To Better Stand on the Shoulder of Giants

Rui Yan[†, ♮], Congrui Huang[†], Jie Tang[§], Yan Zhang[†, *] and Xiaoming Li[†, ‡]
[†] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[‡] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100083, China
[§] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[♮] Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan
{r.yan,hcr,lxm}@pku.edu.cn, jietang@tsinghua.edu.cn, zhy@cis.pku.edu.cn

ABSTRACT

Usually scientists breed research ideas inspired by previous publications, but they are unlikely to follow all publications in the unbounded literature collection. The volume of literature keeps on expanding extremely fast, whilst not all papers contribute equal impact to the academic society. Being aware of potentially influential literature would put one in an advanced position in choosing important research references. Hence, estimation of potential influence is of great significance. We study a challenging problem of identifying potentially influential literature. We examine a set of hypotheses on what are the fundamental characteristics for highly cited papers and find some interesting patterns. Based on these observations, we learn to identify potentially influential literature via Future Influence Prediction (FIP), which aims to estimate the future influence of literature. The system takes a series of features of a particular publication as input and produces as output the estimated citation counts of that article after a given time period. We consider several regression models to formulate the learning process and evaluate their performance based on the coefficient of determination (R^2). Experimental results on a real-large data set show a mean average predictive performance of 83.6% measured in R^2. We apply the learned model to the application of bibliography recommendation and obtain prominent performance improvement in terms of Mean Average Precision (MAP).

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.4 [**Information Systems Applications**]: General

General Terms

Algorithms, Experimentation, Performance

Keywords

Citation pattern analysis, influence prediction, digital libraries

*Corresponding Author.

JCDL'12, June 10–14, 2012, Washington, DC, USA.

1. INTRODUCTION

Effective scientific research requires keeping up with a large, ever-growing body of literature because scientists need to "stand on the shoulder of giants" (knowledge learnt from *Isaac Newton*). But how? Searching for meaningful work is tedious and it is very possible to miss important developments in areas outside a researcher's specialty. More critically, it may miss some potentially influential literature. In fact, considerable research work starts from a small number of initial papers and then explores papers near (citing or cited by) those papers. Therefore, to identify influential literature has long been viewed as one of the most important challenges in data mining for scientific literature. The rapid evolution of scientific research has been creating a huge volume of publications every year, and the explosive trend continues. Figure 1 shows statistics on a large literature database in Computer Science[1]. Figure 1.(a) visualizes the explosive increase on the volume of publications in the past years, in particular recent years. The number of publications in 2009 almost triples than that of 10 years before [28].

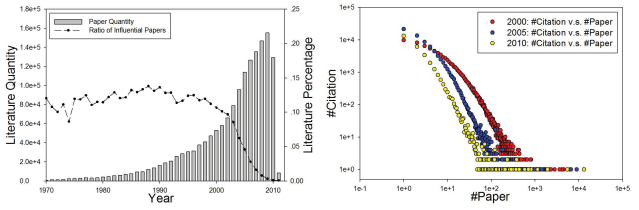
Given the large literature population, however, it is natural that not all publications contribute equal impact to academia. It is useful to identify influential literature to make better utilization of "giant shoulders" against unbounded publications. To measure such literature influence is non-trivial due to the numerous complicated methodologies from different perspectives, among which *citation count* is one of the most simple, standard and objective measurements. Citation count is calculated by how many times a particular publication is cited by other articles. In this study, we use the *citation count* to represent *influence* of literature for a simple and quick start, and both terminologies are used interchangeably.

The assumption of "different influence" is verified by showing the highly skewed citation pattern which follows a power law distribution: a huge number of research papers attract only a few citations, and a few research papers accumulate a large number of citations [28]. In all, there are 2.36% papers with more than 2000 citation counts and 68% papers have less than 5 citations! We also illustrate the changes of the citation distributions in different years in Figure 1.(b). The plots indicate more and more citations agglomerate to smaller percentage of literature from 2000 to 2010. This phenomenon coincides with the dotted line shown in Figure 1.(a) where the ratio of influential papers is decreasing year by year[2].

To measure the current influence by citation counts is not difficult but has been proved to be useful. In several bibliography search systems, current citation count is listed as one of the major factors to rank the candidate articles for the retrieval purpose [1]. Intuitively, we might consider the large probability of high future influence will also help select worthy candidates for literature search or recommendation systems (and indeed is verified by

[1] `http://arnetminer.org`.
[2] Here "**influential papers**" denotes papers with citations over 100.

our experiments), but no previous system is able to examine the effect of future impact. Another ambition for future influence prediction is that this technique encourages researchers to pay attention to and keep up with future influential works at an early stage. Our statistics shows that more than 20 papers received more than 100 citations merely after 5 years. Many of these influential papers finally lead to new research sub-fields. Obviously, being aware of these influential literature would put one in an advanced position in choosing new research topics and avoiding missing important references. One step ahead probably enables better and earlier "stand on the shoulder of giants". To answer the question we claimed in the beginning, identifying potentially influential literature via Future Influence Prediction (FIP) provides a possible option.



(a). The growing volume of literatures. (b). Distribution of literature citation.

Figure 1: Statistics of literature data from ArnetMiner.

It is challenging for FIP to identify fundamental factors correlated with future citations and to combine them under a unified predictive model. For identification of influential literature in the future, this paper has following contributions:

• The **1st** contribution is that we have explored a series of effective features important to future influence from several aspects, such as paper content, author expertise, venue impact, and more especially, **temporal dimension**, which has never been considered before. Furthermore, we examine these features based on experimental observations, not merely based on theoretical assumptions and show interesting discoveries in citation behaviors. We also analyze different roles in the combination of all features.

• The **2nd** contribution of our work is to model all relevant features to identify the potentially influential papers. Unlike previous studies, we put more focus on identification of highly cited literature, instead of treating them equally, i.e., giving emphasis on "giant shoulders". Given multiple features relevant to influence, i.e., citation counts in this study, we incorporate literature weights into regression models to estimate future impacts of scientific papers.

• The **3rd** contribution is that we integrate the factor of future influence into real practical application of bibliography recommendation system. The improved system outperforms its rivals and hence demonstrates the practical utility and benefits of FIP.

2. RELATED WORK

Influence, for evaluating research achievements, has long been heavily discussed by fundamental research journals (e.g. *Science*, *Nature* and *PNAS*). Impact Factor by Eugene Garfield is a measurement reflecting articles influence and is still pervasive [25, 10]. Journals with higher impact factors are deemed to be more important than those with lower ones [8]. However, impact factor can not reflect the influence of individual papers [7, 23] and hence needs a normalization from the audience of citing sides [30]. As to author aspects, the *h-index* is a useful index that attempts to measure both the productivity and impact of the published work of a scientist or scholar [13, 12]. However, both impact factor and h-index reflect the macro characteristics but the influence of a specific collection (all papers from a particular author or venue) may be skewed by individuals from our observations.

Citations indicate the influence of authors, papers and venues, and several works have conducted to analyze citation behaviors [1, 22, 26] and have perceived interesting discoveries for impact [21, 27]. In recent years, several researchers have investigated the prediction of citation counts. Their work differs primarily as regards the features used for prediction.

