. "The incidence and role of negative citations in science." Proceedings of the National Academy of Sciences 112.45 (2015): 13823-13826.
3. Derek J. de Solla Price. Networks of Scientific Papers. Science (149:3683), pages 510–515 (1965).
4. Leo Egghe. Theory and practise of the g-index. Scientometrics (69:1), pages 131–152 (2006).

5. Leo Egghe. The Hirsch index and related impact measures. Annual review of information science and technology (44:1), pages 65–114 (2010).
6. Lawrence D. FU, and Constantin F. Aliferis. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics 85.1 (2010): 257-270.
7. C Lee Giles, Kurt D Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. Proceedings of the third ACM conference on Digital libraries, pages 89–98 (1998).
8. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter (11:1), pages 10–18 (2009).
9. Nick Haslam, Lauren Ban, Leah Kaufmann, Stephen Loughnan, Kim Peters, Jennifer Whelan, and Sam Wilson. What makes an article influential? Predicting impact in social and personality psychology. Scientometrics 76, no. 1 (2008): 169-185.
10. Jorge E. Hirsch. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America (102:46), pages 16569–16572 (2005).
11. Jorge E. Hirsch. Does the h index have predictive power? Proceedings of the National Academy of Sciences 104, no. 49 (2007): 19193–19198.
12. Timothy A. Judge, Daniel M. Cable, Amy E. Colbert, and Sara L. Rynes. What causes a management article to be citedarticle, author, or journal?. Academy of Management Journal 50, no. 3 (2007): 491-506.
13. Michael Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. String Processing and Information Retrieval, pages 1–10 (2002).
14. Lotka, A. J. (1926). The frequency distribution of scientific productivity. Journal of Washington Academy Sciences.
15. Robert K. Merton. The Matthew effect in science. Science 159.3810 (1968): 56–63.
16. Panchanan Mitra. Hirsch-type indices for ranking institutions scientific research output. Current Science 91, no. 11 (2006): 1439.
17. Mark E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. Complex Networks, pages 337–370 (2004).
18. Mark E. J. Newman. The first-mover advantage in scientific publication. EPL (Europhysics Letters) (86:6), (2009).
19. MEJ Newman. Prediction of highly cited papers. EPL (Europhysics Letters) (105:2), (2014).
20. Margaret W. Rossiter. The Matthew Matilda effect in science. Social studies of science 23.2 (1993): 325–341.
21. Andrs Schubert, Andrs Korn, and Andrs Telcs. Hirsch-type indices for characterizing networks. Scientometrics 78, no. 2 (2008): 375–382.
22. Christian D Schunn, Kevin Crowley and Takeshi Okada. The growth of multidisciplinarity in the Cognitive Science Society. Cognitive Science (22:1), pages 107–130 (1998).
23. Irving H. Sher and Eugene Garfield. New tools for improving and evaluating the effectiveness of research. In Research program effectiveness, proceedings of the conference sponsored by the Office of Naval Research, Washington, DC, pp. 135–146. 1965.
24. Xiaolin Shi, Belle Tseng and Lada A Adamic. Information diffusion in computer science citation networks, arXiv preprint arXiv:0905.2636 (2009).
25. Steinbach, Michael, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. KDD workshop on text mining. Vol. 400. No. 1. 2000.
26. Teja Tscharntke, Michael E. Hochberg, Tatyana A. Rand, Vincent H. Resh, and Jochen Krauss. Author sequence and credit for contributions in multiauthored publications. PLoS Biol 5, no. 1 (2007): e18.
27. Hendrik P. Van Dalen and Kène Henkens. What makes a scientific article influential? The case of demographers. Scientometrics (50:3) pages 455–482 (2001).
28. Van Raan, Anthony FJ. Sleeping beauties in science. Scientometrics 59.3 (2004): 467-472.
29. Dashun Wang, Chaoming Song, and Albert-Lszl Barabsi. Quantifying long-term scientific impact. Science 342, no. 6154 (2013): 127–132.

Table 2 Model Features.

