Towards a Stratified Learning Approach to Predict Future Citation Counts

Tanmoy Chakraborty[1], Suhansanu Kumar[2], Pawan Goyal[3],
Niloy Ganguly[4], Animesh Mukherjee[5]
Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur, India – 721302
{[1]its_tanmoy, [2]suhansanu.kumar, [3]pawang, [4]niloy, [5]animeshm}@cse.iitkgp.ernet.in

ABSTRACT

In this paper, we study the problem of predicting *future citation count* of a scientific article after a given time interval of its publication. To this end, we gather and conduct an exhaustive analysis on a dataset of more than 1.5 million scientific papers of computer science domain. On analysis of the dataset, we notice that the citation count of the articles over the years follows a diverse set of patterns; on closer inspection we identify six broad categories of citation patterns. This important observation motivates us to adopt *stratified learning* approach in the prediction task, whereby, we propose a *two-stage prediction model* – in the first stage, the model maps a query paper into one of the six categories, and then in the second stage a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper. Experimental results show that the categorization of this huge dataset during the training phase leads to a remarkable improvement (around 50%) in comparison to the well-known baseline system.

1. INTRODUCTION

A common consensus in the scientific community is that all published articles have a similar itinerary of receiving citations – an initial growth in the number of citations within the first two to three years after publication followed by a steady peak of one to two years and then a final decline over the rest of the lifetime of the article [6, 8]. In most cases, the above observation has been drawn from the analysis of a very limited set of publication data [3, 13], thus, obfuscating the true characteristics. In this paper, we identify for the first time, at least six different such itineraries through a rigorous analysis of a massive data of 1.5 million papers from the computer science domain. Examples of itineraries include an early peak of citation in the first half (within first five years after the publication) of the itinerary followed by an exponential decay, multiple peaks at different time points, monotonic increase of the number of citations etc. (see Figure 2 for the detailed categorization).

The leading objective of this paper is to show that the above finding has significant consequences to early prediction of citation itinerary of scientific papers. Such a prediction scheme can be of significant interest not only for the scholars at universities and research institutes but also for the engineers and policy makers in business and government domains. The very limited number of studies on this topic [23, 24] have mostly modeled the problem as a learning task – given a set of features and a particular time interval, a regression model is trained on the entire set of the training population, and accordingly, the future citation count of a query paper is estimated. A common underlying implicit assumption in these approaches is that the citation itinerary of all published papers have similar characteristics. However, we observe that such an assumption is flawed and therefore seriously affects the accuracy of the prediction. Consequently, we propose to categorize the complete set of data samples into different subparts each of which corresponds to one type of citation itinerary observed. This approach is commonly termed as *stratified learning* [10] in the literature where the members of the stratified space are divided into homogeneous subgroups (aka strata) before sampling. This indeed reduces the extent of variability and increases the representativeness of the data samples in each individual strata thus enhancing the learning scheme [22].

The massive dataset that we use for this work contains papers with complete bibliographic information such as the title, author(s), affiliation of author(s), year of publication, references, abstract, keywords and the field information (subareas within the computer science domain). The identification of the different categories of citation itineraries through an exhaustive and systematic analysis of the data constitutes the key observation in this paper and allows us to make a bunch of contributions, some of which are described as follows. It is important to note that all these contributions follow quite naturally from the key observation, however, having a remarkable impact on the overall accuracy of the prediction.

The major contributions of our paper are manifold:

- We formulate a set of heuristic rules to automatically classify the citation itineraries into the six categories (or strata) (see Section 4).
- We develop a *two-stage prediction model* – in the first stage, a query paper is mapped into one of the strata using a Support Vector Machine (SVM) approach that learns from a bunch of features related to the author,

978-1-4799-5569-5/14/$31.00

the venue of publication and the content of the paper; in the second stage, only those papers corresponding to the strata of the query paper are used to train a Support Vector Regression (SVR) module to predict the future citation count of the query paper. For the same set of features available at the time of publication, the two-stage prediction model remarkably outperforms (to the extent of 50% overall improvement) the well-known baseline model (see Sections 5 and 6).

- Our two-stage prediction model produces significantly better accuracy in predicting the future citation count of the highly-cited papers that might serve as an useful tool in early prediction of the seminal papers that are going to be popular in the near future (see Section 7).
- We conduct an extensive analysis of the features revealing that those related to the author and the venue of the publication are very crucial for the purpose of prediction. However, the features related to the content of the paper are more effective in long term citation prediction (see Section 7).
- Finally, we show that including the first few years of citations of the paper into the feature set can significantly improve the prediction accuracy especially in the long term (see Section 7).

2. RELATED WORK

Several works have been conducted to analyze citation behavior of authors, papers and venues [1, 18]. Sun et al. [20] have investigated impacts of author, venue and content features for clustering in different heterogeneous networks. Early work on citation count prediction focused on a limited set of features and applied simple models such as linear regression and decision trees on relatively small datasets. For example, Callaham et al. [3] study 204 publications and using decision trees they report that the journal's impact factor is the most effective feature for predicting the citation counts 3.5 years after publication. Kulkarni et al. [13] use linear regression to study 328 articles published in 2000 and report R^2 (see the formula in Section 7) of 0.2 (which is pretty ordinary) in predicting citation counts 5 years ahead. Castillo et al. [4] use linear regression and decision tree to predict citation counts 4.5 years ahead. They observe that future citation counts are highly correlated with the citation counts accumulated within the first year after publication and that by adding features describing author's reputation they are able to improve their predictions. A recent work [12] considers the usage of content to improve prediction, i.e., by identifying keywords in the text that are associated with high citations. Didegah and Thelwall [7] study several features and find venue prestige to be the strongest feature, followed by the number of citations attracted by the references of a paper. Teufel et al. [21] propose a scheme to classify citations based on their usage in different contexts. Recently, Yan et al. conduct two similar experiments [23, 24], to study features covering venue prestige, content novelty and diversity and authors' influence and activity. They also account for the temporal dynamics by taking a recent version of each feature calculated on a limited time window. To the best of our knowledge, this is the latest and the most accurate future citation count prediction model, and therefore serves as the baseline system in this paper. We conduct an extended examination of all these factors related to citation counts, with many new features added. Unlike most of the previous studies, this paper is different in two aspects – (i) we adopt a stratified sampling approach to categorize training population that indeed proves to be very effective in the final prediction, (ii) we identify an extensive set of features which can impact citation and meticulously collect information about those features. We find that with the help of these features we can predict with high accuracy the future citation pattern of a paper at the time of its publication; the citation prediction accuracy becomes even better with the inclusion of the first year's citation count.