The 2003 KDD Cup includes a citation prediction task resembling the one we undertake on this paper. The citation prediction task includes estimating the *change* in the number of citations of papers between two different periods of time [11]. Several papers predict the number of citations using information gathered *after* publication. Brody et al. used download data within 6 months after publication as a predictive feature [4]. However, the aim was to show the Open Access advantage. Castillo et al. used a linear regression for the number of citations, the authors' reputation and the source of the paper citations (author related information) after a short period of time as predictive features [5]. Lokker et al. used features related to the article and journal, like number of authors, pages, references and so on [16]. These three works used measures taken after the paper was published to predict its citation count in the future. The main disadvantage of using this feature is that the required values are not available until after publication, and such features are difficult to access.

On the other hand, Fu et al. attempt to forecast citation counts using features which are are available at the time of publication [9]. Support vector machine regression models are used as the learning algorithm. Ibáñez et al. take into account several regression methods and analyze which one provides better predictions for the problem especially to predict annual time horizons [14]. Predictions in these works are made for a simple binary response variable that is defined by a set of citation thresholds to determine if an article is labeled positively or negatively [6]. Unlike their works, we do not deal with the response variables as several fixed thresholds, but to predict the exact citation count for each individual article. We also exploit the information output by the model, like the identification of key features that increase the chances of citation. This method can actually inform publishers about which articles will have larger influence in the future **before** they are published.

Unlike previous studies, we formally research into a predictive task of FIP and we add more relevant features into consideration. The relevance of features is investigated based on real world observations. We also integrate future influence into real world application of bibliography and obtain prominent improvement. To the best of our knowledge, we are the first to formally research into identification of potentially influential literature and to incorporate FIP into real applications such as literature recommendation.

3. FUTURE INFLUENCE PREDICTION

3.1 Problem Definition

In this section, we first present several necessary definitions and a formal representation of the influence prediction problem.

Influence. *Given the literature corpus* ***D****, the Influence* ($INF(.)$) *of a literature article* $\boldsymbol{d} \in D$ *is defined as the citation count of d:*

$$\begin{aligned} citing(d) &= \{d' \in D : d' cites\ d\} \\ INF(d) &= |citing(d)| \end{aligned} \tag{1}$$

Learning task: Given a set of article features, $\vec{X} = x_1, x_2, \ldots, x_n$, our goal is to learn a predictive function $\mathbb{F}(.)$ to predict the influence of an article d after a give time period Δt. Formally, we have

$$INF_{FIP}(d|\Delta t) = \mathbb{F}(d|\vec{X}, \Delta t) \tag{2}$$

Given the literature collection *D*, the learned influence predictive

function is actually to minimize the following objective cost function $\mathcal{O}(D)$, and the FIP task changes into an optimization problem for an optimal predictive function $\mathbb{F}^*(.)$:

$$\begin{aligned}\mathbb{F}^*(d) &= \underset{\mathbb{F}}{\operatorname{argmin}}\ \mathcal{O}(D) \\ &= \underset{\mathbb{F}}{\operatorname{argmin}}\ \sum_{d\in D} |INF_{FIP}(d) - INF(d)| \\ &= \underset{\mathbb{F}}{\operatorname{argmin}}\ \sum_{d\in D} |\mathbb{F}(d) - INF(d)|\end{aligned} \tag{3}$$

The cost function $\mathcal{O}(D)$ in Equation (3) is to some extent insufficient because it assumes that all papers are equally important. However, our goal is to identify the highly influential ones. Due to the influence distribution in Figure 1.(b), we model the skewed literature weight into cost function $\mathcal{O}(D)$, i.e., to increase the cost for errors on highly influential literature during training, and we change $\mathcal{O}(D)$ into:

$$\mathcal{O}(D) = \sum_{d\in D} \Psi(d) \cdot |INF_{FIP}(d) - INF(d)| \tag{4}$$

where $\Psi(d)$ is the weight of the paper *d*. The weight can be defined as a normalization to [0, 1] based on the influence distribution on data set D. We emphasize the weight for highly influential papers which are known, and punish the ordinary papers. The weight $\Psi(d)$ is defined as follows:

$$\Psi(d) = \frac{log(1 + INF(d)) - log(1 + INF_{MIN})}{log(1 + INF_{MAX}) - log(1 + INF_{MIN})} \tag{5}$$

where $INF_{MAX/MIN}$ means the maximum and minimum citation counts which are known to us.

The optimal predictive function $\mathbb{F}^*(.)$ becomes:

$$\mathbb{F}^*(d) = \underset{\mathbb{F}}{\operatorname{argmin}}\ \sum_{d\in D} \Psi(d) \cdot |\mathbb{F}(d) - INF(d)| \tag{6}$$

Before proposing our approach for FIP, we first probe a series of analysis by focusing on the following aspects of input literature articles, and then present the interesting citation patterns observed. Ideally, we ought to consider as many factors as possible. All features in consideration (except for temporal information which is special and is discussed in details) can be grouped into three facets: (1) paper contents, (2) author expertise, and (3) venue impact. Finally, it is important to find unified models which are able to consider all the features simultaneously.

3.2 Contents Feature Definition

3.2.1 Novelty

Novelty is a key criterion to evaluate paper quality, and is measured by similarity between a particular article and the other publications. An assumption is that a low similarity means a high novelty. We investigate whether higher novelty attracts more citations.

Novelty can be measured by similarity against all other literatures. However, such metric leads to an overestimated novelty because most papers from different research fields have naturally low similarity. Therefore, we measure an article's novelty against all its references: these papers are generally from the same sub-area and are supposed to have strong relevance. If the article d is significantly different from its references, we presume this phenomenon secures prominent novelty, which is calculated by Kullback-Leibler divergence D_{KL}:

$$Novelty(d) = \frac{\sum_{d'\in D_R} D_{KL}(\Theta_d||\Theta_{d'})}{|D_R|} \tag{7}$$

where $D_R = \{d'|d' \in referring(d)\}$, which is the collection of the reference papers cited by article d, and $referring(d) = \{d' \in D : d\ cites\ d'\}$. Θ_d is the word distribution of article d. V is the vocabulary set and $p(w|\Theta_d) = \frac{tf(w,\Theta_d)}{\sum_{w'\in V} tf(w',\Theta_d)}$ where *tf* denotes the term frequency for word w.

$$D_{KL}(\Theta_d||\Theta_{d'}) = \sum_{w\in V} p(w|\Theta_d) \log \frac{p(w|\Theta_d)}{p(w|\Theta_{d'})} \tag{8}$$

KL-divergence is asymmetric. To measure such novelty, it makes more sense to use $D_{KL}(\Theta_d||\Theta_{d'})$ than $D_{KL}(\Theta_{d'}||\Theta_d)$ because article d is inspired and motivated by all its references $d' \in D_R$.

Figure 2.(1) is interesting: the plot increases in the beginning, showing that the citation counts positively correlate with novelty and then after a certain threshold, the plot decays. This phenomenon indicates that **for articles which are divergent too far away from mass focus, they are unlikely to attract many citations**.

