The Role Of Citation Context In Predicting Long-Term Citation Profiles: An Experimental Study Based On A Massive Bibliographic Text Dataset

Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty,
Animesh Mukherjee, Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, WB, India
mayank.singh@cse.iitkgp.ernet.in,vikaspatidar859@gmail.com
{suhansanu.kumar,its_tanmoy,animeshm,pawang}@cse.iitkgp.ernet.in

ABSTRACT

The impact and significance of a scientific publication is measured mostly by the number of citations it accumulates over the years. Early prediction of the citation profile of research articles is a significant as well as challenging problem. In this paper, we argue that features gathered from the **citation contexts** of the research papers can be very relevant for citation prediction. Analyzing a massive dataset of nearly 1.5 million computer science articles and more than 26 million citation contexts, we show that average **countX** (number of times a paper is cited within the same article) and average **citeWords** (number of words within the citation context) discriminate between various citation ranges as well as citation categories. We use these features in a stratified learning framework for future citation prediction. Experimental results show that the proposed model significantly outperforms the existing citation prediction models by a margin of 8-10% on an average under various experimental settings. Specifically, the features derived from the citation context help in predicting long-term citation behavior.

Categories and Subject Descriptors

H.2.8 [**Database Application**]: Data mining

Keywords

Long-Term Impact; Citation Prediction; Citation Context; Stratified Learning

1. INTRODUCTION

Citation count of a publication is among the most commonly accepted metric by the research community for evaluating the impact and quality of a research article. Citation count refers to the number of citations received by an article within a specified time-period [2]. Highly-cited works remain as one of the most important criteria for various organization (e.g. companies, universities and governments) to identify the best talents, especially at their initial stages. An early estimate would help in identification of promising articles that could accelerate research and dissemination of new knowledge. This has motivated the interest in the field of future citation prediction [17, 19].

Prediction of future citation counts, however, is difficult because of the nature and dynamics of citations [8, 10]. The citation ranges for the papers published by the same authors or the same venues show a lot of variation. The same can be said about the field of the papers as well. A very recent study [7] has shown that all the scientific papers do not follow the same trajectory and found 6 different citation patterns.

The existing works have used various venue and author centric features, along with the citation information from the initial years for the task of citation prediction. In this paper, we argue that the features extracted from the citation contexts can be extremely helpful for the future prediction. Citation context refers to textual descriptions of a given scientific paper found in other papers in the document collection which cites it [1]. A citation context is, in principle, a set of sentences where a paper is referred to. The intuition behind using the citation context features comes from the hypothesis that citation contexts reflect the opinion of the scientific community about the particular work. We show that even using some very simplistic features extracted from the citation context can boost the performance of a citation prediction system significantly.

Towards this objective, we use a massive dataset consisting of more than 26 million citation contexts for nearly 1.5 million research papers in the computer science domain, crawled from Microsoft Academic Search (MAS)[1]. We extract two features from the citation contexts – average countX (number of times a paper is cited within the same article, averaged over all the citing papers) and average citeWords (number of words within the citation context, averaged over all the citing papers). We show that these features are quite discriminative and exhibit different trends not only for different citation ranges but also for the citation categories identified in [7]. We then append these features along with various other features in an earlier framework based on *stratified learning* [6]. Experimental results show that addition of these two features gives an R^2-correlation of 0.84, 0.81 and 0.78 towards predicting the citation count at 5, 7 and 9 years after publication, improving the prediction accuracy by 8-10% on an average over the nearest baseline. Specifically, these features help in predicting the long term citation behavior of the research papers. We would like to stress here that this study brings forth the tremendous potential of

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.
© 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806566.

[1] http://academic.research.microsoft.com/

the content of a scientific article in predicting future citation counts; the huge success of only two very simple content related features proposed here makes the authors believe that deeper analysis of the content can lead to further significant improvements in the related areas of research.

The rest of the paper has been organized as follows. We discuss the related previous works in section 2. Section 3 describes the citation context dataset used for this experimental study. The two citation context features utilized for our study have been described in section 4. The citation prediction model has been described in section 5. The experiments to evaluate our system under different settings have been reported in section 6 along with a detailed comparison and feature analysis. Finally, conclusions and future works have been presented in section 7.

2. RELATED WORK

In recent years, several researchers have investigated the problem of future citation count prediction [17, 19, 20, 21]. Most of the past works have proposed a set of features and used a supervised learning model to predict the citation count at a later time point. Many works use only the information available at the time of publication to predict future citation count, while other works also use information available from the initial years after publication. For instance, Fu and Aliferis [9] predict citation count with the information available at the time of publication. They incorporate features like number of authors, number of articles for the first author, number of citations for the first author, number of affiliations, the journal impact factor, title, abstract, MeSH terms[2] etc. Support Vector Machines (SVM) have been used to predict citation count after 10 years of publication. Similarly, Livne et al. [14] use features like authors, author institutions, venue and references to train a Support vector regression (SVR). They observe that venue and the references are the most significant features for citation count prediction.

Callaham et al. [16] use decision trees to predict citation counts of 204 publications from emergency medicine specialty meeting. They use features like impact factor of journal, research design, number of subjects, rated subjectively for scientific quality, newsworthiness etc. Kulkarni et al. [11] use linear regression and achieve an R^2 of 0.2 for the prediction of citation count for five year ahead window using 328 medical articles. They use features like journal name, month of publication, study design, clinical category of the article etc.

Brody et al. [3] use information after the publication to forecast citation count. Download data within the first 6 months after publication is used as a predictive feature. Similarly, Lokker et al. [15] use features related to the article and journal, like number of authors, pages, references etc. for the prediction task. Castillo et al. [5] use the number of citations, authors' reputation and the source of paper citations as the predictive features.

Liangyue et al. [13] propose a joint predictive model to forecast the long-term scientific impact problem, formulated as a regularized optimization problem. Their work addresses four key algorithmic challenges, including the scholarly feature design, the nonlinearity, the domain-heterogeneity and dynamics. Further, they propose a fast online update algorithm to adapt joint predictive model efficiently over time. They observe that citation history is a strong indicator of long-term impact and using additional contextual or content features brings little marginal benefits in terms of prediction performance. An analysis on 463,348 papers from Physical Review (PR) corpus suggests high heterogeneity in the citation histories [18]. They present three fundamental mechanisms that drive citation history, namely preferential attachment, aging and novelty, and importance of a discovery (fitness). Combining these mechanisms allows to collapse the citation histories of papers from different journals and disciplines into a single curve, indicating that all papers tend to follow a similar universal temporal pattern.