Table 1: Percentage of papers in various fields of computer science domain.

Fields	% of papers	Fields	% of papers
AI	12.64	Algorithm	9.89
Networking	9.41	Databases	5.18
Distributed Systems	4.66	Comp. Architecture	6.31
Software Engg.	6.26	Machine Learning	5.00
Scientific Computing	5.73	Bioinformatics	2.02
HCI	2.88	Multimedia	3.27
Graphics	2.20	Computer Vision	2.59
Data Mining	2.47	Programming Language	2.64
Security	2.25	Information Retrieval	1.96
NLP	5.91	World Wide Web	1.34
Education	1.45	Operating Systems	0.90
Embedded Systems	1.98	Simulation	1.04

3. THE PUBLICATION DATASET

We have crawled one of the largest publicly available datasets from Microsoft Academic Search (MAS)[1] which houses over 4.1 million publications and 2.7 million authors with updates added every week. We collected all the papers specifically published in the computer science domain and indexed by MAS[2]. The crawled dataset contains more than 2 million distinct papers altogether which are further distributed over 24 fields of computer science domain (see Table 1). Moreover, each paper comes along with various bibliographic information – the title of the paper, a unique index for the paper, its author(s), the affiliation of the author(s), the year of publication, the publication venue, the related field(s) of the paper, the abstract and the keywords of the papers. Each paper is also annotated by MAS with a set of keywords to characterize the paper. The dataset is available at `http://cnerg.org` for the research community to facilitate further investigations.

Table 2: General information of raw and filtered datasets.

	Raw	Filtered
Number of valid entries	2,473,171	1,549,317
Number of entries with no venue	343,090	–
Number of entries with no author	45,551	–
Number of entries with no publication year	191,864	–
Partial data of the years before 1970 and 2011-2012	343,349	–
Number of authors	1,186,412	821,633
Avg. number of papers per author	5.18	5.04
Avg. number of authors per paper	2.49	2.67
Number of unique venues	6,143	5,938
Percentage of entries with multiple fields	9.08%	8.68%

In order to remove the anomalies that crept in due to crawling, the dataset was passed through an initial preprocessing stage where we filtered out all such papers that did not have the bibliographic attributes required for our study such as the unique index of the paper, the year of publication, the list of authors, the publication venue and the

[1] `academic.research.microsoft.com`

[2] The crawling process was completed in August, 2013.

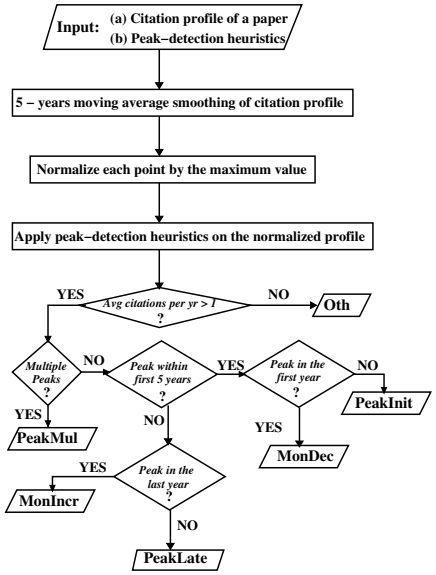

Figure 1: A systematic flowchart demonstrating the rules for classifying the training samples.

abstract and keywords of the papers. Further, we consider only those papers published in between 1970 and 2010 because this set of papers contains most reliable and significant entries. In the filtered dataset, 8.68% papers belong to multiple fields (act as interdisciplinary papers). Apart from the metadata information of all the papers, another advantage of using this dataset is that the ambiguity of named-entities (authors and publication venues) has been completely resolved by MAS, and a unique identity is associated with each author, paper and publication venue. Some of the general information pertaining to the filtered dataset is presented in Table 2.

4. CITATION ITINENARIES OF RESEARCH PAPERS

Since the primary focus of our study is to predict future citation count of a paper, an indepth understanding of how the number of citations after publication of a paper varies over the years is necessary. We therefore conduct an exhaustive analysis of the citation patterns of different papers present in our dataset. Some of the previous experimental results [6, 8] show that the trend of citations received by a paper after its publication date is not linear in general; rather there is a fast growth of citations within the initial few years, followed by an exponential decay. Here for an extensive analysis, we first take all the papers having at least 10 years of citation history and to avoid aging factor in citation analysis [9], we only consider maximum 20 years of their citation history.

In order to decipher the trends of citation, we perform various processing on the data set. First of all, to smoothen the time series data points in the citation profile of a paper, we use five-years moving average filtering; then, we scale the data points by normalizing them with the maximum value present in the time series (i.e, maximum citations received by the paper in a particular year); finally, we run local peak detection algorithm[3] to detect peaks in the citation profile. Over and above, we apply the following two heuristics to specify peaks: (i) the height of a peak should be at least 75% of the maximum peak-height, and (ii) two consecutive peaks should be separated by more than 2 years, otherwise they are treated as a single peak. A systematic flowchart to detect each category is shown in Figure 1.

Surprisingly, we notice that a major proportion of papers do not follow the traditional citation pattern mentioned in the earlier studies; rather there exist six different patterns of citation profiles of research papers based on the count and position of peaks in the citation profile. The six types of citation profiles are described below:

(i) PeakInit: Papers whose citation count peaks within 5 years of publication (but not in the first year) followed by an exponential decay (Figure 2(a)).

(ii) PeakMul: Papers having multiple peaks in different time periods of the citation itinerary (Figure 2(b)).

(iii) PeakLate: Papers having very few citations at the beginning and then a single peak after at least 5 years of the publication which is followed by an exponential decay in citation count (Figure 2(c)).

(iv) MonDec: Papers whose citation count peaks in the immediate next year of the publication followed by a monotonic decrease in the number of citations (Figure 2(d)).

(v) MonIncr: Papers having a monotonic increase in the number of citations from the very beginning of the year of publication till the date of observation (i.e., it can be after 20 years of its publication) (Figure 2(e)).

(vi) Oth: Apart from the above types, there exist a large number of papers which on an average have received less than one citation each year. For these papers, the evidences are not significant enough for assigning them into one of the above categories, and, therefore, they remain as a separate group altogether.