3.2.2 Topic Rank

Topics have long been investigated as a significant feature for literature contents [15, 18]. We utilize the unsupervised Latent Dirichlet Allocation [2] to discover topics[3]. Ideally, topics should be trained on the set of full contents of all papers. Because no such data is readily available, here we use the following proxy: we treat the title and the abstract of an article as the approximation.

We empirically train a 100-topic models and obtain the probability distribution over topics assigned to a literature article d, i.e., $p(topic_i|d)$, the inferred probability of topic *i* in document *d*. To calculate the influence of a particular topic from article d, denoted by $INF(topic_i|d)$, we distribute the influence of the article $INF(d)$ according to the topic distribution, i.e., $INF(topic_i|d) = INF(d) \times p(topic_i|d)$ and we obtain the influence of all topics by:

$$INF(topic_i) = \sum_{d\in D} INF(topic_i|d) \tag{9}$$

where D is the whole literature collection. We rank topics by average citation counts. From Figure 2.(5), we see different topics have different expected average citation counts. **Popular topics accumulate more citation counts than unpopular ones: topic popularity is relevant to literature influence**.

3.2.3 Diversity

Diversity indicates the breadth of an article from its topic distributions. This is important for identifying methodology papers, which are often cited by a wider topical range of articles. When an article has a vast range of audience, it is likely to be cited by authors from various research fields, and hence attract high citation counts. To measure the topical breadth of an article, we calculate the entropy of the document's topic distribution:

$$Diversity(d) = \sum_{i=1}^{|\mathcal{T}|=100} -p(topic_i|d) \cdot \log p(topic_i|d) \tag{10}$$

We hereby calculate the correlation between paper diversity and corresponding average citation counts shown in Figure 2.(3), which is quite interesting because the plot indicates **few papers belong a narrow focused area and in general expected citation counts increase as diversity enlarges**.

3.3 Author Feature Definition

In this section, we will answer the question that how an author's expertise correlates with the number of citation and how to measure

[3]We use Stanford TMT (`http://nlp.stanford.edu/software/tmt/`), with default settings for all parameters.

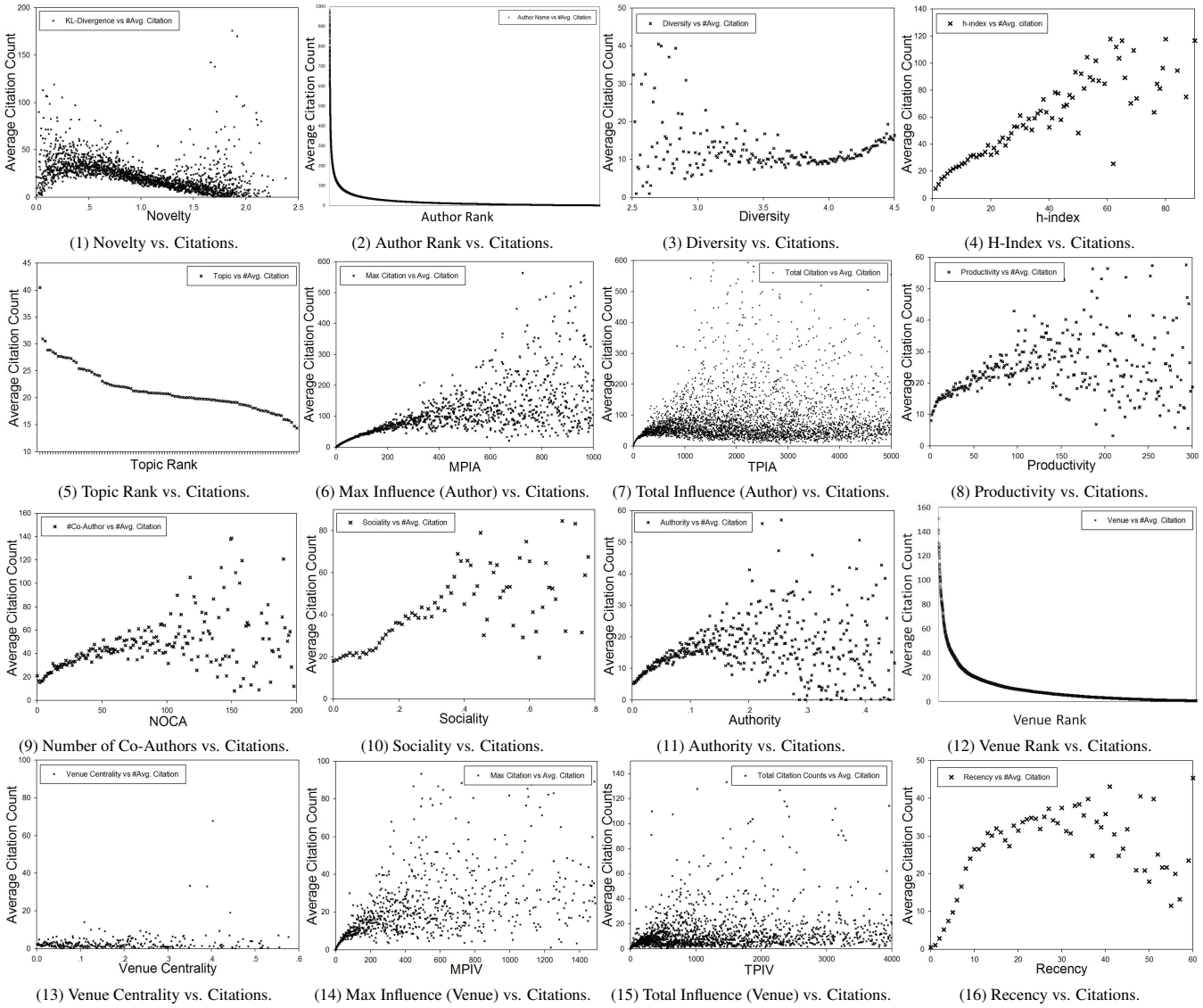


Figure 2: The feature-average citation correlations: x-axis denotes the value of a particular feature (e.g. sociality, authority, etc.), and there are a group of papers with the same value of the examined feature. We take the average citations of these papers as y-value.

such expertise quantitatively. To be self-contained, we first briefly review some of the features introduced in [28], together with our new insights, and then analyze these citation patterns.

3.3.1 Author Rank

Yan *et al.* assume that the "fame" of an author ensures the amount of citations [28]. We calculate the average citations for every author, and then assign an author rank value to him/her according to the rank number by his/her expected citation counts. As shown in Figure 2.(2), the plot of **expected citations is highly skewed, but for most of the authors, their expected citation counts are nearly the same and are rather small**.

3.3.2 H-index

The h-index is useful which attempts to measure both the productivity and impact of the published work of a scientist [12]. Therefore, we examine the correlation between h-index and citation counts. From Figure 2.(4), we observe a highly positive correlation between h-index and average citation counts: the correlation is almost linear, which proves **h-index an effective indicator of influence**.