Yan et al. present two similar works on citation prediction problem [19, 20]. They have introduced features covering venue prestige, content novelty and diversity, and authors' influence and activity. Pobiedina and Ichise [17] introduce a new feature GERscore (Graph Evolution Rule score), based on frequent graph pattern mining techniques, for citation prediction. Yu et al. [21] propose a new data structure namely discriminative term buckets to capture both document similarity and potential citation relation. They also propose metapath based feature space to interpret structural information in citation prediction. Along with these novel ideas, they present an extensive analysis on differences between citation prediction problem and the related work, e.g., traditional link prediction solution.

One of the previous works [6] suggests that *stratified learning approach* leads to good prediction accuracy. They observe that there exist six different patterns of citation profiles of research papers based on the number and position of peaks in the citation profile. Further, a two-stage prediction model was proposed, which maps a query paper into one of the six categories in the first stage, 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. They achieve a superior performance just by using the features at the time of publication. Motivated by this study, we also use a stratified learning framework for citation prediction. However, since the prime objective of this work is to show the utility of citation context features which are available only after publication, we utilize the citation context features derived from the first two years after publication along with the publication time features to improve the prediction accuracy.

To the best of our knowledge, this is the first work that attempts to use citation context based features in the citation prediction problem. We use a massive dataset of more than 26 million citation contexts from computer science research articles towards this goal. The next section describes the dataset used in this study.

3. DATASET

In this paper, we use two computer science datasets, both crawled from Microsoft Academic Search (MAS)[3]. First dataset (bibliographic dataset) consists of bibliographic information of papers, 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, references, citation contexts, the related field(s)[4] of the paper, the abstract and the keywords of the papers [6]. Second dataset (citation context dataset) consists of more than 26 million citation contexts pre-processed and annotated with cited and citing paper information.

Table 1 details various statistics for both the datasets.

Definition I: We will say that paper P *cites* paper C, if paper P refers to paper C in the text. P is termed as **citing paper** while C is termed as **cited paper**. P can refer to C at many places in the text. In our present work, we only consider the sentence as citation context where the reference to the paper is explicitly present.

[2] http://www.nlm.nih.gov/mesh/

[3] http://academic.research.microsoft.com

[4] Note that the different sub-branches like Algorithms, AI, Operating Systems etc. constitute different "fields" of computer science domain.

For an example, Chakraborty et al. [6] cites Yan et al. [19] as

> *Recently, Yan et al. conduct two similar experiments [**25**, 26], 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.*

Although, the above context consists of four sentences, we only consider the first sentence as the citation context since it explicitly refers to Yan et al. [19].

Definition II: countX for a cited paper C with respect to a citing paper P is defined as the number of citation contexts, when a paper P cites paper C. Citation context count for a paper C denotes the sum of countX from all the citations for C.

Table 1: General information about the datasets

Dataset I	Year range	1960-2010
	Number of computer science fields	21
	Number of publications	1,359,338
	Number of authors	138,923
	Avg. number of papers per author	5.43
	Avg. number of authors per paper	2.40
Dataset II	Number of citation contexts	26,197,440
	Avg. number of citation contexts per paper	19.27
	Avg. number of words per citation context	26
	Number of papers having at least one citation context	1,279,104

Each paper has a specific citation context count. Figure 1 shows the distribution of papers having specific citation context count in our dataset. Long tail depicts that many papers have less number of citation context count while a small number of papers have high citation context count.

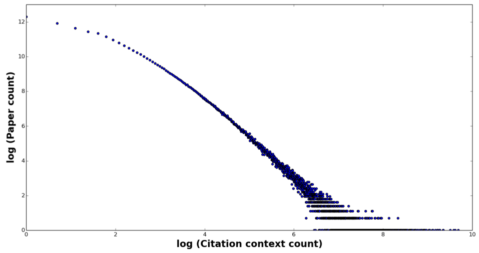

Figure 1: Distribution of citation context count in our dataset

Definition III: citeWords for a cited paper C with respect to a citing paper P is defined as the number of words in the citation context, when a paper P cites paper C. If multiple papers are cited within the same citation context, the number of words are equally divided among all the cited papers. If a paper is cited multiple times within the same paper, citeWords is computed by summing over the words in all the citation contexts.

In the next section, we discuss in detail how the average values of countX and citeWords behave for papers with various citation ranges.

4. AVERAGE COUNTX AND CITEWORDS

After identifying countX and citeWords as two features from the citation contexts, we study in detail as to whether these features are discriminative with respect to the number of citations. To normalize with respect to various ranges of citations, we only used the average values of countX and citeWords for a publication in each year starting from its publication.

To give a working example as to how these features are computed, Table 2 presents citation contexts for paper *P* titled as "*On Relaxed Dynamic Programming in Switching Systems*", published in 2005. The first column gives the citer_ids, which refer to MAS identifier for the papers citing paper *P*. Publication year of the citing paper is shown in column 2. Column 3 gives the exact citation context(s) in the citing paper for paper *P*. Below, we describe as to how the average countX and average citeWords features are computed for *P* over the years.

4.1 Average countX

For a citation edge from paper Q to paper P (i.e., Q citing P), countX denotes the number of times paper P is cited in paper Q. A high value of countX implies that paper P is cited multiple times by paper Q and thus, P might be quite relevant for paper Q. Possibly, Q has cited P for its different aspects. Our hypothesis is that if we consider all the papers citing paper P and find the average value of countX for P, it may serve as a very strong feature to measure the importance of P.

Let us assume that the papers $Q_1, Q_2, \ldots, Q_n$ are citing paper P for $N_1, N_2, \ldots, N_n$ times respectively in the t^{th} year after publication of P. We define the average countX metric for paper P for the t^{th} year as

$$average\ countX(P, t) = \frac{\sum_{j=1}^{n} N_j}{n} \tag{1}$$

Using the example in Table 2, average countX value for paper *P* for the first year after publication (year 2006) can be calculated as: $average\ countX(P, 1) = \frac{2+1}{2} = 1.5$. This is because there are two citing papers in year 2006, one of which cites P twice within the same paper while the other cites it only once.