5. DISTINCTIVE FEATURES

In this Section, we provide a brief description of the set of features learned by the classifiers. The features can be broadly classified into three classes, namely the author-centric features, the venue-centric features and the paper-centric features. Note that for a particular paper, all the features are calculated with respect to the year of its publication. For feature values which are still unobserved, e.g., new authors or new venues, we do not assign zero values; instead we set them to the minimum value observed across all the samples available at that particular time point.

5.1 Author-centric Features

For all the author-centric features mentioned here, we measure both the average (**Avg**) and the maximum (**Max**) values for each paper to incorporate the notion of both team-effect and individual leadership respectively in the final citation count prediction.

(a) Author productivity: Yan et al. [23] noticed that the more papers an author publishes (productivity of the author), the higher average citation counts she can expect. Therefore, for each paper, we calculate the productivity of

[3]The peak detection algorithm is available in Matlab Spectral Analysis package - http://www.mathworks.in/help/signal/ref/findpeaks.html; we use 'MINPEAKDISTANCE'=2 and 'MINPEAKHEIGHT'=0.75 and the default values for the other parameters.

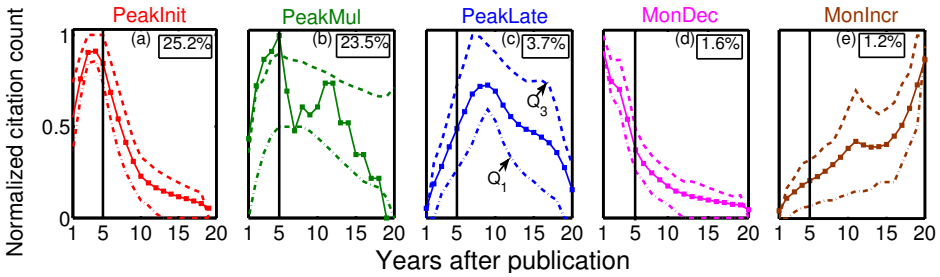

Figure 2: (Color online) Citation itineraries for the first five categories. In each frame, the belt bounded by the lines Q_1 and Q_3 represent the first and third quartiles of the data points respectively. For each category, one representative citation itinerary is shown at the middle of the belt. The percentage mentioned in each frame indicates the proportion of papers in each category. The major proportion of papers (44.8%) lies in category 'Oth' which does not have any specific pattern and is not shown in this diagram.

its authors (**ProAuth**) that indeed indicates how the influence of productive authors regulates the citation profile of a paper.

(b) Author h-index: H-index is a standard metric to measure both the productivity and the impact of the published work of an author [11]. Therefore for each paper, we measure the h-index (**Hindex**) of authors.

(c) Author diversity: The diversity of an author a denoted by $AuthDiv(a)$, indicating the breadth of expertise of a is measured by the entropy of the research fields where she publishes and is given by

$$AuthDiv(a) = -\sum_{i=1}^{24} p(n_i|n) \times log(p(n_i|n)) \qquad (1)$$

where n_i denotes the number of papers written by author a belonging to the field i (total 24 fields are available in the dataset), and n denotes the total number of papers written by a. For each paper, we include the diversity of authors (**AuthDiv**) as a feature. Note that, these two features have not been considered in earlier works [23, 3, 4], and they prove to be quite efficient in predicting future citation count (see Table 6).

(d) Sociality of author: Since the authors tend to cite papers of their previous collaborators [1], it is natural to assume that the paper from a widely connected author has a larger probability to be cited by her coauthors. A simple measurement is to count the number of coauthors (**NOCA**) of each author present in a paper [23].

5.2 Venue-centric Features

We consider the three features listed below to signify the importance of venue.

(a) Long term venue prestige: To measure the prestige of a publication venue (**VenPresL**), we calculate the average citations received by the papers published so far in that venue.

(b) Short term venue prestige: It is measured as the average number of citations received per paper published in that venue during *at most* two preceding years (**VenPresS**). The basic difference between **VenPresL** and **VenPresS** is that while the former one measures the overall impact of a venue by considering all the papers published so far in that venue, the latter one only measures the recent impact of the venue. **VenPresS** is similar to the impact factor of a journal as defined in [8].

(c) Venue diversity: **VenDiv** can be measured by considering the different fields covered by the papers published in that venue. A formula similar to Equation 1 gives a quantitative measure of **VenDiv**. This is another new feature introduced in this study for the first time.

5.3 Paper-centric Features

Among the paper-centric features mentioned below, third and fourth features are newly introduced in this study.

(a) Team-size: It has been observed that there exists a critical value of team-size corresponding to which the citation accumulation is maximum. Hence, we directly take into account the number of authors of a paper (**Team**).

(b) Reference count: Sometimes, only the number of references serves as a feature to distinguish regular and survey papers. Therefore, we directly use the reference count of a paper (**RefCount**) as a feature in our study.

(c) Reference diversity: Chakraborty et al. [5] propose a measure called Reference Diversity Index (**RDI**) as a measure of interdisciplinarity that attempts to quantify the diversity in terms of the number of fields being cited by a paper. It is also measured similarly using Equation 1; here n (n_i) indicates the total number of references (number of references to the papers belonging to field i).

(d) Keyword diversity: As mentioned in Section 3, MAS assigns keywords, from a global set of keywords, against each paper in order to characterize it properly. For each paper, we measure how diverse its keywords are (**KDI**) similarly by Equation 1; here n_i indicates the fraction of keywords of paper x belonging to the field i. Note that, a keyword may appear in multiple fields. For them, we consider multiple instances one for each field.

(e) Topic diversity: We use the unsupervised Latent Dirichlet Allocation[4] [2] as mentioned by Yan et al. [23] to discover topics for each paper. We empirically set the number of topics as 100, i.e., for each of our 100 topics, the topic model calculates $p(topic_i|d)$, the inferred probability of topic i in document d (**Topic**). The topic distribution $\tau(d)$ over all topics in the document d is then: $\tau(d) = \{p(topic_1|d), p(topic_2|d), ..., p(topic_{100}|d)\}$.

6. PROPOSED FRAMEWORK

The prediction is done through a two-stage learning process where the learning task is defined as follows:

[4] We use GibbsLDA++ (http://gibbslda.sourceforge.net/) with the default settings.

