High Impact Academic Paper Prediction Using Temporal and Topological Features

Feruz Davletov[*]
Dept. of Computer Science
Istanbul Sehir University
Istanbul, Turkey
feruzdavletov@
std.sehir.edu.tr

Ali Selman Aydin
Dept. of Computer Science
Istanbul Sehir University
Istanbul, Turkey
aliaydin@std.sehir.edu.tr

Ali Cakmak
Dept. of Computer Science
Istanbul Sehir University
Istanbul, Turkey
alicakmak@sehir.edu.tr

ABSTRACT

Predicting promising academic papers is useful for a variety of parties, including researchers, universities, scientific councils, and policymakers. Researchers may benefit from such data to narrow down their reading list and focus on what will be important, and policymakers may use predictions to infer rising fields for a more strategic distribution of resources. This paper proposes a novel technique to predict a paper's future impact (*i.e.,* number of citations) by using temporal and topological features derived from citation networks. We use a behavioral modeling approach in which the temporal change in the number of citations a paper gets is clustered, and new papers are evaluated accordingly. Then, within each cluster, we model the impact prediction as a regression problem where the objective is to predict the number of citations that a paper will get in the near or far future, given the early citation performance of the paper. The results of empirical evaluations on data from several well-known citation databases show that the proposed framework performs significantly better than the state of the art approaches.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Information Networks; H.3.7 [**Digital Libraries**]: Dissemination; I.2.6. [**Learning**]; I.5.3 [**Clustering**]

General Terms

Algorithms; Design; Experimentation; Verification.

Keywords

Citation count prediction; clustering; time series; regression; network analysis.

1. INTRODUCTION

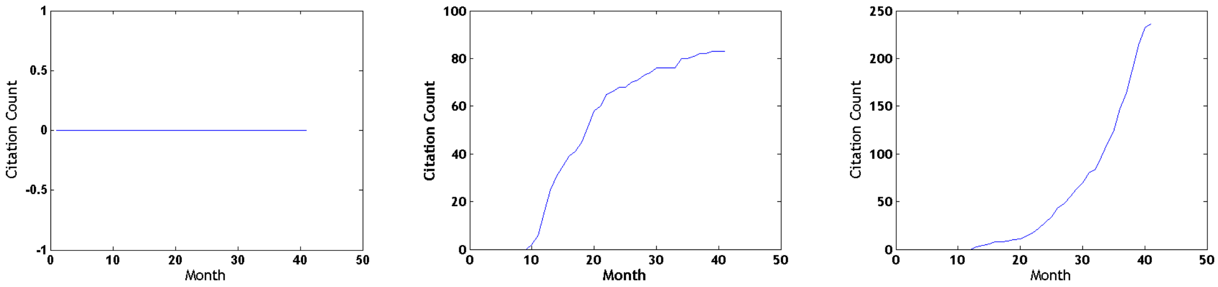
In today's academia, publish or perish policy results in an enormous body of publications. Hence, given the limited time, researchers have to be selective while putting a paper into their reading list, as well as prioritizing the articles in that list. Ideally, many would be more interested in reading papers that are likely to have high impact in their fields so that they can get ahead of their peers in contributing to an emerging field, and possibly become a leading figure in that area. However, it is almost impossible to decide whether a paper would really make a high impact ahead of time before reading it (even after reading it, it would be challenging to mark such papers). In order to tackle with reading list building and prioritization challenge, researchers practice some implicit rough filters, such as following top publication venues and prominent researchers constantly in their field. The implicit assumption is that high impact papers would be published in top venues and/or by prominent researchers. This assumption holds true in many cases, but even with such rough filters, the number of candidate papers to read may still be in the order of hundreds (if not in thousands) per year.

In addition, for policy makers and funding agencies, it may be essential to determine which papers will gain more attention. This is because such information may help them foresee which fields are more likely to be important in the future so that they can allocate resources more strategically. A direct approach for determining the future impact of a paper or field is to use expert knowledge. However, this is time consuming and human effort is not scalable to keep up with the current rate of academic production. Also, this method involves personal opinions of experts which may be subjective and prone to differ significantly. In addition, expert opinions are shown to be fallible for numerous times, as it is difficult to estimate the future impact of a paper just based on its content. One example of this is on the use of Neural Networks [16] for machine learning problems. In early 80s, Neural Networks were very popular in the field of machine learning, until they were replaced by linear classifiers, such as Support Vector Machines [23]. Until the popularity of the Neural Networks diminished, it may be highly likely that an expert would underrate the development in linear classifiers. Thus, the evaluation would be biased.