4.2 Average citeWords

For a citation edge from paper Q to paper P, citeWords denotes the number of words in the citation context(s), where P has been referred to. Since more than one paper might be cited within the same citation context, number of words are divided among all the cited papers. Similar to countX, a high value of citeWords implies that paper P has been discussed in more details by paper Q and therefore, paper P might be relevant for paper Q. Dividing by the number of papers cited within the citation context takes care of the fact that the words in citation contexts have been used to describe multiple papers. Similar to countX, our hypothesis is that finding the average number of words that other papers use to describe P could be indicative of the importance of paper P.

Let us assume that paper P is cited by another paper Q_i in m different citation contexts, $S_1, \ldots, S_m$. For this citation edge, citeWords is computed as

$$citeWords(P, Q_i) = \sum_{i=1}^{m} AW(S_i, P, Q_i) \tag{2}$$

where $AW(S_i, P, Q_i)$ denotes the average number of words used in sentence S_i to describe P. In general, if $k \geq 1$ papers are cited within the sentence S_i, the average words for each of these k papers (including P) is given by:

$$AW(S, P, Q_i) = \frac{Len(S)}{k} \tag{3}$$

where $Len(S)$ denotes the length of a sentence S and is simply computed by counting the number of words appearing in it. Now, assume that the papers $Q_1, Q_2, \ldots, Q_n$ are citing paper P in the t^{th} year after publication of P. We define the average citeWords metric for paper P for the t^{th} year as:

$$Average\ citeWords(P,t) = \frac{\sum_{j=1}^{n} citeWords(P,Q_i)}{n} \qquad (4)$$

Using Table 2, average citeWords value for paper *P* for the third year after publication (year 2008) can be calculated as:

$(citeWords(P, 6413388) + citeWords(P, 5052733))/2$

To compute $citeWords(P, 5052733)$, we see that paper 5052733 cites P in one citation context when a total of two papers are cited. Thus

$citeWords(P, 5052733) = \frac{11}{2} = 5.5$, where 11 is the length of the citation context.

Similarly, paper 6413388 cites paper P twice but in both the citation contexts, two papers are cited. Therefore,

$$citeWords(P, 6413388) = \frac{25}{2} + \frac{16}{2} = 20.5$$

Thus, $Average\ citeWords(P, 3) = \frac{20.5+5.5}{2} = 13$.

4.3 Correlation between citation counts and citation content features over the years

We investigate whether the average countX and average citeWords values over the years are correlated with the number of citations a paper receives. We reiterate that both average countX and average citeWords are normalized with respect to the number of citations received by the paper. We divide the set of papers in our dataset into 6 buckets based on the following criterion on the number of citations.

Bucket 1: Top 0.1% papers – citations 389-7859

Bucket 2: Top 0.1 - 1% papers – citations 95-389

Bucket 3: Top 1 - 5% papers – citations 29-95

Bucket 4: Top 5 - 10% papers – citations 16-29

Bucket 5: Top 10 - 25% papers – citations 6-16

Bucket 6: Rest of the papers – citations 0-6

For each of the Citation buckets, we plot the temporal profile for the average countX values, averaged for all the papers within that bucket, in Figure 2. The $X-$axis denotes the year after publication for the paper, ranging from 0 (same year as publication) to 10 (10^{th} year after publication). While averaging for a citation bucket for a particular year, we consider only those papers which have non-zero citations in that year. Minimum value of countX can be 1 for any citation edge. Interestingly, as per our hypothesis, various citation ranges show differences in terms of the average countX values. Some important observations from Figure 2 are:

1. There is an increase in value of countX in initial years irrespective of the citation bucket, and it further decreases continuously over the years. A slight increase is observed for the 10^{th} year after publication.

2. Highly cited papers are cited more number of times in a single paper.

We clearly see a correlation between the number of citations and the average countX profiles of the papers. Further, we investigate whether the countX values can discriminate between the 6 citation categories identified in [7]. Accordingly, we divided the set of papers into 6 categories mentioned in [7]. For readability, the six categories are described below:

(i) PeakInit: Papers whose citation count peaks within 5 years of publication followed by an exponential decay.

(ii) PeakMul: Papers having multiple peaks in different time periods of the citation history.

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

(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.

(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.

(vi) Oth: Papers not belonging to any of the above mentioned categories belong to this category.

Figure 3 presents the temporal profile of average countX values for each of these 6 categories. Again, we can see that the average countX values are the highest for the $MonIncr$ and $PeakLate$ categories, which have been identified as having the categories corresponding to high number of citations in [7]. Similarly, average countX values are the lowest for the $MonDec$ and $Others$, which have been identified as the categories corresponding to the low number of citations (see [7] for details).

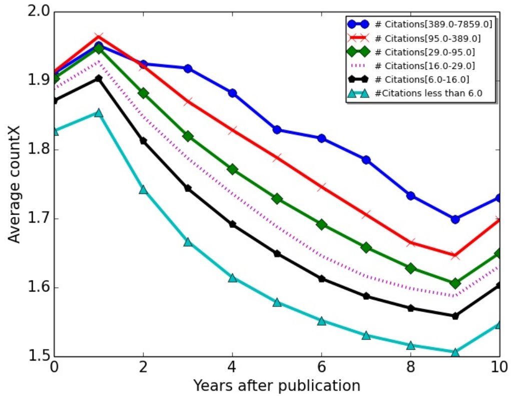

Figure 2: Average countX: temporal profiles for six citation buckets over the publication age

We now plot the temporal profile for the average citeWords values for the six citation buckets in Figure 4. Similar to average countX, while averaging for a citation bucket for a particular year, we consider only those papers which have a non-zero citation in that year. Average citeWords also shows a very similar trend as that seen with the average countX values, an initial increase and then a decreasing trend over the years. Interestingly, differences are observed between various citation ranges with the papers having the highest citations also earning a high number of average citeWords over the years.