Learning task: Given a set of features $F = \{f_1, f_2, ..., f_n\}$, our goal is to learn a function ψ to predict the citation count of an article d at a give time period Δt after its publication. Formally, this can be written as:

$$\psi(d|F, \Delta t) \rightarrow C_T(d|\Delta t) \quad (2)$$

where *citation count*, C_T is as defined below.

Citation count: As defined by Yan et al. [23], given the set of scientific articles D, the citation count ($C_T(.)$) of an article $d \in D$ is defined as:

$$citing(d) = \{d' \in D : d' \ cites \ d\}$$
$$C_T(d) = |citing(d)|$$

Note that in this paper, we consider $\Delta t \in [1, 5]$.

We now elaborate the two-stage learning process undertaken to accomplish the above mentioned task.

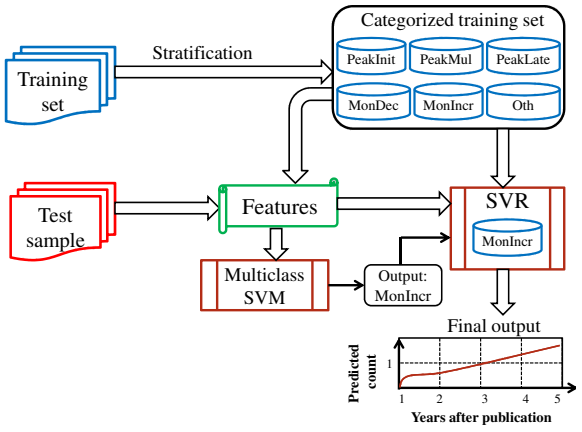

Figure 3: (Color online) A schematic of our proposed framework (SVM: Support Vector Machine, SVR: Support Vector Regression). Here we assume that the query paper is mapped to 'MonIncr' class by the SVM module.

6.1 Two-stage Prediction Model

The schematic diagram of our proposed two-stage model for predicting future citation count is shown in Figure 3. In the first stage, a sample (paper) is classified into one of the six identified categories which is done by using a multi-class SVM. In the second stage, the actual citation prediction of the paper is computed by employing a customized SVR model. In the rest of the Section, we explain each of the stages separately.

6.1.1 Support Vector Machine (SVM)

For each training sample, we identify its category among the six defined categories using the set of rules shown in Figure 1. We also extract the features (mentioned in Section 5) for each training sample. Hence the multi-class Support Vector Machine (SVM) [14] receives the category and the feature (author-centric, paper-centric, venue-centric) information of each member in the training set. Subsequently, given a test sample (query paper) along with its set of features, the multi-class SVM outputs the category of the sample. For training and classification phases of SVM, we use Weka-LibSVM toolkit[5] applying pairwise multi-class decision approach. The best results are obtained for the polynomial kernel setting. The overall accuracy and the importance of each feature in the classification task are reported in Section 7.

6.1.2 Support Vector Regression (SVR)

Support Vector Machine can be applied not only to classification problems but also to the case of regression often termed as *Support Vector Regression* (SVR) [19]. We use LibSVM (epsilon-SVR)[6] for this analysis with the default parameter settings. We train separate SVR model for each category C as well as for each time instance Δt; each SVR is identified by the notation $SVR(C, \Delta t)$. Recently, Yan et al. [23] used four prediction models, namely Linear Regression, k-Nearest Neighbor, CART and SVR and showed that SVR outperforms other models in predicting future citation counts. Therefore, in our experiment we use SVR for the final prediction.

The training set for SVR pertaining to a certain category (say, MonIncr) contains the papers whose citation patterns follow that category. Besides taking the features of the papers as input, $SVR(C, \Delta t)$ also takes as input the number of citations the constituent training papers in that category have received at Δt time after their publication. That is, if $\Delta t = 5$, the citation count of a paper at the fifth year of its publication available from the training sample is taken into consideration. For example, if a paper has been published in 1975 (1978), the number of citations it received in the year 1980 (1983) is taken as input.

Handling information leakage: In order to make predictions for the query paper, we always consider the information available before the publication of the query paper (i.e., we avoid any information *at or after* the publication year of the query paper). For instance, when predicting the future citation count of an article (published in 1996) 5 years after its year of publication, all the articles published in the year 1990 or before are processed in the training samples; all the other articles published after 1990 are discarded. The reason is that for 5-years future citation prediction of the papers published in 1996, if we use the papers published in 1992 in the training phase, their citation counts in the year 1997 would become the data points in the training space of the regression model for $\Delta t = 5$. This implies that we are using the information of the citations at 1997 in order to predict the citation count of the paper published in the year 1996, which leads to information leakage.

7. PERFORMANCE EVALUATION

In this Section, we analyze the performance of the baseline system and our proposed model in predicting future citation count of a given paper at the time of publication. For the baseline system, we design a model which is similar to that proposed by Yan et al. [23] (except that we are using a lot more features). It is identical to our proposed model except that it does not include the first stage of our model. Thus, for a query paper, it takes into account all the training samples and the set of features discussed in Section 5, and applies SVR to predict the citation count of the paper.

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Essentially, we intend to show the significance of the preprocessing stage (first stage of our model) in the task of future citation prediction.

Evaluation Metrics: For the evaluation purpose, we use the following metrics: coefficient of determination (R^2)[7] [23], mean squared error (θ) [15] and Pearson correlation coefficient (ρ) [17]. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true.

7.1 Dataset

The filtered dataset contains 1,549,317 scientific articles which need to be divided further for training and testing. However, for the evaluation of SVM, we need those papers whose true categorizations are known to us, i.e., those papers which have at least 10 years history (published between 1970-2000); though for measuring SVR accuracy, this might not be the criteria. Therefore for the sake of uniformity, we consider the papers published between 1970-1995 for training (505,149 papers), and the papers published between 1996-2000 (146,620 papers) for testing (for baseline as well as our algorithm) throughout the evaluation (unless explicitly mentioned). However, we also report the final prediction accuracy for the papers published between 2001-2005.

7.2 Performance of the Baseline Model

The predictive performances of the baseline system for each of the consecutive five years after publication are shown in Table 3 (columns 2-4). We observe that the baseline system achieves the highest accuracy (R^2=0.55, θ=5.45 and ρ=0.59) at the immediate next year after publication of a paper. We also observe that the accuracy of the predicted citation count is moderately overestimated for longer number of years which in turn decreases the accuracy of the baseline system in the later time periods.