Hence, there is a need for automated tools that can accurately predict high impact papers (and, indirectly, the corresponding future hot research fields) shortly after their pub-

[*]This researcher has been supported by TUBITAK grant number 2215.

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2662066.



(a)

(b)

(c)

Figure 1: Citation behaviors of three papers extracted from Arxiv HEP-TH dataset

lication. In this paper, we propose an impact prediction framework for academic papers in which we model the behavior of the papers in terms of their citation performance during their initial lifetime. Given the citation behavior of a paper in the first few years after it gets published, our model assigns the paper to a cluster, hence captures its citation behavior. Our model then uses the paper's topological properties (such as various centrality measures within the citation network) to improve the prediction accuracy. The main novelty of our approach is the use of behavioral models that are defined by clustering the citation time series of the papers to analyze their citation patterns.

Measuring the impact of a publication is usually done in an appeal to popularity fashion, in which the number of citations a paper gets is used to quantify the attention that the paper gets; hence, the impact of the paper in the academic world [21]. In spite of the studies that point out the problems of the citation-based analysis of impact [13], the number of citations has been widely used as a measure of impact of publications or scientists [7, 9]. Hence, the impact of a paper in our approach is also based on the number of citations that a paper already got or will get. Previous research [2] shows that the number of citations in the initial years after the publication of a paper is a good indicator of that paper's citation performance in the long term. Hence, citation performance of a paper during its early life-time is a promising feature, and has been employed previously [14,17]. However, none of the previous research has focused on a paper's citation behavior which captures the patterns in citation count changes of a paper. More specifically, the citation behavior of a paper tells about whether the number of citations increases steadily, or it saturates after some time, or whether the paper seems to get no citations at all at the beginning, and its citation count explodes later, and so on. Figure 1 shows three different types of citation patterns, where x axis represents the time and y axis represents the cumulative number of citations. In Figure 1a, the number of citations forms a straight line, since the number of citations over time does not increase, and the paper ends up having zero citations at the end of the analysis period. Figures 1b and 1c show a more steady increase, followed by a saturation period. The difference between Figures 1b and 1c is in terms of the scale of the graph, *i.e.,* the number of citations that the papers get at the end of the analysis period: the paper in Figure 1b ends up having around 85 citations, while the paper in Figure 1c gets around 230 citations. These three patterns indicate two important parameters about the citation behavior of a paper: (i) the scale, which corresponds to the number of citations, and (ii) the temporal change with respect to the number of citations. While considering the current citation behavior of a paper, one may take into account these two parameters.

It is relatively straightforward to employ the number of citations as a predictor of future importance, when the problem is considered as a regression task. However, it is not as intuitive to incorporate the citation behavior into such a regression-based prediction scheme. For this problem, we propose to cluster papers based on their citation behaviors (*i.e.,* change of the number of citations over time), and assign a polynomial to each cluster for regression. Given a new paper, we assign the paper to a cluster by using its citation behavior in the initial stages after its publication, and perform citation prediction based on the polynomial associated with the paper's cluster.

In addition to the temporal citation patterns, topological properties of papers in the citation network are shown to be helpful in predicting future importance of a paper [17, 21]. Hence, we also use topological properties of a paper, such as betweenness centrality, closeness centrality, PageRank, and eigenvector centrality in the citation network for improving prediction performance of our model.

Experiments in two well-known datasets reveal that our model outperforms the state of the art by a significant margin. Also, we show that using behavioral models, rather than considering all papers to have the same citation behavior improves the prediction performance significantly. We also provide several sensitivity analyses on the parameters we use in our model.

Contributions. Our contributions in this paper are as follows:

- Clustering papers during the initial stage based on their citation behaviors.
- Creation of regression models specific to each citation behavior.
- Improvement of the prediction based on the distribution of topological measures.
- Experimental evaluation of the proposed scheme on real data from several citation databases.

Organization. The rest of the paper is organized as follows: In Section 2, we review the related work, and present the features and models that have been used for predicting a paper's importance. Then, in Section 3, we present the proposed scheme. We also present a complexity analysis in this section. In Section 4, various experiments that we performed will be presented to validate the proposed scheme. Several parameter sweeps will also be presented. Then, in Section 5, we conclude by summing up the overall findings and giving insights about future work.