We further use six citation categories to plot the temporal profiles in Figure 5. The trends are again very similar to those observed for the case of average countX values, with the MonIncr and PeakLate categories having a higher value of average citeWords than the other categories and MonDec category having the lowest values.

Table 2: Example citation contexts for paper (*P*) titled as *On Relaxed Dynamic Programming in Switching Systems*, published in 2005. Citer_id represents MAS identifier of the paper citing paper *P*. Publication year represents year of publication of citing paper. Finally, Context column contains the citing sentence. There are several instances where a paper is cited more than once in a citing paper. Also, a citing sentence might cite more than one paper. Citations are sorted in the descending order of the year of publication of citing paper. Bold face text represents cited paper reference.

citer_id	Publication year	Context
5330841	2010	Our approach relies on the following result from relaxed dynamic programming [**12**, 15], which is a straightforward generalization of proposition [5, Proposition 2.4], cf. [7] for a proof
6899965	2009	In [18, **19**] a relaxed dynamic programming procedure is proposed
		The existence of a solution is assumed in [18,**19**], which is different from the objective in this paper; to obtain a sub-optimal solution only when the minimum does not exist
5179891	2009	Recently, this has been studied by Lincoln and Rantzer in [**11**,17]
5977376	2009	The paper [**6**] (see also [5]) uses controllability conditions and techniques from relaxed dynamic programming [13, 18] in order to compute explicit estimates for the degree of suboptimality, which in particular lead to bounds on the stabilizing optimization horizon N which are, however, in general not optimal
6006644	2009	Our approach relies on results on relaxed dynamic programming [9], [**13**] already used in an MPC context in [7] which we adapt to our variable control horizon setting
6413388	2008	Inequalities of such type have been used frequently in the optimal control literature, however, a systematic study seems to have performed only recently in [**14**, 18]
		The approach we take in this paper relies on recently developed results on relaxed dynamic programming [**14**, 18]
5052733	2008	These general algorithms are also used to study switched systems in [**11**], [12]
5433268	2007	Some are based on a newly elaborated condition of optimality see e.g., [1], [**2**],[3], others are more related to semi-classical approaches see e.g., [4], [5], [6], [7]
4971068	2007	A novel approach to overcome some of the difficulties mentioned above was recently proposed in [**4**], [5], [3], see also [14] for examples from switching systems
12659162	2006	For further details on the theoretical foundations, the reader is referred to [**13**]]
		Further discussion of the implicit algorithm is given in [**13**]
50488928	2006	In a recent work, it is shown that the optimal control problem can be reformulated as an approximate linear-quadratic problem, whose complexity grows only polynomially [**10**]
2992574	2005	Konda and Tsitsiklis [KoT03], Marbach and Tsitsiklis [MaT01], **Rantzer [Ran05]**,
12894773	2005	In [16], This led some to develop relaxed DP techniques, e.g. (**Rantzer, 2005**)

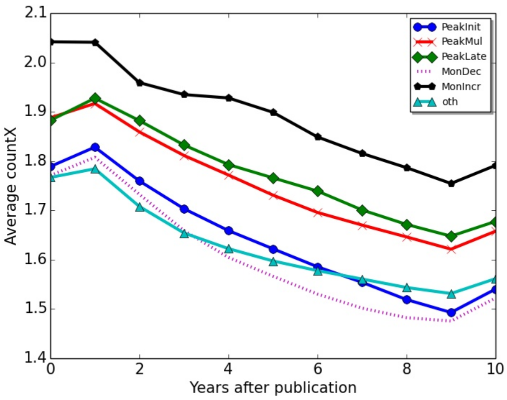

Figure 3: Average countX: temporal profiles for the six citation categories [7] over the publication age

4.4 Correlation between citation counts and citation context features for the initial years

To motivate the importance of average countX and citeWords as features for future citation prediction, Table 3 shows some specific examples of papers having the same citation count in the first two years after publication but different average countX and citeWords values. What we observe is that in both the cases, among the papers having the same citation count, the paper having a high countX (and citeWords) value in the initial two years receives a much higher citation count in the future. Thus, the average countX feature from the initial years of publication can serve as an important feature towards predicting future citations.

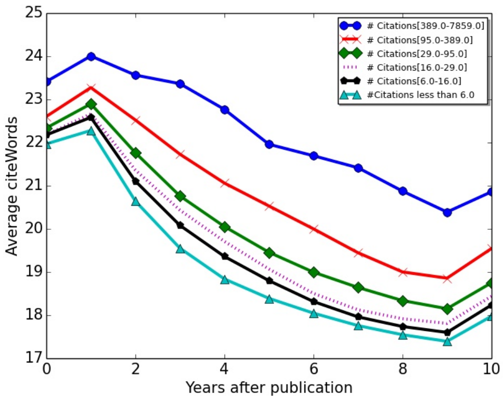

Figure 4: Average citeWords: temporal profiles for the six citation buckets over the first 10 years of publication age

Table 3: Example paper-pairs having a similar citation count in the initial 2 years of publication but different countX values.

Paper ID	Initial citation count	Initial Avg. countX	Initial Avg. citeWords	Final Citation count
349111	4	1.75	41.75	140
25	4	1	10.58	47
1911	7	3.29	47.9	155
349954	7	1.42	16.35	38

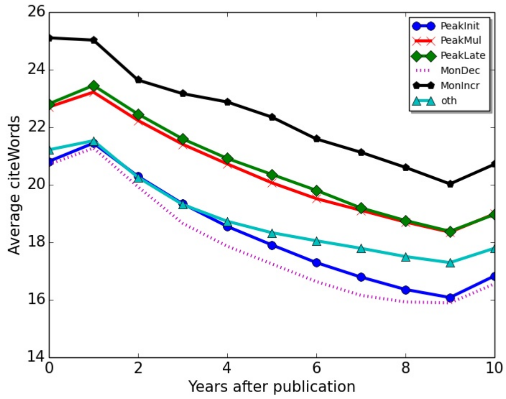

Figure 5: Average citeWords: temporal profiles for the six citation categories [7] over the first 10 years of publication age