Category	Feature	Description
Author	Coauthors	Number of authors
	max H-index	Largest H-index for publication's authors
	min H-index	Smallest H-index for publication's authors
Community	Community ID	Unique ID for the community
	Impact Factor	Number of citations to all publications in the community divided by number of publications in that community in the two years prior to date in question
Longevity	Max Longevity	Largest longevity of the publication's authors
	Min Longevity	Smallest longevity of the publication's authors
	Max Community Long.	Largest community longevity of the publication's authors
	Min Community Long.	Smallest community longevity of the publication's authors
Loyalty	Max Loyalty	Largest loyalty (cf. Section 5.1) for publication's authors
	Min Loyalty	Smallest loyalty (cf. Section 5.1) for publication's authors
Publication	Publication ID	Unique ID for the publication
	Year Published	Year in which the publication appears in community
	References	Number of references made from the publication
Title and Abstract	Title Word Count	Number of words in title
	Abstract Word Count	Number of words in abstract
	Abstract/Title Ratio	Number of words in abstract divided by number of words in title
	Specific words in abstract	census, facebook, google, html, http, investigate, investigation, overview, probe, review, social, study, survey, twitter, web
	Specific words in title	census, facebook, google, html, http, investigate, investigation, overview, probe, review, social, study, survey, twitter, web

A Features

Table 2 lists all 48 features used in our system. We consider different functionals (example, min or max of a set of numbers) to be different features.

B Clustering Performance

Table 3 shows the different aliases for the Quantitative Evaluation of Systems (QEST) conference. We chose this conference because of its relatively small number of entries but its relatively high number of aliases. ID numbers uniquely identify an alias within our database. Lines separate clusters of aliases. Note that there is one large cluster of 15 aliases and many clusters with a single alias.

Table 3 Aliases for the QEST community.

ID	Alias
2216	QEST
5025	In International Conference on the Quantitative Evaluation of Systems (QEST). IEEE Computer Society
10938	Proceedings of the 2nd International Conference on the Quantitative Evaluation of Systems (QEST
12326	In Proc. 1st QEST
15105	In Quantitative Evaluation of Systems - (QEST?06
17352	In QEST 2006 (3rd International Conference on the Quantitative Evaluation of SysTems
6784	In International Conference on Quantitative Evaluation of Systems (QEST
38546	In To appear in the Proceedings of QEST?07
39720	of Systems, First International Conference on (QEST?04), 00:304?313
44534	In: Proc. of the 1st Int. Conf. on Quantitative Evaluation of Systems (QEST) 2004. Twente, The Netherlands
45724	in QEST?05, Proc. 2nd Intl. Conf. on the Quantitative Evaluation of Systems
48025	In Proc. 1st International Conference on Quantitative Evaluation of Systems (QEST?04
52576	In Proc. 1st International Conference on the Quantitative Evaluation of Systems (QEST
55762	in Proc. 2nd International Conference on Quantitative Evaluation of Systems (QEST
182532	in 2nd International Conference on the Quantitative Evaluation of Systems (QEST
91245	In Proceedings of the 3rd International Conference on Quantitative Evaluation of Systems (QEST 2006). IEEE
91246	in Proc. of the 1st Int. Conference on Quantitative Evaluation of Systems (QEST 2004), IEEE-CS
94902	In Proceedings of QEST
100828	In Proc. of QEST: Quantitative Evaluation of Systems
100831	In QEST
105179	In QEST ?04: Proceedings of the The Quantitative Evaluation of Systems, First International Conference on (QEST?04
147598	INTERNATIONAL CONFERENCE ON QUANTITATIVE EVALUATION OF SYSTEMS (QEST
147989	in: Proceedings of the First International Conference on Quantitative Evaluation of SysTems (QEST-2004
173560	In Proceedings of QEST. IEEE Computer Society
180732	in Proceedings of the 2nd International Conference on the Quantitative Evaluation of Systems (QEST

Because none of the clusters have aliases belonging to other conferences, the purity of each cluster and of the set of clusters is 1.0. The entropy of this set of clusters is 0.0269, slightly higher than that of the other communities we sampled (0.0231).