2. RELATED WORK

Impact prediction for academic papers has been an active research area. Most of the literature uses features that are extracted from graph structure [21, 24]. [2] states the importance of the citations in the initial years after the publication of a paper, hence it is commonly used [5, 14, 17]. [3] models citation behaviors of researchers in computer science, and defines two publication popularity phases, namely, the population growth phase and the population decay phase. However, none of the papers mentioned here uses clustering in terms of citation behavior to have a better prediction. In our model, we exploit the differences between the citation behavior of papers, and cluster similar papers together for training regression models for each cluster, rather than training a single regression model for all papers.

In addition to number of citations, measures of network centrality, such as clustering coefficient, average shortest path length, and betweenness centrality are used [21]. The idea in [21] is that there is a pattern of topological features between the papers that get high number of citations. For testing their hypothesis, they analyzed the correlation between the number of citations in the future and the topological measures stated above. Similar to these work, our model uses topological features. However, the way our model uses topological features is to use it as a means of improvement, rather than using it as the main component used for prediction. Various topological features are also studied in [20].

In addition to temporal and topological features described above, a variety of contextual features are used. A reasonable indicator of a paper's impact is the previous works of the paper's authors. People in academia tend to cite papers that are published by *celebrities*, *i.e.,* people who are well-known in these fields. Hence, an author with high number of citations is more likely to get citations in his/her following papers. This phenomenon was taken into account in [5]. Some other works consider the content of the paper. [24] uses contextual features, in addition to a large pool of other features, to predict for citation number prediction in ArnetMiner dataset. [25] uses metadata information of papers, and they found out that text features significantly improve the prediction performance, compared to the baseline methods.

Models that are employed for the impact prediction problem are also worth mentioning. One approach is to consider citations in an information diffusion context, in which citations are spread like a disease. Preferential attachment [4] is another model that is used to study the problem, especially in the context of link prediction [12]. The concept of preferential attachment suggests that new nodes favor connections to existing nodes that are highly connected. There exist other models proposed for multidisciplinary networks that are based on structural holes, but these are outside the scope of our study.

The next section is devoted to the explanation of our proposed method.

3. METHODOLOGY

In this section, we present the temporal and topological features that are used to analyze the paper's current situation, and how these features are employed to predict the future characteristic of the paper. Then we present the regression model that performs the prediction, based on the extracted features.

3.1 Problem Description

Given paper's time series of citation counts and topological features; predict citation count in future. Formally we can define this problem as below.

Let G be directed citation graph of papers. Where nodes are papers and edges $p \rightarrow q$ means p cites q

Let p be the paper that we want to predict citation count.

Let C_t^p be citation count time series of a paper p where *t=0,1,2,3,...*

Let F^p be the topological feature set of a paper p where $F^p = \{f_1^p, f_2^p, f_3^p, f_4^p, ...\}$. An example of features might be $f_1^p = PageRank(p, G)$, $f_2^p = Closeness(p, G)$ and $f_3^p = Eigenvector(p, G)$.

Predict citation count c_{t^*} given (C_t^p, F^p, t^*) where t^* is time in future.

3.2 Time Series Classification

We consider the citation pattern of each paper as a time series, and investigate the relationship among different time series. We do this by defining paper types in the context of incoming citation performance. We first create snapshots of the network with a constant interval determined by considering the properties of the dataset, and in each snapshot, we compute the number of citations for each paper for six years after the publication of paper for training purposes. Then, we construct a distance matrix D between all the papers that are used in the training, which is defined as follows:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (X_i^k - X_j^k)^2}, \qquad (1)$$

where X_i and X_j are citation time series for paper i and paper j, and k is the snapshot id starting from the publication of the paper. Here, the sizes of both time series are assumed to be the same, hence Euclidean distance [10] alone is used for computing the overall distance. However, it may not be the case in real life scenarios: for each paper, there exist a time interval where it has no citations, since the papers that cite a paper will go through a peer-review process which usually takes months, if not years. Hence, one might want to consider the citation behavior only after the first citation (if exists) by considering the related timespan. To this end, one may extend the scheme above by employing Dynamic-Time Warping (DTW) [19]. DTW can be used to compare time series that vary in time. Hence, using DTW makes it possible to compare papers even their citation behavior has differences caused by different peer-review lengths among the papers that cite each paper.