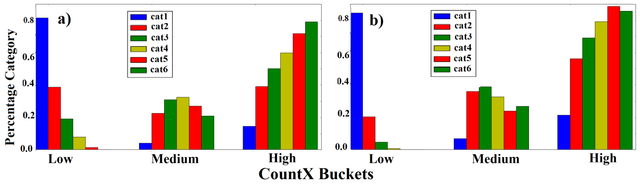

Figure 6: Correlating citation count and countX buckets. (a) Correlation at 5 years after publication; (b) Correlation at 9 years after publication

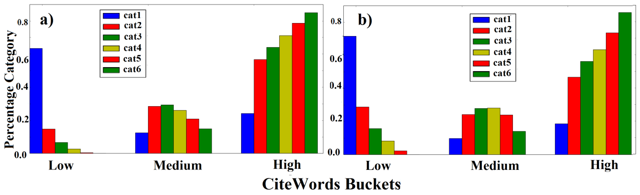

Figure 7: Correlating citation count and citeWords buckets. (a) Correlation at 5 years after publication; (b) Correlation at 9 years after publication

We further study whether the average countX and average citeWords values from the initial two years after publication can serve as discriminating features to predict citations at a later point of time. We, therefore, divide all the papers in 3 ranges as per the average countX values ($\{1\}$, $(1, 1.5]$ and $(1.5, -)$) and as per the average citeWords ($(0, 10.5]$, $(10.5, 16.5]$ and $(16.5, -)$ in the initial two years of publication. We call these ranges as low, medium and high respectively. We now take citation counts of the papers for the time points, corresponding to 5 and 9 years after publication. For each such time point, we create 6 different citation buckets (top 0.1% etc.) and plot the distribution of the papers falling into these 6 citation buckets on various countX and citeWords ranges. For example, 5 years after publication, 75% of the papers in the lowest citation category have an average countX value=1 (see Figure 6(a)). On the other hand, more than 75% of the papers in the top two categories (top 0.1% and top 0.1-1%) have a countX value ≥ 1.5. The trend becomes much more prominent for 9 years after publication (Figure 6(b)), with the probability of a paper having countX ≥ 1.5 increasing with increasing citation counts.

Very similar trends are observed for the average citeWords as well (see Figure 7). From these figures as well as examples in Table 3, it is clear that information from average countX and average citeWords in the initial years of publication acts as a discriminating factor for the future citation counts, such that most of the highly cited papers have high values of average countX and citeWords in the initial years, which is not true for the low-cited papers.

Motivated by these examples, we now use these citation context features for the task of future citation prediction. The model is described in the next section.

5. CITATION PREDICTION MODEL

We extend the two-stage stratified learning framework proposed in [6] with the addition of three features. In the first stage, a query paper is classified into one of the citation profile category using Support Vector Machine (SVM) learning model. Further, for each category, a Support Vector Regression (SVR) model is learned for predicting citation counts. Thus, given a query paper, we first classify it into one of the six citation profile categories. Post classification, category based SVR is used to predict citation count. Our citation prediction model uses features at the time of publication, along with the citation information from the first 2 years after publication. Features from the time of publication are the same as reported in [6]. These features can be divided into three categories: features based on the paper content, features based on author information and features based on venue information. For the sake of completeness, we describe these features in brief below. For more details, the reader is requested to look into [6].

5.1 Features based on paper content

We used five paper-centric features as proposed in [6]. Last three among these are entropy based features.

(a) Team-size (Team): The number of authors in a paper.

(b) Reference count (RefCount): The number of references mentioned in the reference section of a paper.

(c) Reference diversity (RDI): RDI measures the diversity in the fields of the referred papers. A paper citing papers of various fields has a high value of RDI.

(d) Keyword diversity (KDI): Keyword diversity refers to diversity in the keywords mentioned in the paper.

(e) Topic diversity (Topic): Each paper is assigned a set of probable topics inferred from LDA. Topic diversity gives a diversity over these probable topics.

5.2 Features based on Author information

The author of a publication plays an important role in its popularity. The following four author-centric features were used for citation prediction.

(a) Author h-index (HIndex): H-index is a standard measure for author productivity and impact. This feature measures average h-index of the authors at the time of publication.

(b) Author productivity (ProAuth): Author productivity refers to the count of his publications. A more productive author will produce more. The feature is an average of the productivity of the all the co-authors of a paper.

(c) Author diversity (AuthDiv): Author diversity refers to the diversity in the research fields of author publications. A highly diverse author will publish in different domains. The feature is an average of all the authors taken together.

(d) Sociality of author (NOCA): This feature counts the number of co-authors in all the publications of each author present in the paper.

5.3 Features based on venue information

We also used certain features based on the prestige as well as the diversity of the venue, where the paper has been published. These features are described in detail below.
(a) Short term venue prestige (VenPresS): Short term venue prestige measures the average number of citations for the papers published in a venue during the two preceding years.
(b) Long term venue prestige (VenPresL): Long term venue prestige measures the average number of citations for the papers published in a venue so far.
(c) Venue diversity (VenDiv): This feature measures the diversity in the research fields of the papers published in a venue.

5.4 Features after the publication year

In addition to these features, we also utilize the two features derived from the citation context, the average countX and average citeWords for the first 2 years after publication, as well citation count received after the first 2 years of publication.

In the next section, we report the experiments using our citation prediction model.

6. EXPERIMENTS

We perform experiments using the stratified learning framework for citation prediction. We selected papers having at least 10 years of history and published in between 1970 - 2005. We divided this dataset into training and testing sets. For training, we consider papers published in between 1970 - 2000. For testing purpose, we took the range as 2001 - 2005. First of all, we learn a stage-I classification model using our training dataset. We also learn separate regression models for each citation category, for each time point, for which the citation count is to be predicted. Given a query paper, first the classification model is used to assign a citation category (stratum) to it (stage I). In stage II, a regression model trained on the assigned category is used for citation count prediction for the specified time periods. We use all the features described in section 5. We have used three different time points $\Delta t = 5$, 7, and 9 for prediction.

We evaluate our model on two baselines. The first baseline [19] is similar to our model except that it does not include the classification stage. Thus, all the features are directly used in a regression model for citation prediction. We use Chakraborty et al. [6] as the second baseline. While the authors conducted experiments both with and without the initial year of publication information, we use the citation count of first two years for their method for a fair comparison. Thus, this baseline is very similar to our model with the only difference being that we use two citation context features identified in this paper, average countX and average citeWords, for the t^{th} year after publication, with $t = 0, 1, 2$.