Table 3: Performance of the baseline model (columns 2-4) and our proposed system at various time intervals for the test papers published between 1996-2000 (columns 5-7) and test papers published between 2001-2005 (columns 8-10). All the fractional values obtained from the regression model are suitably converted into the nearest integer values since the citation count of a paper can not be a fractional value. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true.

	Baseline 1996-2000			Our model 1996-2000			Our model 2001-2005		
	R^2	θ	ρ	R^2	θ	ρ	R^2	θ	ρ
Δt=1	**0.55**	**5.45**	**0.59**	0.87	2.66	0.86	0.89	1.95	0.88
Δt=2	0.54	6.36	0.57	**0.90**	**1.46**	**0.88**	**0.91**	**1.20**	**0.90**
Δt=3	0.53	7.67	0.56	0.83	3.11	0.85	0.82	3.22	0.80
Δt=4	0.50	9.16	0.52	0.77	3.86	0.84	0.77	3.76	0.79
Δt=5	0.48	12.09	0.49	0.74	4.18	0.75	0.71	4.08	0.73

7.3 Performance of Our Model

Table 3 shows the final performance of our model in each time interval after the time of publication. In this table, apart from the citation prediction for the papers published between 1996-2000, we also show the accuracy for the papers published between 2001-2005 (in that case, the training set consists of papers published between 1970-2000, and papers published between 2001-2005 constitute the test samples). Contrary to the performance of the baseline model where the highest accuracy is achieved at the immediate next year after publication, we achieve the best performance of our model 2 years after the year of publication. We shall analyze the reason behind the highest accuracy at Δt=2 in the Section 7.7. Remarkably, we observe that for all the cases, our model achieves nearly 50% higher accuracy compared to the baseline system (especially for θ and R^2). Note that, the performance in 2001-2005 is also quite significant - even better than the previous regime as the system is getting trained with more data. On observation, we find that the typical situation where the system performs poorly is when a new venue gets introduced and quickly becomes popular; it takes certain number of years of learning for the system to predict accurately. Another important observation is that the predicted citation counts are almost always overestimated (not underestimated) for the later years. The reason behind this is not exactly clear but the result itself provides an opportunity to estimate a linear offset to predict more accurately. However, this issue is out of the scope of this paper.

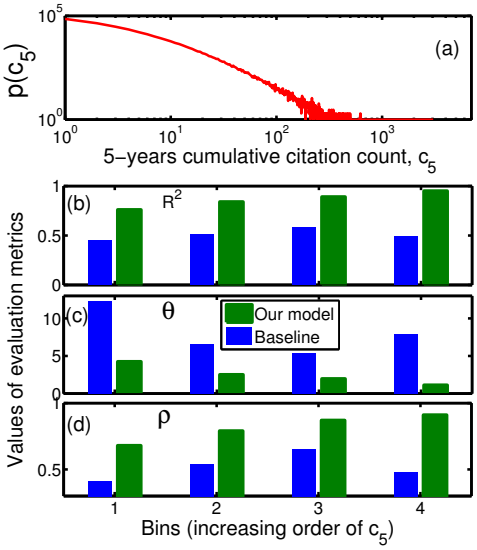

Figure 4: (Color online) (a) Distribution of 5-years cumulative citation count (c_5); values of (b) R^2, (c) θ and (d) ρ in four different buckets of c_5 for both baseline and our models. The range of c_5 in each bin is as follows: 1:0-2; 2:3-5; 3:6-10; 4:11-3045. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true.

7.4 Performance Evaluation Considering Different Citation Ranges

In order to compare the performances of these two models in different ranges of citation, we further look into the results obtained from the test set. First, we plot the distribution of cumulative 5-years citations (denoted by c_5) for test samples in Figure 4(a). Then we divide the entire range

[7] R^2 is defined as: $R^2 = 1 - \frac{\sum_{d\in D}(C(d)-C^{'}(d))^2}{\sum_{d\in D}(C(d)-C(D))^2}$, where D is the set of test documents, $C(d)$ is the actual citation count for article d, $C^{'}(d)$ is the predicted citation count for article d in the test set D, $C(D) = \frac{1}{|D|}\sum_{d\in D} C(d)$ is the mean of the actual citation count for an article present in D. $R^2 \leqslant 1$, and a larger R^2 indicates a better performance.

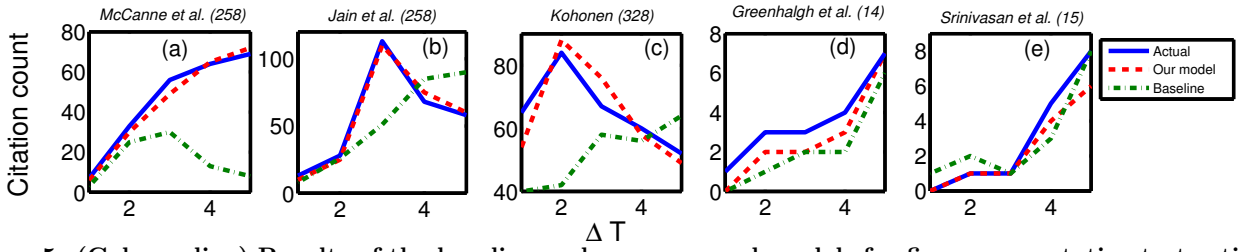

Figure 5: (Color online) Results of the baseline and our proposed models for five representative test articles. For each article, the cumulative citation count within the first five years after publication is reported in parenthesis in the title of each frame.

of c_5 into four bins such that all bins contain nearly equal number of papers, and for each individual bin we measure the values of the three evaluation metrics. Note that, here for each bin we measure the average value of each evaluation metric over five different values of Δt. It is apparent from Figures 4(b)-4(d) that the performance of our system increases with the increase of c_5; whereas the performance of the baseline model seems to be the best in the middle range of c_5. Therefore, we believe that our model might serve as an useful tool in early prediction of the important papers that are going to be popular in the near future. We take some example papers and individually compare their original and predicted citation counts.