After D matrix is calculated, we apply spectral clustering [18] to determine citation models. Spectral clustering involves the application of eigen decomposition over the dis-

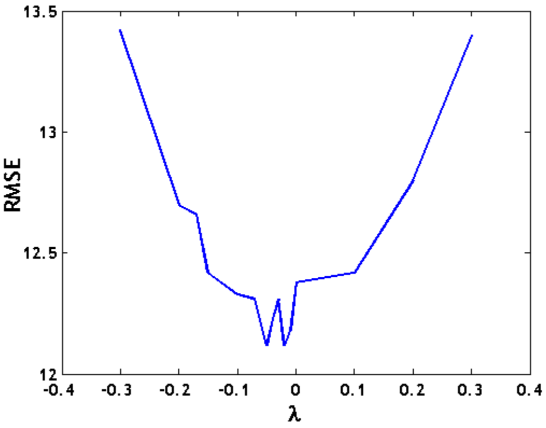


Figure 3: RMSE vs λ

tance (or similarity) matrix. The resulting eigenvectors are considered as cluster labels. We expect the citation behaviors to act as an indicator of citations in the future, given its citation performance in the first few(usually two or three in practice) years after the publication.

One important aspect that needs to be determined is the number of clusters, each of which correspond to a different behavioral model. If less number of clusters is used, the model results in underfitting, coming short in correctly modeling the citation behavior. On the other hand, having too many clusters is likely to overfit, which is another source of error. There are several options available for determining the number of clusters in given data. A rule of thumb is to use $\sqrt{\frac{n}{2}}$ as the number of clusters, where n is the number of data points [15]. There also exist different types of predictions that are based on the rate of change in cost function (*i.e.,* the elbow method), or information criterion. The idea in our model is to choose the number of clusters that minimizes the prediction error in our training dataset.

Figure 2 shows two of the clusters when the number of clusters are equal to 5. In each subfigure, the upper plot corresponds to individual citation behaviors, while the middle plot corresponds to the average of the cluster, and the bottom plot is the predicted polynomial for that cluster.

Note that the normalization plays an important role in the preprocessing stage prior to clustering, since our goal in the clustering is to capture the behavior, rather than to record the number of citations. We perform normalization by dividing the number of citations in each instant by the number of citations after time used for clustering; hence, the last element of a normalized sequence always becomes 1. The numbers of citations at the end of the timespan used for clustering are kept separate from the behavior, and used during the prediction stage.

When a new paper arrives, the paper is assigned to the cluster that minimizes the distance between its mean citation behavior of the cluster and the paper's normalized citation behavior, that is:

$$c_i = \arg\min_j \|X_i - \mu_j\|. \tag{2}$$

3.3 Topological Features

In addition to the clustering scheme that is described in the previous section, it is possible to use topological features as *improvement* terms. In this context, improvement is used to introducing information obtained from the topological properties of a paper with the intend of improving prediction accuracy. We either increase or decrease the predicted citations by a sum of topological features.

The way that we employ topological features is as follows: We first calculate topological features, such as betweenness centrality, closeness centrality, PageRank, and eigenvector centrality, and create a f^i vector out of these features. After f^i vector is calculated for each paper, we calculate the following score for each feature:

$$\alpha_j^i = \frac{f_j^i - \mu_j}{\sigma_j}, \tag{3}$$

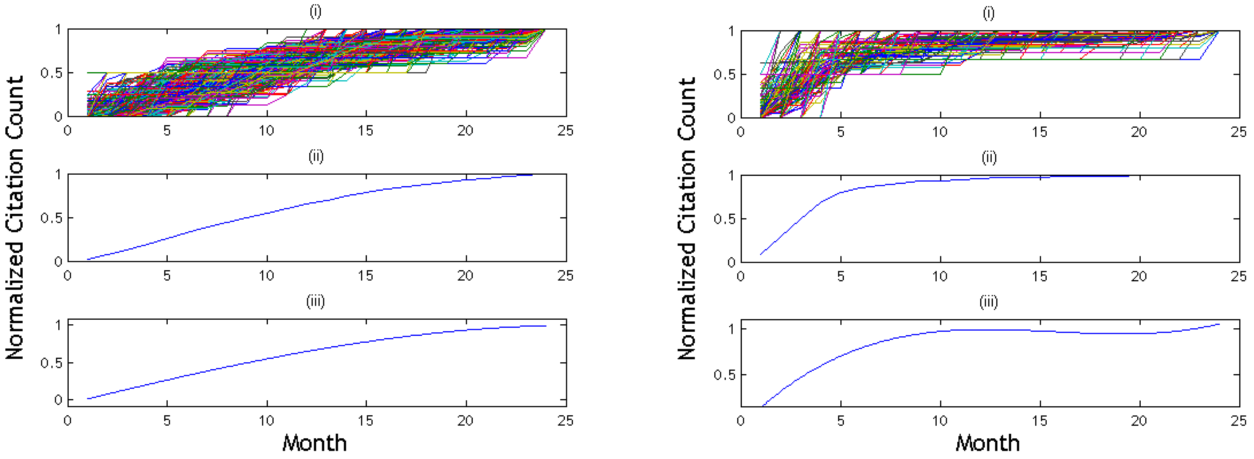
where μ_j is the mean and σ_j is the standard deviation of training samples for jth feature, and α_j^i is the multiplier of ith test sample for $j$the feature. Normalizing the features with respect to the distribution of these features along the training samples makes it possible for us to approach the problem in an anomaly detection sense. The anomalous samples, *i.e.,* the samples that deviate more from the mean, are considered as advantageous features in terms of topology. The sum of these values are then multiplied with a carefully tuned λ parameter to improve citation predictions. The tuning is performed based on the change in the prediction error in the training data(see Figure 3).