6.1 Evaluation metrics

We use the following three metrics for evaluating our results.

6.1.1 *Coefficient of determination (R^2)*

Coefficient of determination (R^2) [4] is a number that indicates how well data fit a statistical model of future outcome prediction. It measures the variability introduced by the statistical model. It is defined as the proportionate reduction in uncertainty, measured by Kullback-Leibler divergence, due to the inclusion of regressors. Let d be the document in test document set D_T, we calculate R^2 as:

$$R^2 = \frac{\sum_{d \epsilon D_T} (C_{T_{ccp}}(d) - C_T(D_T))^2}{\sum_{d \epsilon D_T} (C_T(d) - C_T(D_T))^2} \quad (5)$$

Here, $C_{T_{ccp}}(d)$ denotes the predicted citation count for document d. $C_T(D_T)$ denotes the mean of observed citation counts for documents in D_T. ($C_T(d)$ denotes actual citation count for document d. R^2 values ranges from 0 to 1. A larger value indicates better performance.

6.1.2 *Pearson correlation coefficient (ρ)*

Pearson correlation co-efficient (ρ) [12] measures the degree of linear dependence between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (6)$$

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \quad (7)$$

Here, cov(X,Y) denotes covariance between X and Y, σ_X and σ_Y denote standard deviation values for X and Y respectively. Similarly, μ_X and μ_Y denote mean values for variables X and Y respectively. E represents the expected value. ρ ranges from -1 to 1, where $\rho = 1$ corresponds to a total positive correlation, 0 corresponds to no correlation, and -1 corresponds to total negative correlation. A larger value indicates better performance.

6.1.3 *Mean squared error (θ)*

Mean square error (θ) measures the expected value of the squared error loss in estimation. It is a risk function corresponding to the expected value of the squared error loss. For n number of observations, we define mean squared error as:

$$\theta = \frac{\sum_{i=1}^{n} (\hat{Y_i} - Y_i)^2}{n} \quad (8)$$

Here, $\hat{Y}$ and Y denote the vectors of predicted and actual values respectively. A smaller value indicates better performance.

6.2 Comparisons with the baseline models

Next, we compare the performance of the two baselines with our model. We also present performance statistics for stage I (classification) and stage II (prediction). Along with performance analysis, we compare categories and analyze results.

Table 4 compares the performance of these baselines with our model. Columns 2-4 in Table 4 show the predictive performance for baseline I using three metrics, while columns 5-7 show the predictive performance of baseline II. Columns 8-10 show the performance of our model.

We observe that for all the three systems, performance decreases with the increase in time period for prediction, with the best performance achieved for $\Delta t = 5$. While baseline I performs the worst among the three models, the R^2 value of 0.56 obtained for $\Delta t = 5$ is in itself significantly better that some previous works. For example, Kulkarni et al. [11] achieved an R^2 value of 0.2 using 328 medical articles. Baseline II performs better than baseline I for all of the three time-periods. This performance improvement can be credited to the stratified learning approach used in baseline II, as was established in [6]. Our model performs better than both the baselines for all the three time-periods. While the improvements over the first baseline are almost over 50% in terms of R^2, improvement of the order of 8-10% are achieved over baseline II as well. Improvement in terms of θ are of the order of 20-25% over the baseline II. Since the only difference between baseline II and our model are the average countX and average citeWords features identified in this paper, this improvement can be credited to the use of initial year information from the citation context of the paper.

Table 4: Performance comparison between Baseline I, Baseline II, and our model. Three evaluation metrics – θ, R^2 and ρ are used. A low value of θ and high values of R^2 and ρ represent an efficient model. Prediction is made over three time periods – $\Delta t = 5$, $\Delta t = 7$ and $\Delta t = 9$.

	Baseline I			Baseline II			Our Model		
	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ
Δ t=5	0.56	0.59	14.56	0.78	0.76	10.45	**0.84**	**0.79**	**7.86**
Δ t=7	0.54	0.57	15.90	0.74	0.72	12.57	**0.81**	**0.75**	**9.70**
Δ t=9	0.51	0.54	17.22	0.73	0.68	14.89	**0.78**	**0.74**	**12.43**

6.3 Category-wise performance analysis

Since we use the six categories as strata, we further analyze the prediction results for each of these categories. Table 5 presents category-wise performance metrics (except the category *Oth*) values for the three time-periods. Figure 8 gives the scatter plots for each category for the prediction task for the three time periods. $X-$axis denotes the actual citation count, while the $Y-$axis denotes the predicted citation count.

From Table 5, we observe that for $\Delta t = 5$, the performance is the best for the *PeakLate* category on all the three metrics. Figure 8 also confirms this observation with most of the points densely accumulated around $x = y$ line. For $\Delta t = 7$, *PeakLate* performs the best on ρ, while *MonDec* and *MonIncr* perform well on R^2 and θ respectively. For $\Delta t = 9$, *MonIncr* performs the best among all the categories for all the three evaluation metrics. Overall, *PeakLate* and *MonIncr* categories perform the best. This is very crucial for the citation prediction model, as these categories correspond to the highly cited papers [6].

From Figure 8, we observe that for $\Delta t = 5$, all categories show roughly the same pattern. Majority of the papers lie below the line, which denotes that in the initial years after publication our model slightly under-estimates the citation counts. The only cases of over-estimation are for the *PeakMul* category, $\Delta t = 9$ (majority papers above the line) and for the *MonIncr* category for $\Delta t = 7$.

6.4 SVM classification analysis

The first stage SVM model classifies each paper into one of the six categories. Table 6 presents the confusion matrix of SVM classification. Each entry in the first column represents a ground truth category of the paper. Similarly, each entry in the first row represents predicted category. We observe that around 50% of *Oth* category paper are wrongly classified into *PeakMul*. While *MonDec* has the highest accuracy (0.989), more than 29% *PeakInit* are classified into *MonDec*, which in turn decreases the accuracy for *PeakInit* category. As our dataset is highly biased towards *Oth* category (highest % of papers), SVM overestimates *Oth* category in the classification. Classification inaccuracy in the first stage decreases prediction accuracy in the second stage, with the *Oth* category playing a significant role in lowering the precision.