3.3.3 Past Influence of Authors

Past influence probably indicates future influence. It is reasonable to assume the previous high influence for an author may result in the future high influence, which explains how reputation establishes. There are two ways to measure author past influence: previous (1) maximum citation counts and (2) total citation counts.

• *Maximum Past Influence of Authors (MPIA).* In the real world, one well-known publication helps to recognize academic reputation. The correlation between average citation counts and maximum citation counts is shown in Figure 2.(6), indicating expected citations strongly correlate with MPIA and **one widely acknowledged paper indeed benefits author reputation**.

An interesting observation is that for authors with more than 500 MPIA, the correlations seems to be a chaos with large variances. There are two possible reasons for this situation: 1) the distribution of MPIA per author follows the power law, which means most of

the authors have a relatively small MPIA while few authors have a large MPIA. Therefore, for few authors with high MPIA, there is probably lack of average effect and as a result, large variances emerge. 2) The noise of author statistics for high MPIA ranges.

• *Total Past Influence of Authors (TPIA).* Another measurement of past influence is to calculate total citations per author. From Figure 2.(7) we conclude the weak correlation between expected citations and TPIA: the fitted curve is almost parallel to the x-axis, indicating **total citation of an author might not be predictive**.

3.3.4 Productivity

According to [1], authors have tendencies to cite their own papers: the more productive an author is, the larger chances for his/her papers to be cited. Due to such self-citation behavior analysis, we examine the predictive power of productivity in Figure 2.(8). We notice that productivity and citation counts have a positive correlation: **the more papers an author have published, the higher average citation counts he/she could expect**. Similar to Figure 2.(6), the large variances in the high value range might be due to noisy data and lack of average effect for these few authors.

3.3.5 Sociality

Researchers tend to cite papers from whom the author(s) have co-authored [1]. Hence a widely connected author is more likely to be highly cited by his/her wide variety of co-authors. A straightforward social measurement of Number of Co-Authors (NOCA) is introduced in [28] and we examine the correlation between NOCA and average citation counts. As shown in Figure 2.(9), the plot implies a prominent positive correlation.

However, sociality measured by the co-author numbers is simple and reflects insufficient social relationships: authors have different social weights. Collaboration with the same amount of authors of high (or low) sociality might lead to different influence but "the number of co-authors" fails to distinguish such difference. We then measure sociality based on the real academic social network. We establish a co-author network graph $G_c(V, E)$ to discover the research communities, where V is the set of vertices and each vertex v_i in V represents an author. E is the set of edges which is a subset of $V \times V$. Each edge denotes the co-authorship and is associated with an author affinity weight $A_{aff}(v_i, v_j)$ between vertex v_i and v_j ($i \neq j$). The weights of the edges are calculated by the times of collaboration, i.e., $A_{aff}(v_i, v_j) = |D_{co}|$ where co-authored paper set $D_{co} = \{d|(author(d) \cap (v_i \cup v_j)) \neq \varnothing\}$. The transition probability between v_i and v_j is then defined by normalizing the corresponding affinity weight [28].

We use the row-normalized matrix $M = M_{i,j_{|V|\times|V|}}$ to describe G_c with entry corresponding to the transition probability, i.e., $M_{i,j} = p(v_i, v_j)$. Based on the matrix M, the sociality of an author v_i (denoted as $Sociality(v_i)$) can be deduced from all other authors linked with him/her, which can be formulated in a recursive form as in the PageRank algorithm.

$$Sociality(v_i) = \mu \sum_{j \neq i} Sociality(v_j) \cdot M_{j,i} + \frac{1-\mu}{|V|} \quad (11)$$

where μ=0.85. We conduct an examination of the correlation between sociality and average citation counts. **The positive effect of sociality has been confirmed in Figure 2.(10)**.

3.3.6 Authority

Besides the co-authorship network, another heterogenous social network for academia is established from the "citing - cited" relationships among literature articles. A widely cited paper indicates peer acknowledgements, and hence indicates author authority. We transmit paper authority to all its authors. We build a graph of $G_a(V, E)$, where V denotes the paper collection and E denotes the *citing-cited* linkage with directions. The out-degrees measure how many times a paper is cited while in-degrees indicate the references of a particular paper. Each paper is represented as a term vector of semantic words from the vocabulary, and we calculate the standard cosine similarity between two papers as the weight of the paper affinity in the graph, i.e., $P_{aff}(v_i, v_j) = sim_{cos}(v_i, v_j)$. After a similar PageRank procedure as Equation (11), we obtain the authority score $Authority(d)$ of each paper d. We define the authority of an author a as:

$$Authority(a) = \sum_{d \in D_a} Authority(d) \quad (12)$$

where $D_a = \{d|a \subseteq author(d)\}$. We examine the correlation between these two variants in Figure 2.(11). **The strong correlation between authority and average citations is hence verified**.

3.3.7 Versatility

Like paper contents, a wide topic breadth of an author's research implies large amount of audience from various research fields, which is similar to *diversity* to identify highly influential papers cited by a wider topical range of articles. When an author has a vast range of audience, his/her papers are likely to be highly cited. To measure the topical breadth of an author a, namely *versatility*, we calculate the entropy of the author's topic distribution:

$$Versatility(a) = \sum_{i=1}^{\mathcal{T}=100} -p(topic_i|a) \cdot \log p(topic_i|a) \quad (13)$$

where

$$p(topic_i|a) = \frac{\sum_{d \in D_a} p(topic_i|d)}{|D_a|} \quad (14)$$

where $D_a = \{d|a \subseteq author(d)\}$. We hereby calculate the correlation between author versatility and corresponding average citation counts. The pattern distribution is quite similar to Figure 2.(3) and due to page limits we omit the near-duplication of Figure 2.(3). **Few authors have a narrow focus and the expected citation counts increase as versatility enlarges**.

3.4 Venue Feature Definition

Like authors, venues also have academic reputations. From our observations, some venues have larger probability to be highly cited than others. We hereby investigate the venue impact on citations.

3.4.1 Venue Rank

Prestigious venues attract more focus just as the author rank pattern. As shown in Figure 2.(12), the reputation of a venue ensures different citations. Compared with Figure 2.(2), although plot in Figure 2.(12) is not that skewed, **the differences of expected citations for different venues are prominent enough**.

3.4.2 Venue Centrality

High citations of a particular conference or journal indicate peer acknowledgement from other venues, and we aim to find such central venues. Venues are connected by paper *citing-cited* linkage. We establish a venue connective graph $G_v(V, E)$ where V denotes the venues and the edges E denote the citing-cited relationships between venues. Like $G_a(V, E)$, $G_v(V, E)$ also has directions: the out-degrees measure how many times a venue is cited by papers from other venues while in-degrees denote citations. The weight of each edge is calculated by the number of citations between two venues. Hence, the venue centrality can be calculated via a similar

PageRank algorithm as Equation (11) and we examine the correlations between venue centrality and average citation counts, illustrated in Figure 2.(13). **To our surprise, the feature of venue centrality seems to imply little relevance with citations**.