feature name	formula
Betweenness centrality	$\sum_{i\neq j\neq k} \frac{totalShortestPath_{jk}(i)}{totalShortestPath(i)}$
Closeness centrality	$closeness(i) = \sum_{i\neq j} \frac{1}{d_{ij}}$
PageRank	$PR(i) = \frac{1-d}{N} + d\sum_{j\in L(i)} \frac{PR(j)}{L(j)}$
Eigenvector centrality	$eigen(i) = \frac{1}{\lambda} \sum_{j\in M(i)} eigen(j)$

Table 1: List of features and formulas

3.4 Prediction

As discussed in the above sections, the evaluation of a paper is twofold: (i)a paper is first assigned to a cluster, and then (ii)the topological features are computed for the paper. The way we combine these two is as follows: Given a paper for which the prediction is to be performed, we first compare the citation behavior of the paper for a short timespan (*e.g.,* 2-3 years) with mean citation behavior of all clusters. As in Equation 2, the cluster that minimizes the distance is chosen as the cluster of the paper. Then, for the rest of the prediction timeline, the chosen cluster's polynomial is used for prediction (*polynomial regression*). Since the polynomial for each cluster is determined based on normalized citations, it only captures the behavior, and it should be amplified with the number of citations the paper has after the first window. Then, we introduce the effect of topological features, which we use as improvement terms. We use a carefully-tuned λ parameter to adjust the effect of α values on prediction. The resulting prediction, y', is calculated as follows:



(a)

(b)

Figure 2: Two of the clusters when the number of clusters is 5. Note that in each subfigure, (i) corresponds to individual citation behavior, while (ii) corresponds to the average of the cluster, and (iii) is the predicted polynomial for that cluster. The plots are based on citation data from Arxiv HEP-TH dataset

$$y'_i = y_i + \lambda \sum_j \alpha^i_j, \tag{4}$$

where y_i is the initial prediction without topological features. The λ value may be tailored for each feature, or may be global, which results in equal weighing of topological features. We used it as a global since all features are equally important.

3.5 Running Time Complexity

If the algebraic solution is used for spectral clustering, it requires computing the eigenvectors, which is $O(n^3)$, where n is the number of papers to be clustered, which is impractical when the number of samples grow. Hence, we use a greedy algorithm that approximates spectral clustering [6]. The algorithm we use is based on sparse similarity values, and the complexity of constructing the similarity matrix is $O(n^2)$, where n is the number of training data points (*i.e.,* papers). The complexity for spectral clustering is

$$(O(m^3) + O(mn) + O(nt)) \times O(m - k), \tag{5}$$

where m is the Arnoldi length in using the eigensolver, n is the number of data points, t is the number of desired nearest-neighbors in the algorithm ($t << n$), and k is the number of desired clusters [6]. m is often set to be several times larger than k [6]. The importance of training complexity further diminishes due to the fact that clustering is done once, and the constructed model is used for many times.

For testing, assigning a paper to a cluster is $O(kw)$, where w is the window size used to compare papers. Using the MATLAB implementation of [6] that runs on commercially-available hardware, training of $2,000$ samples can be performed in around 1 second. Similarly, testing can also be performed in real-time.

The next section is devoted to experiments that we performed to evaluate the performance of our proposed framework.

4. EXPERIMENTS

In this section, we first present a comparison of our method to the state of the art in various datasets. We then elaborate on our method by presenting sensitity analyses. Throughout the experiments, we used root-mean squared error (RMSE), the coefficient of determination (R^2), and correlation (r) as error metrics, depending on the work that we compare our results with. The reported results are averaged over 10 runs(*i.e.*, k-fold cross validation with *k=10*).