6.5 Paperwise analysis

Table 7 presents one best representative paper from each of the five categories. For each paper, we calculate the absolute difference between actual citation count and predicted citation count for our model, baseline I and baseline II for three time periods. As observed from Table 7, our model is closest to the actual values in terms of citations at any time instance.

6.6 Feature Analysis

We now study as to how various features correlate with the actual citation counts. Accordingly, we divide our features into 6 different sets and compute Spearman's correlation for the three time-periods in Table 8. We can see from the table that the last three features, namely average countX, average citeWords and 2-year citations, show a much higher correlation than the other three feature sets. While the correlation for 2-year citation feature is slightly higher than average countX for $\Delta t = 5$, correlation is the highest for average countX for $\Delta t = 9$. Thus, average countX serves as the most important feature for predicting the long term citation behavior of the papers.

Table 8: Average Spearman's rank correlation of each feature category (column 1) with the actual citation count without categorization for Δ t=5,7 and 9 years after publication

Feature category	Δ t=5	Δ t=7	Δ t=9
Author centric	0.387	0.342	0.317
Venue centric	0.343	0.309	0.285
Paper centric	0.429	0.417	0.392
Average countX	0.569	0.543	0.521
Average citeWords	0.512	0.499	0.481
2 year citation	0.571	0.543	0.502

6.7 Comparison with past works

The experimental results clearly confirm that the proposed method for citation prediction outperforms the other baselines for various time-periods. Further, we wanted to put this work in perspective of the previous related works for this problem. Table 9 lists five other works and compares them for the size of the dataset used for the study, year-ranges of the test papers, method used by the papers, as well a time period for which the R^2 values have been reported. Our dataset size is comparable to the other datasets reported in the literature. Also, we achieve a better R^2 value on this massive dataset than the ones reported earlier in the literature. Our prediction time period ($\Delta t = 9$) is the maximum among all these works.

7. CONCLUSION AND FUTURE WORK

In this paper, we have used a massive dataset of citation contexts to show that the features extracted from the citation contexts of the papers, in the immediate years after publication, play a vital role for the task of future citation prediction. We introduced two new features, average countX and average citeWords, and feature analysis showed that these citation context features are highly correlated with the actual citation counts, specifically for the long-range citation prediction.

For the citation prediction task, we used a stratified learning framework, similar to [6]. Experimental results confirm that including the citation context features significantly improves the accuracy over [6] under various experimental settings for different evaluation metrics.

In future, we plan to extend this work along many different dimensions. First, we only used two features from the citation context in this work. More features based on the textual analysis of citation contexts can be investigated. Also, we would like to investigate the classifier further to reduce the errors due to the classification stage. In addition, we plan to make the citation context dataset publicly available with additional insights and properties.

Table 5: Category-wise prediction accuracies using three metrics.

	PeakInit			PeakMul			PeakLate			MonDec			MonIncr		
	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ
Δ t=5	0.76	0.81	7.09	0.79	0.73	8.25	**0.89**	**0.83**	**1.96**	0.88	0.78	12.20	0.79	0.79	11.51
Δ t=7	0.77	0.72	9.91	0.78	0.76	9.78	0.81	**0.78**	9.88	**0.89**	0.77	9.86	0.80	0.76	**9.22**
Δ t=9	0.74	0.75	14.44	0.78	0.73	13.40	0.79	0.75	13.32	0.79	0.75	13.32	**0.79**	**0.79**	**12.61**

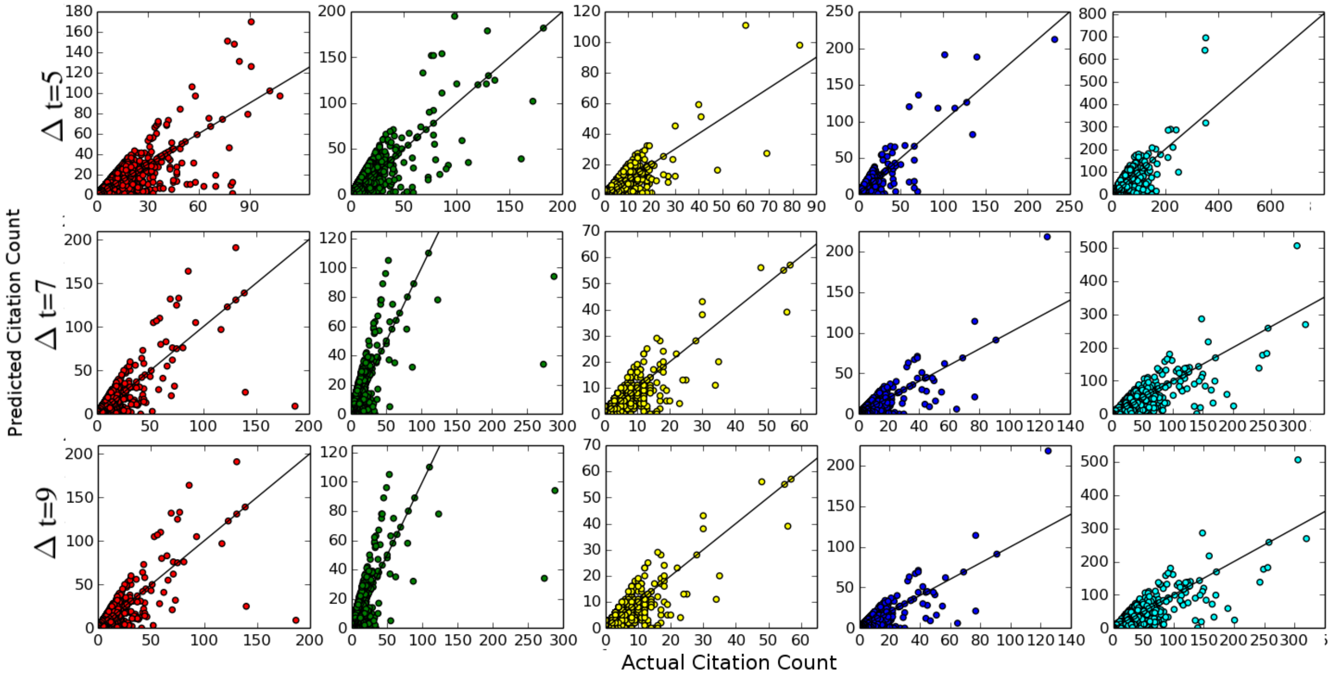

Figure 8: (Color online) Change in prediction over the time-periods for each category. Each scatter plot shows relation between actual citation count with predicted citation count. Here, from left to right, red color represents *PeakInit*, green color represents *PeakMul*, yellow color represents *PeakLate*, blue color represents *MonDec* and cyan color represents *MonIncr*. Black color line represents $x = y$ line passing through origin.