3.4.3 Past Influence of Venues

Similar to author expertise, past venue influence might indicate probable future success. It is natural to assume the previous high influence for a venue may result in the future high influence. We still use (1) maximum past influence of venues (MPIV) and (2) total past influence of venues (TPIV) to measure past venue influence, showing in Figure 2.(14) and 2.(15). **Interestingly, plots of MPIV and TPIV share similar patterns with Figure 2.(6) and 2.(7)**.

3.5 Temporal Feature

Temporal dimension has long been proved to be significant in literature studies [1, 29]. Intuitively the citation counts accumulate as time passes by, thus a measure of the age of an article is important. We include as a feature the number of years since the article was published. We expect a positive correlation on temporal recency - the longer an article is published, the more citations it may receive.

As indicated in Figure 2.(16), the effect of temporal recency is highly prominent for the first 10 to 20 years. The average citation counts sheerly increase for the first decades when the articles get published. The increase rate slows down for the following years: it is natural that articles are attractive when they are to some extent "new" to researchers. Only few classic works accumulate citation counts steadily for decades. We notice that for literatures published more than 40 years ago, the citation counts decay and have a larger variance. This situation may be due to the noise for the aged publications when online literature libraries were not available. The decay may also be due to literatures attracts citations from temporal recent publications, but as shown in Figure 1.(a), the proportion of aged publications (e.g., more than 40 years) is rather small. Large amount of new articles are not likely to cite aged literatures. Therefore, aged literatures have low citations in general.

Growing Factor. Nothing can catch people's eyes more than a rising phenomenon. For instance, an author with rapid accumulation of citations or publications in recent years might indicate high influence to the academia. Therefore, it is reasonable to incorporate growing factor of preference for new scientists, new venues and newly developed research topics. For the growing effect analysis, we create another dataset constituted by literature of recent N years only, namely **RData**. Now we have two datasets: RData and the full dataset by literature of all years (**FData**). We study all features and examine different patterns on these two datasets, as well as the combination of both sets.

Decaying Factor. As the scenario of future influence prediction requires invisibility of future literature patterns, it is natural that the influence calculated from current situation decays after a particular given time period Δt. For simplicity, we use the traditional decay function for the temporal dimension, i.e., the decayed influence is measured as $INF(d) \times e^{-\gamma \Delta t}$, where $\gamma \in [0,1]$ is a scaling factor and $\Delta t = t_f - t_p$. t_p is the publish year and t_f is the future year to predict citations counts.

In all we have 32 features to predict future citation counts (listed in Table 3), including *full-feature* and *recent-feature*. Considering growing factor by distinguishing FData and RData doubles the number of available features. All recent-features are calculated based on the RData. Note that the 100 topics for FData and RData are the same, and hence we do not introduce a duplicate feature of *recent diversity*. Similarly, *recent recency* is fully covered by *recency*. Note that there may be multiple authors for a single paper, and it is inappropriate to use the first author only because all authors are assumed to have contributions to the influence of the paper. We create a **virtual author** whose feature values are averaged based on all authors of the paper. We experiment different combinations of these features.

3.6 Predictive Models

3.6.1 Gaussian Process Regression

Given a compact feature representation amenable for learning, our objective is to estimate an article's expected citation counts. Note that predicting the times that an article will get cited is an extremely hard problem. Due to the complex mechanism of future influence estimation, it is likely that they are a non-linear function of all features used to represent the data. Gaussian Processes provide a Bayesian formulation for non-linear regression, where the prior information about the regression parameters can be easily encoded. This property makes them suitable for our problem formulation.

A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution [20]. Given the feature vector X, the estimated citation $INF(d)$ for d is:

$$INF(d) = K(X, X_T)[K(X_T, X_T) + \sigma^2 I]^{-1} INF(d_T), \quad (15)$$

where K is a kernel function returning a kernel matrix, X_T is a matrix of feature vectors for the training papers, σ is a noise parameter, I the identity matrix and $INF(d_T)$ the vector of citation counts of the training article d_T. Note that we only use the mean and not the variance estimated by the GP. We use a Gaussian kernel in our experiments. We notice that the performance of GP has a weak dependence on σ and set it to 0.5.

3.6.2 CART Model

We then fit a Classification and Regression Tree (CART) model [3], in which a greedy optimization process recursively partitions the feature space, resulting in a piecewise-constant function where the value in each partition is fit to the mean of the corresponding training data. Folded cross-validation [19] is used to terminate partitioning to prevent over-fitting. Our model included 32 features summarized in the last section as predictors.

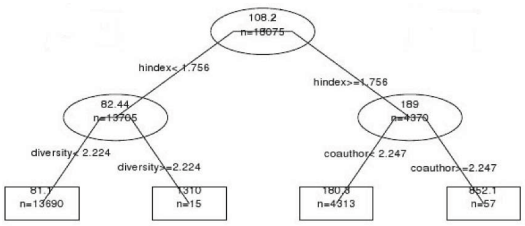


Figure 3: An example of regression tree for citation prediction.

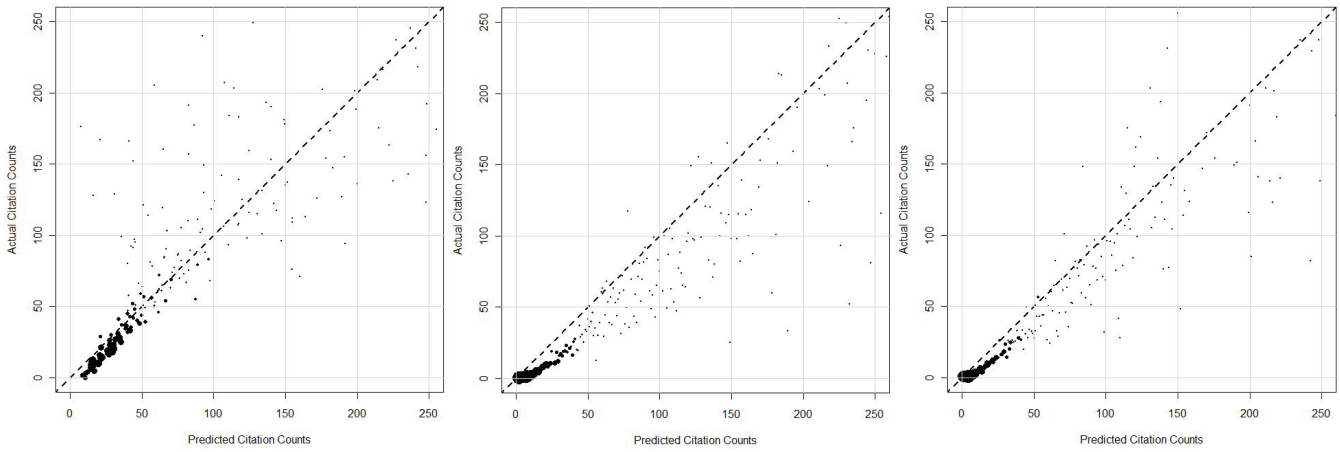
Figure 3 shows the regression tree for one of the folds. Conditions at the nodes indicate partitions of the features, where the left (right) child is followed if the condition is satisfied (violated). Leaf nodes give the function value for the corresponding partition. Thus, for example, one of the leaves indicates that papers with *h-index*$\geq$1.756 and *NOCA*$<$2.247 are predicted to have the influence of 180 citation counts.