4.1 Datasets

Throughout the experiments, we used three datasets. The first dataset is Arxiv HEP-TH (high energy physics theory) dataset [8, 11], which contains citation information for $27,700$ papers that were published between 1992 and 2003, with all papers belonging to high energy physics theory. The second dataset we use is ArnetMiner dataset [22], which contains 1,511,035 papers and 2,084,019 citation relationships. Lastly, we used CiteSeerX [1] dataset which consists of over 2 million papers and 40 million citations. ArnetMiner and CiteSeerX datasets are used for comparison with the state of the art, while Arxiv is used to perform a detailed evaluation of our method, since it has less amount of missing values.

4.2 Comparison with the State of the Art

We compared our method to the two most recent works as representative studies [21] [5] of the state of the art from the literature. The first comparison was performed against [24] on ArnetMiner dataset [22]. In the experiments section of [24], the results were reported on various regression methods such as linear regression, regression trees, and SVR for $t = 1$, $t = 5$, and $t = 10$ years, where t is the time for which the number of citations is predicted. For the sake of brevity, we used the best result for each t. We used the citation information for first three years after the publication.

The error metric used by [24] is the coefficient of determination (R^2), which is defined as follows:

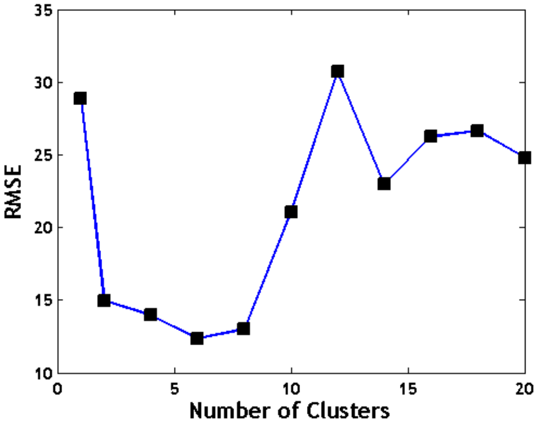


Figure 4: RMSE error vs number of clusters

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2}, \tag{6}$$

where $\overline{y}$ is the mean of the observed data (actual number citations at time t), y_i is each observed value and f_i is the each predicted value.

Table 2 indicates that our method outperforms the method used in [24], for $t = 1$, $t = 5$, and $t = 10$. The proportional difference between two methods is more apparent with $t = 1$ and $t = 10$, compared to $t = 5$. In order to show the positive effect of clustering, the results for regression where clustering is not performed are included as baseline. For the clustered case, the number of clusters is 7 and polynomial degree is 3. Window size is determined to be 2 years.

	Best result in [24]	Ours	Baseline
1 year	0.683	0.784	-9.9
5 years	0.752	0.800	-1.028
10 years	0.786	0.842	0.66

Table 2: R^2 comparison between our method and [24].

The second set of comparison were performed on CiteSeerX dataset against the results of [5]. They used the correlation coefficient r between the predicted and observed values as the error metric, where r is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{7}$$

where x is the predicted citations and y is the actual citations.

The comparison of results is presented in Table 3. Note that their method used the month data, in addition to the year data(which is not available in CiteSeer dataset) and predicted the number of citations for 4.5 years. Since month data is not available to our method for CiteSeer dataset, we provide comparison of results up to 4 years. Similar to ArnetMiner dataset, the number of clusters is 7 and polynomial degree is 3.

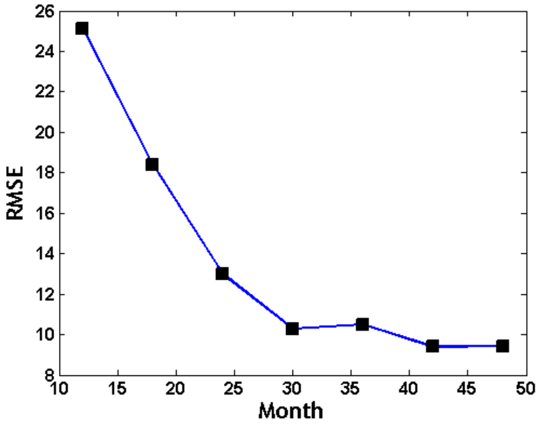


Figure 5: RMSE error vs window size used for classification

Table 3 indicates that our method outperforms [5] when the used window is up-to two years, while both methods provide similar performance when the length of the used window converges to the predicted window.

Used Window	[5]	Ours
1 year	0.76	0.95
2 years	0.92	0.96
3 years	0.97	0.96
4 years	0.99	0.97

Table 3: Correlation coefficient (r) comparison between our method and [5].

In the next subsections, we present the sensitivity analyses. The following results were obtained from experiments performed over Arxiv HEP-TH dataset.