Further analysis of papers in different citation ranges: Figure 5 shows the outputs of the baseline system and our proposed model for five representative scientific articles (with their cumulative citation counts within first 5 years after publication). Note that, the representative articles are chosen to represent various ranges of cumulative citation counts. The idea is to illustrate that our technique outperforms the baseline for high ($\geq$ 300) as well as medium ($<$300 and $\geq$ 100) and low ($<$ 100) citation count ranges. Further, in each range the articles for which the outcomes of our model are significantly different from the baseline have been a more preferable choice. The representative articles are as follows:

- **McCanne et al.**, Receiver-driven layered multicast, CCR, 24:4, pp. 117-130, 1996.
- **Jain et al.**, Statistical pattern recognition: A review, PAMI, 22:1, pp. 4-37, 2000.
- **T. Kohonen**, The self-organizing map, IJON, 21:3, pp. 1-6, 1998.
- **Greenhalgh et al.**, A QoS architecture for collaborative virtual environments, ACM - MM, pp. 121-130, 1999.
- **Srinivasan et al.**, An assessment of submissions made to the predictive toxicology evaluation challenge, IJCAI, pp. 270-275, 1999.

One can clearly notice an almost perfect alignment of the future citation count predicted by our model with the actual citation count in comparison to that for the baseline system for different values of Δt. Moreover, we observe that in many cases the baseline system even fails to estimate the basic citation pattern which is yet to manifest for a particular paper, hence making costly mistakes (see Figures 5(a) - (c)).

7.5 Performance of SVM Classification

We have discussed the accuracy of the prediction model but this in turn depends on the underlying first stage of classification which is done using multi-class SVM. Table 4 shows the confusion matrix describing the performance of the SVM classification model used in the first stage of our model. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Therefore, all correct guesses are located in the diagonal of the table. Bethard and Jurafsky [1] mentioned that 90% of papers that have been published in academic journals are never cited. We have also observed in Figure 2 that our dataset is highly biased towards the population of the low-cited papers (i.e, 'Oth'). Therefore, SVM also slightly overestimates 'Oth' category in the classification. The overall accuracy of the classification system is 0.78 which is quite promising considering the biased training samples and the fact that no feature after the publication of the paper is considered to classify the papers. Besides 'Oth' category, we also observe higher accuracy for class 'MonIncr' (0.87) which is followed by 'PeakInit' (0.79), 'PeakMul' (0.73) and 'PeakLate' (0.73). The lowest accuracy is obtained for category 'MonDec' (0.61). One possible reason could be that this is one of the rarest categories in the dataset. Thus, the lack of enough evidences might have accounted for the low final accuracy of the SVM model in classifying the papers into this category.

Table 4: Confusion matrix depicting the performance of SVM at the first stage of our prediction model. The last column indicates the accuracy of the classification system for each individual category. The correct classification results (diagonal elements) are highlighted in bold font.

	PeakInit	PeakMul	PeakLate	MonDec	MonIncr	Oth	Accuracy
PeakInit	**9550**	70	20	20	0	2419	0.79
PeakMul	29	**15261**	2500	3	0	3000	0.73
PeakLate	7	718	**4842**	2	489	518	0.73
MonDec	398	444	157	**2247**	0	453	0.61
MonIncr	2	403	0	0	**2789**	0	0.87
Oth	55	5142	5	2	0	**154188**	0.96
Overall accuracy							0.78

7.6 Performance Assuming Perfectly Accurate SVM

In Table 4, we notice that in the first stage of our model, we achieve overall 78% accuracy and the error in this stage propagates in the second stage of our model. We believe that this performance can be improved a lot in future with more efficient feature selection and thus remains a potential area of future research. However, one might argue that if the SVM model could have achieved nearly 100% accuracy,

how much improvement one would expect from the final prediction model. This might also answer how the error which propagates from the first stage of the model to the next stage affects the final output of citation prediction. Since we know the true category of each of the test papers, we use only those training samples belonging to the true category for training SVR, thus forcing 100% accuracy in the first stage. Table 5 shows the performance improvements (differences) of our model in comparison to the earlier results shown in Table 3 for different values of Δt (test set constitutes papers published within 1996-2000). One can clearly notice a significant improvement over the baseline model and our earlier results especially for the higher values of Δt. This indicates that the error propagating from the first stage SVM model to the next stage significantly affects long term citation prediction, and improvements in the first stage can highly enhance the overall performance of the system.

Table 5: Performance improvement (differences) of our model in comparison to the earlier results shown in Table 3 for different values of Δt, while considring 100% accuracy in SVM model. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true. Therefore, more negative value in terms of θ indicates better accuracy.

	Improvement over baseline model			Improvement over our earlier results		
	R^2	θ	ρ	R^2	θ	ρ
$\Delta t = 1$	0.34	-3.54	0.31	0.02	-0.75	0.04
$\Delta t = 2$	0.37	-5.09	0.34	0.01	-0.19	0.03
$\Delta t = 3$	0.37	-5.85	0.33	0.07	-1.26	0.04
$\Delta t = 4$	0.35	-7.22	0.36	0.08	-1.92	0.04
$\Delta t = 5$	**0.41**	**-10.19**	**0.37**	**0.15**	**-2.28**	**0.11**

7.7 Feature Analysis

Here we systematically analyze the impact of different feature groups described in Section 5 for SVM classification system as well as the actual citation count prediction. For SVM classification, we drop each feature in isolation and measure the overall performance of the model. The third column of Table 6 (column 3) shows the decrease in overall performance of SVM after dropping each feature in comparison to the case where all the features are present. Author-centric features seem to be the most effective features in the classification model. Among them, the absence of the average productivity of the authors in a paper (AvgProAuth) leads to maximum decrease in performance which is followed by the maximum diversity of the authors in a paper.

For measuring the impact of each feature in future citation count prediction, we use the standard approach adopted by McNamara et al. [16] – Spearman's rank correlation coefficient, an established measure of the dependence of two variables using a monotonic function, is taken for each of the features and the target variable (actual citation count) for each Δt.

Feature analysis in baseline model: Table 6 (columns 4-8) shows the impact of each individual features in the baseline model. We observe that once again the average productivity of authors in a paper turns out to be the best feature for all the time intervals. It is quite understandable since authors are likely to cite papers written by reputed and influential authors. Venue impact is also significant for the first few years. Papers from prestigious venues are likely to be highly cited. Interestingly, most of the paper-centric features which seem to have least significance in the initial few years, appear to be quite effective in the later time periods. This essentially indicates that a good quality paper eventually gets appreciations from the researchers irrespective of the reputation of the authors and the publication venue. However, it might take some time to get noticed by the others.