4. EXPERIMENTS AND EVALUATION

4.1 Data Description

We perform influence prediction on the real-world data set[4], which is extracted from academic search and mining platform ArnetMiner.

[4]Downloaded from `http://arnetminer.org/citation`.



(1). 5-Year unweighted FIP for Year 2005, regression = GPR.(2). 5-Year weighted FIP for Year 2005, regression = GPR.(3). 10-Year weighted FIP for Year 2000, regression = GPR.

Figure 4: Actual vs. predicted citation counts: the performance for FIP with full features. The dotted line $y = x$ means the best result of predicted citation counts = actual citation counts. Figure 4.(2)-4.(3) incorporate literature weights while 4.(1) does not.

It covers 1,558,499 papers from major Computer Science publication venues and has gathered 916,946 researchers for more than 50 years (from 1960 to 2010). Two heterogeneous networks are included: one is the co-author collaboration network and the other is the paper citation network. The full graph of co-author network has 916,946 vertices (authors) and 3,063,257 edges (co-authorship), and the full graph of citation network has 1,558,499 vertices (literature papers) and 20,083,947 edges (citations).

To predict the citation counts after one year, we randomly take 10,000 papers from the literature collection from Year 2009 as the test set, and another random 10,000 papers from the Year 2009 as the development set. Note that for all training and evaluation, we only used features calculated over previous years. For example, when predicting articles published in Year 2009, all the articles up through Year 2008 are processed, and only the articles from the Year 2009 are available (as test set). Thus, these time dependent features would only include papers published in 2008 and earlier. Structuring the evaluation in this way is more realistic - when presented with new coming articles, the system can only predict possible future influence based on the patterns it has previously observed. We take the same procedure to predict citation counts after 5 (and 10) years with 10,000 test papers and 10,000 development papers from Year 2005 (and Year 2000). For unobserved feature values, e.g., new authors or new venues, we use the minimum feature values instead of N/A: anything has a start. We compare predicted citation counts with actual influence from the test data.

4.2 Algorithms for Comparison

We implement the following citation prediction algorithms as the baseline systems. All baselines are originally designed for traditional influence prediction problem rather than identify the influential ones. For fairness we apply the same biased objective function and the same pre-processing procedure for all algorithms.

• **kNN.** *k*-Nearest Neighbor (kNN) is used to predict citation by Ibáñez et al [14], which predicts the influence value for the paper *d* to be the average of the values of its *k*-nearest neighbors and the distance function measuring neighbors is based on a similarity calculation method such as cosine similarity. The neighbors are taken from training set for which real citation counts are known.

• **LR.** Lovaglia et al. propose a Linear Regression model to predict citation counts [17].

• **SVR.** Based on Support Vector Machine (SVM) model, Fu et al. build a Support Vector Regression (SVR) model to predict citation counts of biomedical publications using only predictive information available at publication time [9].

• **CART.** The CART method was used by Yan et al. on a limited feature space without incorporation of literature weight [28].

• **GPR.** The GPR method is our newly proposed approach for FIP problem, using the Gaussian Process for prediction.

4.3 Evaluation Metric

The coefficient of determination R^2 [24] is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of related features. It is the proportion of variability in a data set that is accounted for by the statistical model, which provides a measure of how well future outcomes are likely to be predicted by the model. The definition of R^2 is:

$$R^2 = \frac{\sum\limits_{d \in D_T} (INF_{FIP}(d) - INF(D_T))^2}{\sum\limits_{d \in D_T} (INF(d) - INF(D_T))^2} \tag{16}$$

where $INF_{FIP}(d)$ is the predicted citations for article d in the test set D_T and $INF(D_T) = \frac{1}{|D_T|} \sum_{d \in D_T} INF(d)$ is the mean of the observed citation counts for an article in D_T. $R^2 \in [0, 1]$, and a larger R^2 indicates better performance and hence is desired.

4.4 Performance and Strategy Analysis

The performance of FIP for different years is visualized in Figure 4, and the detailed results are summarized in Table 1. The size of circles indicates the number of points in each predicted citation counts. Most circles are gathered within in the range of [0, 50], indicating most of the papers have relatively low citations. Among different prediction tasks, 10-year prediction has the most remarkable performance. Furthermore, we notice a probable trend among these series of experiments: accuracy increases as Δt increases. We will examine the sensibility of Δt in the next section. The system is not quite ideally performed in predicting short term influence but it is still of great significance because it is likely to estimate the long term influence for a paper more accurately, but the ultimate influence determines the achievements of literature.

Comparing FIP with and without literature weights, we notice that both strategies have merits. When $\Psi(d)$ is not incorporated, the overall accuracy of FIP is slightly better because it predicts low citations very well and the lowly cited literature is the most major

Table 1: The performance of various prediction techniques for different feature combinations on the test set.

	1-Year FIP (Δt=1)			5-Year FIP (Δt=5)			10-Year FIP (Δt=10)		
Methods	FData	RData	Combined	FData	RData	Combined	FData	RData	Combined
kNN	0.515	0.311	0.593	0.681	0.268	0.734	0.649	0.161	0.767
LR	0.625	0.479	0.692	0.798	0.134	0.811	0.885	0.123	0.912
SVR	0.590	0.268	0.644	0.723	0.162	0.771	0.813	0.111	0.861
CART	0.679	0.441	**0.713**	0.797	0.203	0.834	0.852	0.128	0.905
GPR	0.601	0.349	0.668	0.823	0.153	**0.869**	0.894	0.130	**0.927**

Table 2: The performance of various prediction techniques for different feature combinations on the test set. "+" indicates the single feature group in isolation while "-" indicates the drop of the feature group from the full combination.

	1-Year FIP (Δt=1)					5-Year FIP (Δt=5)					10-Year FIP (Δt=10)				
Methods	kNN	LR	SVR	CART	GPR	kNN	LR	SVR	CART	GPR	kNN	LR	SVR	CART	GPR
+Content	0.061	0.097	0.101	0.104	0.110	0.065	0.100	0.107	0.105	0.103	0.114	0.152	0.137	0.174	0.168
+Author	0.563	0.586	0.582	0.603	0.589	0.601	0.619	0.611	0.627	0.631	0.616	0.637	0.623	0.652	0.659
+Venue	0.236	0.331	0.333	0.362	0.331	0.315	0.347	0.340	0.369	0.372	0.345	0.380	0.371	0.402	0.417
-Content	0.623	0.711	0.706	**0.727**	0.719	0.651	0.760	0.735	0.769	0.781	0.684	0.795	0.773	0.820	0.867
-Author	0.267	0.323	0.327	0.412	0.419	0.340	0.409	0.427	0.441	0.432	0.406	0.435	0.412	0.468	0.455
-Venue	0.571	0.583	0.588	0.606	0.597	0.605	0.617	0.609	0.632	0.625	0.621	0.650	0.628	0.667	0.672
Combined	0.593	0.692	0.644	0.713	0.668	0.734	0.811	0.771	0.834	**0.869**	0.767	0.912	0.861	0.905	**0.927**

part of the whole collection. When literature weights modeled, FIP biases to find the highly influential papers. We notice that in Figure 4.(2)-(3), the scattered circles in the high citation range are much less, and are much closer to the criterion line $y = x$ than those in Figure 4.(1). **Particular, few circles are in the upper triangle area (i.e., when $y > x$), which indicates few significant literature is missed.** However, for FIP with literature weights, the prediction effect in the low citation range is compromised. Although for lowly cited publications, the error of mis-prediction does not cost as much as those highly cited ones, it remains to be the insufficiency of FIP and can be further improved.