4.3 Effect of Behavioral Clustering and Number of Clusters

The first study was performed to see the effect of clustering (if any), and how the number of clusters affects the accuracy of predictions. We first evaluated a non-clustered prediction, which we considered as baseline, and then increased the number of clusters gradually. The results of this study can be seen in Figure 4. Note that introducing clustering significantly decreases the prediction error, which shows the effectiveness of clustering. The baseline predictor, which does not use clustering, results in an RMSE error of 28.889, while the predictor with 6 clusters results in 12.365. However, it can be seen that increasing the number of clusters results in overfitting after some point, which increases the prediction error.

4.4 Changing Window Size

The second experiment was performed on the window size that we use to predict a paper's citation behavior. The window size is of great importance since using a very large window size would make the scheme impractical, requiring the citation records over long years, in which case, the citations of a paper is already stabilized. On the other hand, using a very small window size would be noise-prone, since it takes variable amount of time for papers to get their first citations. Figure 5 shows the average RMSE error for varying window sizes.

Looking at Figure 5, it can be seen that the error constantly decreases until a window size of 42 months, after which increasing the window size does not make any significant contribution. Taking the necessities of today's academic world into consideration, we concluded that a window size of 24 months is suitable(which is what we use in our experiments), since having a long window size will decrease the usefulness of the scheme, especially in fields like computer science.

4.5 Citation Lifetime

Another goal of our study is to identify a point for which the number of citations stabilize. This makes it possible to determine prediction interval which limits the range of prediction. In this experiment, we trained our model by using the citation records of 10 years. The prediction was also performed over a 10 years interval. Figures 6a and 6b show two of the three models that are observed in long term. The model that is not shown in the above figure corresponds to the papers that got no citations in the whole timespan. The first model, which is shown in Figure 6a, corresponds to papers that stabilize after some time. These papers possibly lost their connection with the state-of-the-art, hence they are less cited or no more cited at all. Obviously, this needs to be verified based on data, which is part of our future work. On the other hand, model in Figure 6b corresponds to the papers that sustain a steady growth in terms of the number of citations. These papers are more likely to be seminal papers, *i.e.,* papers that preserve their popularity even after years. The plots presented above show that these two types can be easily discriminated by using the data of 10 years, which allows for the possibility of long-term citation prediction. However, our analysis on this topic is limited due to constraints introduced by the dataset, which covers citation data from only 11 years, hence making a long-term analysis impossible. Using different datasets that cover a longer timespan for further analysis is part of our future work.

4.6 Error over Time

An other experiment was performed to see how prediction error changes over time. Figure 7 shows the prediction error over time.

Looking at Figure 7, it can be seen that the prediction error increases as the predicted interval becomes larger up until 40 months. After that point, the prediction error declines. This is possibly caused by degree of the regression. A quick solution to decrease error might be to increase the order of polynomial. However, we noticed that increasing the degree of polynomial results in higher prediction error, since a polynomial of higher degree is more likely to overfit.

5. CONCLUSIONS AND FUTURE WORK

We proposed an academic impact prediction framework based on the first years' citations and topological position of a paper. Our model uses time series approach to predict the number of citations, which is successfully employed before. Additionally, our model makes use of the citation behavior, *i.e.,* the pattern in the increase of the number of citations. In the training phase, the papers are clustered according to citation behaviors. Then, when a new paper arrives, it is assigned to a cluster and the prediction is performed accordingly. We also employ topological features of the paper, such as various centrality measures, to increase the prediction performance. Using topological features were also used before, however, we use them as a contributor to the prediction task, rather than using them as the main features. The experimental evaluation shows the high accuracy and robustness of our framework. Also, several parameter sweeps are performed to see the effect of parameters.

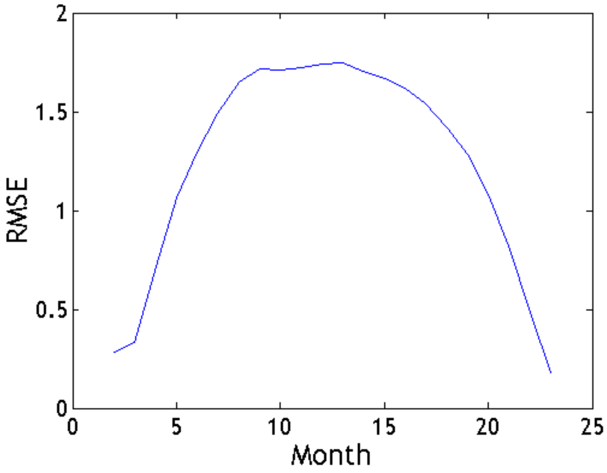


Figure 7: RMSE over time

As part of future work, we are planning to include contextual features described in the related work section to further improve our prediction framework. Also, we are planning to use a better weighing method for employing topological features to decrease prediction error. Finally, we plan to extend our prediction interval and prediction context further by using datasets that span a longer amount of time and a variety of disciplines.