Observations about strong and weak features: In Section 7.3, we have noticed that our model seems to perform well at Δt=2. We hypothesize that this might happen because of the early categorization in the training phase. In Table 6 (columns 4-8), we notice that though the most prominent features such as AvgProAuth, AvgAuthDiv, Team show higher correlation with the actual citation count at Δt=1, other weak features such as MaxProAuth, AvgHindex, VenPres, RDI and KDI tend to attain maximum correlation at Δt=2. We believe that due to the categorization, these weak features tend to become prominent in the final prediction, resulting in highest accuracy at Δt=2. To strengthen this hypothesis, we again measure Spearman's rank correlation for all the features in our model.

Feature analysis in our model: Table 6 (columns 9-13) reports the rank correlation of each feature with the actual citation count averaged over all the categories. Besides the improvement of the absolute value of the correlation, one can also notice a large overall improvement of the correlation for most of the features at Δt=2 which indeed strengthens our hypothesis. Interestingly, although the relative ordering of the features in terms of the average correlation for all values of Δt remains almost same without (baseline model) and with categorization (our model), the weak features tend to rise significantly after categorization with reasonably higher improvement in rank correlation and serve an important role in the final citation count prediction.

7.8 Robustness of Categories

Earlier results show that the systematic categorization of the training samples improves the performance of the prediction system in comparison to the baseline system. A pertinent question could be that how robust are these categories for the final prediction, i.e., if the (near-)similar/dissimilar categories are merged together, how does it affect the final output of the model. Note that in Figure 2, the categories 'PeakInit' and 'MonDec' ('PeakLate' and 'MonIncr') are nearly similar in terms of the number of peaks and whether the peak occurs in the first/last half of the citation profile; others are reasonably different. Now the question is that if we merge the near-similar categories together to reduce the total number of categories, how does it affect the final prediction. The extreme case would be the baseline system itself where all the categories are combined. Apart from this, we reconfigure the categorization in two different ways: **[Cat-1]** combining near-similar categories and keep others separate ([PeakInit + MonDec], [PeakLate + MonIncr], [PeakMul], [Oth]), **[Cat-2]** combining one pair of dissimilar categories ([PeakInit + PeakMul], [PeakLate], [MonDec], [MonIncr], [Oth]). In this case also, we use the default set of training and test samples as mentioned in Sec-

Table 6: Feature analysis in two different stages of our prediction model. SVM classification: the third column indicates the decrease in overall accuracy when dropping each feature in isolation in comparison to the case when all the features are present (i.e., 0.78). SVR model: each subsequent column from the columns 4-8 (columns 9-13) indicates the Spearman's rank correlation of each feature with the actual citation count without categorization - Baseline Model (with categorization - Our Model). For each feature-group, the highest value in each column is highlighted in bold font.

Features		Decrease in performance of SVM	Correlation with the actual citation count - Baseline Model					Correlation with the actual citation count - Our Model				
			Δt=1	Δt=2	Δt=3	Δt=4	Δt=5	Δt=1	Δt=2	Δt=3	Δt=4	Δt=5
Author-centric	AvgProAuth	**0.21**	**0.67**	**0.65**	**0.63**	**0.61**	**0.55**	**0.76**	**0.74**	**0.70**	**0.67**	**0.68**
	MaxProAuth	0.05	0.19	0.20	0.20	0.20	0.19	0.28	0.38	0.30	0.33	0.21
	AvgHindex	0.06	0.17	0.27	0.18	0.16	0.15	0.32	0.41	0.36	0.38	0.32
	MaxHindex	0.04	0.11	0.13	0.12	0.10	0.11	0.34	0.42	0.39	0.32	0.29
	AvgAuthDiv	0.12	0.56	0.55	0.46	0.43	0.43	0.71	0.70	0.69	0.61	0.56
	MaxAuthDiv	0.18	0.62	0.55	0.49	0.26	0.14	0.63	0.62	0.56	0.38	0.34
	AvgNOCA	0.10	0.33	0.21	0.17	0.19	0.18	0.43	0.52	0.42	0.38	0.21
	MaxNOCA	0.16	0.43	0.32	0.31	0.30	0.23	0.51	0.45	0.42	0.39	0.32
Venue-centric	VenPresL	0.12	0.47	**0.42**	**0.23**	**0.14**	0.01	0.57	**0.58**	**0.51**	**0.49**	**0.36**
	VenPresS	0.13	0.48	0.22	0.13	0.04	0.02	0.49	0.44	0.38	0.32	0.29
	VenDiv	**0.18**	**0.50**	0.37	0.20	0.13	**0.03**	**0.59**	0.48	0.43	0.36	0.32
Paper-centric	Team	**0.11**	**0.47**	0.36	0.22	0.18	0.11	0.37	**0.49**	**0.46**	0.41	0.37
	RefCount	0.01	0.18	0.20	0.21	0.29	0.31	0.38	0.43	0.41	0.37	0.32
	RDI	0.07	0.36	**0.37**	**0.36**	0.35	0.33	**0.40**	0.46	0.42	0.39	0.31
	KDI	0.03	0.17	0.19	0.13	0.12	0.12	0.27	0.36	0.29	0.22	0.21
	Topic	0.10	0.23	0.21	0.25	**0.38**	**0.40**	0.31	0.35	0.37	**0.41**	**0.45**

tion 7.1 and run the two-stage prediction model separately for two types of categorization.

Table 7 shows the final performance of the two-stage model for the two categorization schemes. One can easily notice two immediate consequences of these schemes – (i) combining two near-similar categories (as followed in Cat-1) does not make much effect on the final prediction in comparison to combining two different categories (as followed in Cat-2), since the decrease in accuracy from the actual results (shown in Table 3) is significantly less for Cat-1 than that for Cat-2; (ii) while combining two major categories in Cat-2, the accuracy of the final prediction decreases drastically from the actual results of Table 3, and it tends to be closer to the baseline system. The results for Cat-1 are still worse (although slightly) than the original six categories system. Hence, a natural question stays whether dividing the data into further categories would improve performance. We have tried different variations, all show more noise begin to enter in the SVM classification model thus net decreasing the performance. However a more systematic category study is an important future work.

Table 7: Performance of the two-stage prediction model for two different types of categorization schemes. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true.