Different predictive models have different performances on these three individual tasks in our experiments. In general, non-linear regression achieves better performance. From Table 1, we notice that kNN has the worst performance. The result is as expected because kNN merely seeks the most similar neighbors and takes the neighbors' citation counts as the predictive influence while utilizes little information from the enormous training data. LR, by linear combination of all features, and CART and GPR by non-linear regressions have comparable performances and proves the generality of our extracted features. GPR is the best and CART, compared with GPR, has a little unstable performance. We also examine the growing factor in Table 1. Generally, the introduction of recent feature distributions benefit the performance of future influence prediction. Although the RData alone does not bring with excellent results, the combination of features from FData and from RData enhance the performance of what FData is capable of.

We then examine the different aspects of feature groups: paper content, author expertise and venue impact in Table 2. Author expertise is proved to be the most influential feature group in influence prediction, with the highest performance of R^2=0.659 in isolation and the lowest performance of R^2=0.419 for GPR when left out from full feature combination. It is understandable that authors are likely to cite papers written by reputable and influential authors. Venue impact is also influential. Papers from prestigious venues are likely to be highly cited. Unexpectedly, paper content is proved to have the least significance, with the average performance of R^2=0.130 in isolation. We assume (1) authors have biases to choose their bibliography: they sometimes merely consider author/venue reputation; (2) it seems that paper quality is represented by author/venue which create the paper. Influential authors or venues seem to overwhelm the impact of paper content itself; (3) it might also be due to the insufficient feature distilling for contents, e.g. using abstracts as approximation may not be enough for topic/diversity discovery.

Table 3: Feature analysis: R^2 result when with the pending feature ("+Add"), and result in R^2 when dropped from the all-features model ("-Drop").

	FData		RData	
Feature	+Add	−Drop	+Add	−Drop
Novelty	0.059	0.754	0.066	0.751
T.Rank	0.079	0.783	0.135	0.678
Diversity	0.157	0.661		
A.Rank	**0.593**	**0.406**	0.227	0.626
H-Index	0.244	0.611	0.186	0.663
Productivity	0.198	0.652	0.187	0.684
MPIA	0.585	0.419	**0.363**	**0.596**
TPIA	0.048	0.805	0.037	0.811
NOCA	0.056	0.794	0.158	0.643
Sociality	0.249	0.597	0.181	0.632
Authority	0.155	0.668	0.178	0.615
Versatility	0.160	0.649	0.139	0.665
Recency	0.101	0.738		
V.Rank	0.337	0.603	0.225	0.648
V.Centrality	0.049	0.793	0.067	0.776
MPIV	0.329	0.616	0.196	0.667
TPIV	0.023	0.815	0.021	0.823

We also conduct to a detailed experiment on all separate features in Table 3, and the visualization is presented in Figure 5. We mark the most prominent changes of performance in bold characters in Table 3. For the FData, the absence of *Author Rank*, *MPIA* and *Sociality* lead to unfavorable decrease; for the RData, the absence of *MPIA*, *Authority* and *Author Rank* results in similar effects. Hence, the performance is different for the same feature from RData and FData. For instance, the recent productivity is superior than the overall productivity according to the experiments. The visual example of Figure 5 better illustrates the comparisons between different features and their combinations. *TPIA* and *TPIV* are the least powerful prediction factors for both FData and RData, and the results confirm Figure 2.(7) and Figure 2.(15). Another interesting discovery is that *Novelty* and *Recent Novelty* lead to similar per-

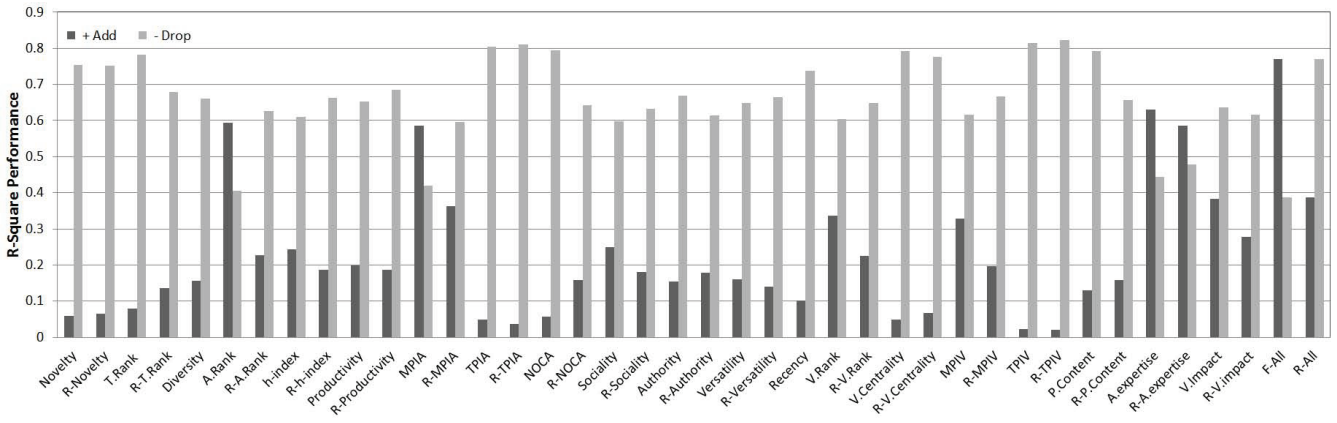


Figure 5: Performance comparison in R^2 for feature analysis. *"R-"* denotes the feature measured on the RData. *A.Rank/V.Rank* denotes author/venue rank, and *F-All/R-All* denotes all features measures on the FData or on the RData.

formances, perhaps due to the temporally recent bibliographies for reference papers. We also find graph-based *sociality* is better than simple *NOCA*.

4.5 Parameter Tuning

We have two free parameters in the temporal features. N is to control the size of RData and when N=0, the consideration of growing factor is off. γ controls the decay of the influence value after a given period of years. When γ=0, the decaying factor is off. We have tested several values, summarized in Table 4. The consideration of feature decay brings positive outcomes, but the penalization should better not be too much: γ=1 harms the performance. We choose N=5 in our experiments because a RData where N=1 hardly makes any difference. From our observations and experiments, the larger N is, the similar feature pattern it would be to the pattern on FData. Based on performance tuning on development set, we set *k*-NN as 5-NN empirically.

Table 4: Tuning temporal influence on growing factor/decaying factors. $N = 0$ means no growing factor of RData and $\gamma = 0$ means no decaying factor for future feature values.