6. ACKNOWLEDGMENTS

We would like to thank Prof. C. Lee Giles and his team from Pennsylvania State University for sharing the CiteSeerX data with us.

7. REFERENCES

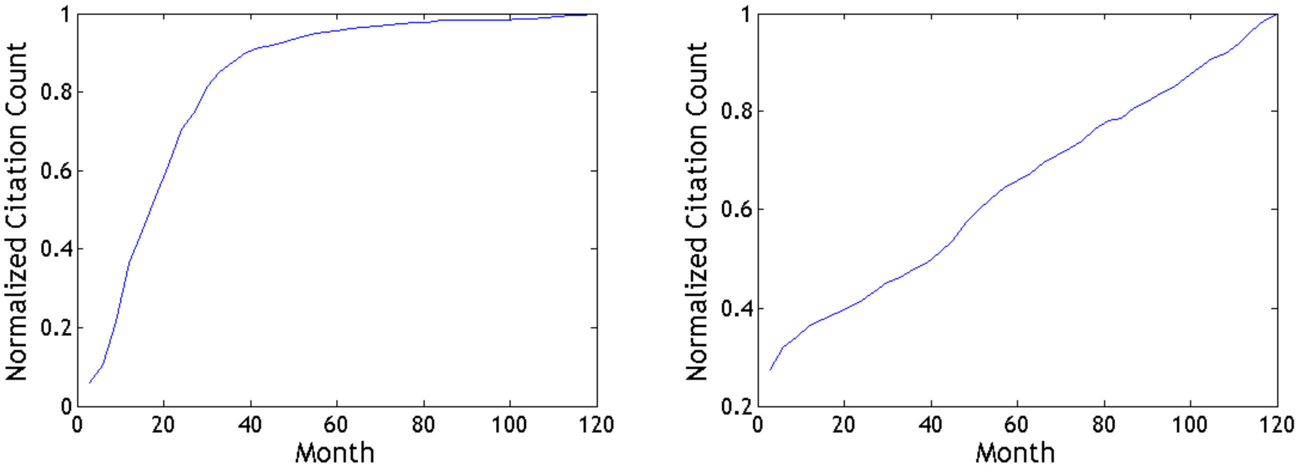
[1] CiteSeerX Digital Library. `http://citeseerx.ist.psu.edu/`, 2008. [Online; accessed 23-March-2014].

[2] J. Adams. Early citation counts correlate with accumulated impact. *Scientometrics*, 63(3):567–581, 2005.

[3] S. Bani-Ahmad and G. Ozsoyoglu. On popularity quality: growth and decay phases of publication popularities. In *Innovations in Information Technology, 2009. IIT'09. International Conference on*, pages 165–169. IEEE, 2009.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[5] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *String*



(a) (b)

Figure 6: Two models of long term citations

processing and information retrieval, pages 107–117. Springer, 2007.

[6] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang. *Parallel Spectral Clustering in Distributed Systems*, 2010.

[7] E. Garfield. The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

[8] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

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

[10] E. F. Krause. Taxicab geometry. *The Mathematics Teacher*, pages 695–706, 1973.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[12] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[13] M. H. MacRoberts and B. R. MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342–349, 1989.

[14] J. Manjunatha, K. Sivaramakrishnan, R. K. Pandey, and M. N. Murthy. Citation prediction using time series approach kdd cup 2003 (task 1). *ACM SIGKDD Explorations Newsletter*, 5(2):152–153, 2003.

[15] K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. 1980.

[16] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[17] D. McNamara, P. Wong, P. Christen, and K. S. Ng. Predicting high impact academic papers using citation network features. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 14–25. Springer, 2013.

[18] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

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

[21] N. Shibata, Y. Kajikawa, and K. Matsushima. Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, 58(6):872–882, 2007.

[22] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In *ICDM'07*, pages 292–301, 2007.

[23] V. Vapnik. *The nature of statistical learning theory*. springer, 2000.

[24] 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.

[25] D. Yogatama, M. Heilman, B. O'Connor, C. Dyer, B. R. Routledge, and N. A. Smith. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 594–604. Association for Computational Linguistics, 2011.