Performance of two-stage prediction model						
	Cat-1			Cat-2		
	R^2	θ	ρ	R^2	θ	ρ
Δt=1	0.87	2.05	0.85	0.59	5.23	0.63
Δt=2	**0.88**	**1.94**	**0.88**	**0.61**	**4.67**	**0.68**
Δt=3	0.79	3.38	0.80	0.55	6.86	0.61
Δt=4	0.75	4.01	0.76	0.51	8.89	0.54
Δt=5	0.71	4.10	0.72	0.50	9.58	0.49

7.9 Impact of Early Citation Information

In earlier papers [4, 13], it has been shown that the citation count of a paper in the initial few years after publication plays an important role in predicting the future citation count of the paper. However, in our experiments, we have only considered those features of a paper that one can get at the time of its publication since our objective is to predict the future impact of a paper as early as possible. However, we also believe that the initial few years' citation counts can boost up the prediction of the final citation counts since these initial citations seem to be the early crowd-sourced feedback of the scientific community about the paper. Therefore, to see its impact in the final prediction, we conduct another set of experiments – we include the citation count of a paper in the immediate next year (Δt=1) of its publication as a feature and predict the citation count of each paper for Δt between 2 and 5 years. In this case, we use the default set of training and test samples as mentioned in Section 7.1 and run the baseline system and the two-stage prediction model. Table 8 shows the accuracy for both the baseline system and the two-stage prediction model. As compared to Table 3, we can see a clear improvement of the system mostly in the higher values of Δt. Moreover, this also improves the SVM classification where we achieve 84% overall accuracy. With this information, the baseline system also improves a lot as mentioned in [23].

Table 8: Performance of the baseline model and our proposed system at various time intervals after including the first year's citation count as another feature. Note that, the more the value of R^2 and ρ, the more the accuracy of the model; but for θ, the reverse argument is true.

	Baseline			Our model		
	R^2	θ	ρ	R^2	θ	ρ
Δt=2	**0.60**	**4.92**	**0.65**	**0.92**	**1.02**	**0.90**
Δt=3	0.59	5.06	0.64	0.85	2.56	0.82
Δt=4	0.58	5.44	0.62	0.83	3.16	0.81
Δt=5	0.54	6.56	0.56	0.81	3.88	0.79

8. CONCLUSION AND FUTURE WORK

In this paper, we presented a two-stage framework to predict the future citation count of a published article in different time intervals after publication. We observed that

the inclusion of a stratified learning approach in the traditional citation prediction model remarkably enhances the overall performance of the prediction model. More importantly, the differences in the relative importance of different features provide insight to the differences in dynamics in different time intervals. We introduced a bunch of features in this task that prove to be effective in predicting citation count. We observed that author-centric features are the most distinguishing ones; among these, average productivity of authors seem to make a paper attractive. We further showed that adding the citation counts accumulated within the first year after publication as a feature can improve the prediction accuracy. However, the performance of the first stage classification and the choice of the number of categories are two vital areas that need to be carefully tackled during the prediction. As a final comment, the superior prediction accuracy has become possible due to the availability of the massive bibliographic dataset which we painstakingly collected. We plan to make the dataset publicly available soon for future research.

Since the information of different research fields in computer science domain is also available in our dataset, we plan to extend this work by looking into these fields separately. We also plan to explore new features that can provide additional signals not captured by the features used in this study. We suspect that the content features seem to provide weak signals because of the coarse representation of the content in terms of topic modeling. A more sophisticated and systematic mining of meaningful features from the content is an immediate future task. We also plan to investigate whether similar techniques could be used to predict the scholarly impact of higher level entities (e.g., researchers and universities). Beyond such future tasks, we believe that the concept of categorization of citation profile which is introduced for the first time in this paper, can prove to be very effective and can be instrumented in the design of more accurate bibliometric measurement and ranking schemes.

9. REFERENCES

[1] S. Bethard and D. Jurafsky. Who should i cite: learning literature search models from citation behavior. In *CIKM*, pages 609–618, New York, NY, USA, 2010. ACM.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] M. Callaham, R. L. Wears, and E. Weber. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287(21):2847–2850, 2002.

[4] C. Castillo, D. Donato, and A. Gionis. Estimating the number of citations using author reputation. In *SPIRE*, volume 4726 of *LNCS*, pages 107–117, 2007.

[5] T. Chakraborty, S. Kumar, M. D. Reddy, S. Kumar, N. Ganguly, and A. Mukherjee. Automatic classification and analysis of interdisciplinary fields in computer sciences. In *SocialCom*, pages 180–187, 2013.

[6] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee. Computer science fields as ground-truth communities: Their impact, rise and fall. In *ASONAM*, pages 426–433, Canada, 2013.

[7] F. Didegah and M. Thelwall. Determinants of research citation impact in nanoscience and nanotechnology. *JASIST*, 64(5):1055–1064, 2013.

[8] E. Garfield. Impact factors, and why they won't go away. *Nature*, 411(6837), May 2001.

[9] Y. Gingras, V. Larivier, B. Macaluso, and J.-P. Robitaille. The Effects of Aging on Researchers' Publication and Citation Patterns. *PLoS ONE*, 3(12):e4048, 2008.

[10] G. Haro, G. Randall, and G. Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. In *NIPS*, pages 553–560. 2006.

[11] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

[12] A. Ibanez, P. Larranaga, and C. Bielza. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309, 2009.

[13] A. Kulkarni, B. J. W., and S. I. Kulkarni. Characteristics associated with citation rate of the medical literature. *PLoS ONE*, 2(5):e403, 05 2007.

[14] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *J. Am. Statist. Assoc.*, 99(465):67–82, 2004.

[15] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.

[16] D. McNamara, P. Wong, P. Christen, and K. S. Ng. Predicting high impact academic papers using citation network features. In *PAKDD Workshops*, pages 14–25, 2013.

[17] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.

[18] A. Siddharthan and S. Teufel. Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*, pages 316–323, 2007.

[19] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.

[20] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *SIGKDD*, pages 797–806, New York, USA, 2009.

[21] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *EMNLP*, pages 103–110, Stroudsburg, PA, USA, 2006.

[22] D. Wu, L. Lu, J. Bi, Y. Shinagawa, K. L. Boyer, A. Krishnan, and M. Salganicoff. Stratified learning of local anatomical context for lung nodules in ct images. In *CVPR*, pages 2791–2798, 2010.

[23] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *JCDL*, pages 51–60, New York, USA, 2012.

[24] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *CIKM*, pages 1247–1252, New York, USA, 2011.