Decay \ Grow	Off: N=0	N=1	N=5	N=10
Off: γ=0	0.799	0.812	0.833	0.808
γ=0.01	0.803	0.817	0.836	0.806
γ=0.1	0.747	0.765	0.791	0.752
γ=1	0.718	0.739	0.767	0.720

Strictly, Δt is not a parameter to tune but we can see a increasing trend for the performance as Δt enlarges. A possible explanation for such a increasing trend along with Δt is that for papers with certain features (such as high *author rank*, high *MPIA*, etc.) are predicted to have high influence. However, as Figure 2.(16) shows, citation accumulation takes time (about 10 to 20 years). The predicted citation counts will be overestimated for a short period of years: hence in Figure 4, many circles are below the line $y = x$.

4.6 Application: Bibliography Recommendation by Re-Ranking Mechanism

Future influence prediction not only estimates literature quality, but also can be used as auxiliary information in practical applications such as literature recommendation. Although our major focus is on citation estimation, we also implement the prototype of literature recommendation/search system proposed in [1] to prove the benefits of FIP. The system calculates the appropriate candidate reference papers for each article d and hence ranks the reference list by the ranking score, but it takes no consideration of future influence. We combine these estimated citation counts with the original ranking score to obtain the re-ranking score. In particular, the ranking score S_{FIP} for future influence is:

$$S_{FIP}(d) = \frac{log(1 + INF(d)) - log(1 + INF_{min})}{log(1 + INF_{max}) - log(1 + INF_{min})} \quad (17)$$

where $INF_{max/min}$ means the maximum/minimum predicted influence calculated. The score is scaled to [0, 1] so as to be comparable with the original ranking score $S_O(d)$ in [1]. Finally we model the re-ranking score S as the linear combination:

$$S(d) = \lambda \cdot S_O(d) + (1 - \lambda) \cdot S_{FIP}(d) \quad (18)$$

Table 5: Result comparison in MAP for original and combined ranking scores. λ is tuned at 0.3 from development set.

Method	Dev MAP	Test MAP
Original Score	12.51%	12.38%
Combined Score with FIP	15.66%	15.87%

As to evaluation, we train the model on the development set until it achieves the highest Mean Average Precision (MAP) [1]. The performance of bibliography recommendation is listed in Table 6, from where we notice that a 3.49% of improvement. This phenomenon does not necessarily indicate that authors intend to cite future influential papers, but can be explained by citation patterns: papers from reputable authors/venues are more likely to be chosen as bibliography and these papers are predicted to be higher influential, which is not a coincidence.

5. CONCLUSION AND FUTURE WORK

In this paper we propose a novel solution for the task of Future Influence Prediction (FIP), which identifies the potentially influential publications. Given a particular paper and its corresponding features relevant with citation patterns (such as paper content, author expertise and venue impact), FIP predicts its possible citation counts. We formally formulate FIP task as a learning problem utilizing several regression models, and evaluate the prediction performance by coefficient of determination (R^2). We also implement

a practical system of bibliography recommendation via re-ranking mechanism. The system outperform the baseline system, indicating papers with influential characteristics are likely to attract author attentions when they choose references.

From our experiments, we find that authors have biases in citing references. Author expertise and venue impact are the distinguishing factors for the consideration of bibliography, among which, *Author Rank*, *Maximum Past Influence of Authors* make paper influential. Content features are not predictive perhaps due to (1) citation bias, (2) paper quality is covered by authors/venues, or (3) insufficient content modeling. As FData and RData have different feature patterns, the combination of both result in better performance. In general, the prediction after a longer period can achieve the best accuracy (R^2=0.927 when $\Delta t = 10$). Currently, we consider a particular paper itself without considering any of its audience (citing papers). However, the impact of audience can also be modeled because once a paper is cited by an influential audience, it is likely to be influential as well. As considering the audience will result in a *multi-step influence diffusion* problem and increase the complexity in measurement. In this study, we do not consider the audience's influence when measuring the influence of the cited literature, while it can be further studied in the future.

6. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. The work was partially supported by the Natural Science Foundation of China (Grant No. 60933004, Grant No. 61073081) and the Open Fund of the State Key Laboratory of Virtual Reality Technology and Systems. Jie Tang was supported by the Natural Science Foundation of China (Grant No. 61073073), Chinese National Key Foundation Research (No. 60933013, No.61035004). Rui Yan was supported by the MediaTek fellowship.

7. REFERENCES

[1] S. Bethard and D. Jurafsky. Who should I cite: learning literature search models from citation behavior. In *Proceedings of CIKM*, CIKM'10, pages 609–618. ACM, 2010.

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

[3] L. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[4] 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.

[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] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of ICML'07*, pages 233–240, 2007.

[7] J. Dimitrov, S. Kaveri, and J. Bayry. Metrics: journal's impact factor skewed by a single paper. *Nature*, 466(7303):179–179, 2010.

[8] A. Fersht. The most influential journals: Impact Factor and Eigenfactor. *Proceedings of the National Academy of Sciences*, 106(17):6883, 2009.

[9] L. D. Fu and C. Aliferis. Models for predicting and explaining citation count of biomedical articles. In *AMIA Annual Symposium*, pages 222–226, 2008.

[10] E. Garfield. Impact factors, and why they won't go away. *Nature*, 411(6837):522–522, 2001.

[11] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *SIGKDD Explor. Newsl.*, 5:149–151, December 2003.

[12] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569, 2005.

[13] J. Hirsch. Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193, 2007.

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

[15] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of CIKM'10*, CIKM '10, pages 199–208, 2010.

[16] C. Lokker, K. McKibbon, R. McKinlay, N. Wilczynski, and R. 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.

[17] M. Lovaglia. Predicting citations to journal articles: The ideal number of references. *The American Sociologist*, 22(1):49–64, 1991.

[18] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceeding of SIGKDD*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[19] R. Picard and R. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:575–583, 1984.

[20] C. Rasmussen. Gaussian processes in machine learning. *Advanced Lectures on Machine Learning*, pages 63–71, 2004.

[21] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, pages 49–58, 2010.

[22] 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.

[23] K. Simons. The misused impact factor. *Science*, 322:165, 2008.

[24] R. Steel and J. Torrie. *Principles and procedures of statistics*, volume 633. McGraw-Hill New York, 1980.

[25] Y. Sun and C. Giles. Popularity weighted ranking for academic digital libraries. *Advances in Information Retrieval*, pages 605–612.

[26] Y. Sun, T. Wu, Z. Yin, H. Cheng, J. Han, X. Yin, and P. Zhao. Bibnetminer: mining bibliographic information networks. In *Proceedings of SIGMOD*, SIGMOD '08, pages 1341–1344, 2008.

[27] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of SIGKDD*, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM.

[28] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *Proceeding of CIKM*, CIKM '11, 2011.

[29] D. Zhou, X. Ji, H. Zha, and C. L. Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of CIKM'06*, pages 248–257, 2006.

[30] M. Zitt. Citing-side normalization of journal impact: A robust variant of the Audience Factor. *Journal of Informetrics*, 4(3):392–406, 2010.