Table 6: SVM classification confusion matrix: Column 1 represents the ground truth categories, column 2 represents total number of papers in each of these categories, columns 3-8 represent the predicted categories and column 9 presents the accuracy values for each category. Correct classification results are highlighted in bold font from column 3-8. In column 9, highlighted bold font represents both the highest and lowest accuracy values

	No. of papers in category	PeakInit	PeakMul	PeakLate	MonDec	MonIncr	Oth	Accuracy
PeakInit	15178	**10987**	12	134	3245	43	757	0.724
PeakMul	30969	6	**27554**	1	1	0	3407	0.889
PeakLate	8946	49	0	**7298**	23	0	1665	0.815
MonDec	5263	1	22	0	**5207**	0	55	**0.989**
MonIncr	4010	1	64	1	0	**3005**	1003	0.749
Oth	142792	13	70618	23	1	0	**72138**	**0.494**

8. REFERENCES

[1] B. Aljaber, N. Stokes, J. Bailey, and J. Pei. Document clustering of scientific texts using citation contexts. *Information Retrieval*, 13(2):101–131, 2010.

[2] L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

[3] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

[4] A. C. Cameron and F. A. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[5] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *String processing and information retrieval*, pages 107–117. Springer, 2007.

[6] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to

Table 7: A best representative paper for each category: each paper is mapped to its MAS paper-id. Column 3 gives the actual citation count for the paper for 3 time points. Columns 4-6, 7-9 and 10-12 give the absolute difference between the actual citation count and the predicted citation count for the three systems for three different time-periods. Bold font represents the best predictions for each time period in each category. Values in parenthesis indicate predicted citation count.

Category	MAS Paper-id	Actual citation count (Δt =5,7,9)	Baseline I			Baseline II			Our Model		
			Δt= 5	Δt= 7	Δt= 9	Δ t= 5	Δt= 7	Δt= 9	Δt= 5	Δt= 7	Δt= 9
PeakInit	73939	35,20,5	7(28)	5(25)	14(19)	10(25)	**1(21)**	10(15)	**5(30)**	2(22)	**5(10)**
PeakMul	1447048	37,43,41	7(30)	11(22)	13(28)	9(48)	9(52)	14(55)	**5(42)**	**5(48)**	**3(38)**
PeakLate	837621	30,33,36	9(21)	8(25)	18(18)	8(38)	9(42)	19(55)	**5(35)**	**5(38)**	**9(45)**
MonDec	23419	8,6,3	7(1)	2(8)	7(10)	1(7)	1(7)	3(6)	**0(8)**	**1(7)**	**1(4)**
MonIncr	9871	18,20,26	6(12)	10(10)	10(16)	3(21)	3(23)	4(30)	**2(20)**	**1(21)**	**2(24)**

Table 9: Comparing related works in citation prediction: column 1 presents the title of the paper, column 2 presents the size of the dataset used in the paper, column 3 lists year range of test papers, column 4 presents the time periods used for prediction, column 5 lists the method/model used for prediction and column 6 presents the R^2 values reported in the paper for a time period, comparable across different methods. Papers are arranged in the increasing order of R^2 values.

Title of paper	Dataset Size	Year range	Time period(s)	Method	R^2(time period)
Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study [15]	1274	2005	3.5	decision trees	0.14(3.5)
Characteristics Associated with Citation Rate of the Medical Literature [11]	328	1999 - 2000	5	linear regression	0.2(5)
Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals [16]	204	1991	2	multiple regression	0.60(2)
Towards a Stratified Learning Approach to Predict Future Citation Counts [6]	1,549,317	2001-2005	1,2,3,4,5	SVR	0.71(5)
		1996 - 2000	1,2,3,4,5	SVR	0.74(5)
Citation Count Prediction: Learning to Estimate Future Citations for Literature [20]	1,558,499	1960 - 2011	5	CART	0.752(5)
The role of citation context in predicting long-term citation profiles: an experimental study based on a massive bibliographic text dataset [our model]	1,359,338	2001-2005	5,7,9	SVR	0.84(5)

predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 351–360. IEEE Press, 2014.

[7] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. On the categorization of scientific citation profiles in computer sciences. *CoRR*, abs/1503.06268, 2015.

[8] D. G. Feitelson and U. Yovel. Predictive ranking of computer scientists using citeseer data. *Journal of Documentation*, 60(1):44–61, 2004.

[9] L. D. Fu and C. Aliferis. Models for predicting and explaining citation count of biomedical articles. *PMC*, 2008:222–226, 2008.

[10] L. Getoor. Link mining: a new data mining challenge. *ACM SIGKDD Explorations Newsletter*, 5(1):84–89, 2003.

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

[12] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[13] L. Li and H. Tong. The child is father of the man: Foresee the success at the early stage. *arXiv preprint arXiv:1504.00948*, 2015.

[14] A. Livne, E. Adar, J. Teevan, and S. Dumais. Predicting citation counts using text and graph mining. In *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.

[15] C. Lokker, K. McKibbon, R. J. McKinlay, N. L. Wilczynski, and R. B. Haynes. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ*, 336(7645):655–657, 2008.

[16] C. M, W. RL, and W. E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287(21):2847–2850, 2002.

[17] N. Pobiedina and R. Ichise. Predicting citation counts for academic literature using graph pattern mining. In *Modern Advances in Applied Intelligence*, pages 109–119. Springer, 2014.

[18] D. Wang, C. Song, and A.-L. BarabÃąsi. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[19] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 51–60. ACM, 2012.

[20] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1247–1252. ACM, 2011.

[21] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, volume 12, pages 1119–1130. SIAM, 2012.