Predicting the Future Impact of Academic Publications*

Carolina Bento, Bruno Martins, and Pável Calado

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva
2744-016 Porto Salvo, Portugal
{carolina.bento,bruno.g.martins,pavel.calado}@ist.utl.pt

Abstract. Predicting the future impact of academic publications has many important applications. In this paper, we propose methods for predicting future article impact, leveraging digital libraries of academic publications containing citation information. Using a set of successive past impact scores, obtained through graph-ranking algorithms such as PageRank, we study the evolution of the publications in terms of their yearly impact scores, learning regression models to predict the future PageRank scores, or to predict the future number of downloads. Results obtained over a DBLP citation dataset, covering papers published up to the year of 2011, show that the impact predictions are highly accurate for all experimental setups. A model based on regression trees, using features relative to PageRank scores, PageRank change rates, author PageRank scores, and term occurrence frequencies in the abstracts and titles of the publications, computed over citation graphs from the three previous years, obtained the best results.

1 Introduction

Citations between articles published in academic digital libraries constitute a highly dynamic structure that is continuously changing, as new publications and new citations are added. Moreover, the graph of citations between publications can provide information for estimating the impact of particular publications, through algorithms such as PageRank, since highly cited papers are more likely to be influential and to have a high impact on their fields. Ranking papers according to their potential impact is thus a highly relevant problem, given that this can enable users to effectively retrieve relevant and important information from digital libraries. Having accurate prediction methods for estimating the impact of recently published papers is particularly important for researchers, since articles with high future impact ranks can be more attractive to read and should be presented first when searching for publications within digital libraries.

* This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT), through project grants with references UTA-EST/MAI/0006/2009 (REACTION) and PTDC/EIA-EIA/109840/2009 (SInteliGIS), as well as through PEst-OE/EEI/LA0021/2013 (INESC-ID plurianual funding).

L. Correia, L.P. Reis, and J. Cascalho (Eds.): EPIA 2013, LNAI 8154, pp. 366–377, 2013.

In this paper, we propose a framework that enables the prediction of the impact ranking of academic publications, based on their previous impact rankings. Specifically, given a series of time-ordered rankings of the nodes (i.e., the individual publications) from a citation graph, where each node is associated with its ranking score (e.g., the PageRank score) for each time-stamp, we propose a learning mechanism that enables the prediction of the node scores in future times. Through the formalism of regression trees, we propose to capitalize on existing trends through the changes in impact rankings between different snapshots of the citation graph, in order to accurately predict future PageRank scores and the future number of downloads. Moreover, we also experimented with the use of features derived from the textual abstracts and titles of the publications, in an attempt to capture trending topics. We evaluate the prediction quality through the correlation between the predicted ranked lists and the actual ranking lists, and through the error rates computed between the predictions and the correct results. The obtained results show that there is a significant correlation between the predicted ranked lists and the actual impact ranking lists, therefore revealing that this methodology is suitable for impact score prediction.

The remaining of this paper is organized as follows: Section 2 presents related work. Section 3 describes the proposed approaches for predicting the future impact of academic publications, detailing both the computation of impact estimates at a given time, and the machine learning models for making predictions. Section 4 presents the experimental validation of the proposed approaches, describing the evaluation protocol and the obtained results. Finally, Section 5 presents our conclusions and points directions for future work.

2 Related Work

The PageRank algorithm is a well-known method for ranking nodes in a graph according to their importance or prestige. It was originally proposed in the context of the Google search engine, and it has been extensively studied [9]. Many authors have proposed the application of PageRank, or of adaptations of this particular algorithm, to measure impact in scientific publication networks encoding citation and/or co-authorship information [2, 18].

An interesting approach that aims at approximating PageRank values without the need of performing the computations over the entire graph is that of Chien et al. [3]. The authors propose an efficient algorithm to incrementally compute approximations to PageRank scores, based on the evolution of the link structure of the Web graph. Davis and Dhillon proposed an algorithm that offers estimates of cumulative PageRank scores for Web communities [4]. In our work we also propose algorithms for estimating PageRank scores, but instead focusing on the prediction of future scores, based on previous PageRank computations.

Specifically focusing on PageRank computation over time-dynamic networks encoding citations, Radicchi et al. divided the entire data period into homogeneous intervals, containing equal numbers of citations, and then applied a PageRank-like algorithm to rank papers and authors within each time slice,

thereby enabling them to study how an author's influence changes over time [12]. Lerman et al. proposed a novel centrality metric for dynamic network analysis that is similar to PageRank, but exploiting the intuition that, in order for one node in a dynamic network to influence another over some period of time, there must exist a path that connects the source and destination nodes through intermediaries at different times [10]. The authors used their dynamic centrality metric to study citation information from articles uploaded to the theoretical high energy physics (hep-th) section of the arXiv preprints server, obtaining results contrasting to those reached by static network analysis methods.

Sayyadi and Getoor suggested a new measure for ranking scientific articles, based on future citations [13]. The authors presented the FutureRank model, based on publication time and author prestige, that predicts future citations. FutureRank implicitly takes time into account by partitioning data in the temporal dimension, using data in one period to predict a paper's ranking in the next period. The FutureRank scores are shown to correlate well with the paper's PageRank score computed on citation links that will appear in the future.

Kan and Thi have partially addressed the problem of predicting impact scores, by presenting a study that focused on Web page classification, based on URL features [6]. In their study, the authors also report on experiments concerning with predicting PageRank scores for graphs of hyperlinks between Web pages, using the extracted URL features and linear regression models.

The works that are perhaps more similar to ours are those of Vazirgiannis et al. [16] and of Voudigari et al. [17]. Vazirgiannis et al. presented an approach for predicting PageRank scores for Web pages, generating Markov Models from historical ranked lists and using them for making the predictions. Voudigari et al. extended this method, comparing models based on linear regression and high-order Markov models. Although both these works aimed at predicting PageRank for Web graphs, the authors evaluated their methods on co-authorship networks built from DBLP data. In this paper, we instead report on experiments made over citation networks built from DBLP data, aiming at the prediction of both PageRank scores and number of downloads for publications. We also relied on a highly-robust regression approach for learning to make the predictions, namely ensemble models based on Gradient Boosting Regression Tress (GBRT) [11].

3 Predicting the Future Impact of Publications

In this section, we present a learning method for predicting the future impact of academic publications, leveraging on patterns existing in the ranking evolution of the publications, and on the textual contents of the titles and abstracts. Given a set of successive snapshots for the graph of citations between publications, we generate, for each publication, a sequence of features that captures the trends of this publication through the previous snapshots. For each publication, we also generate features based on the publication date, and based on the words appearing on the title and abstract. We then use these features of previous snapshots as training data for a learning method, afterwards trying to predict

the future impact of particular publications, based on the previous snapshots. Figure 1 presents a general overview on the proposed methodology.

Let G_{t_i} be a snapshot of the citation graph, capturing citations between papers published before the timestamp t_i that is associated to the snapshot. Let $N_{t_i} = |G_{t_i}|$ be the number of publications at time t_i. In the case of citation networks for academic publications, we have that $N_{t_i} \leq N_{t_{i+1}}$. We also assume the existence of a function $rank(p, t_i)$ that provides an influence estimate for a publication $p \in G_{t_i}$, according to some criterion. In this paper, we used the original PageRank algorithm, over the citation graphs, to compute $rank(p, t_i)$, although other impact estimates could also have been used instead.

The original PageRank formulation states that a probability distribution over the nodes in a graph, encoding importance scores for each of the nodes, can be computed by leveraging links through the following equation, which represents a random walk with restarts in the graph:

$$Pr(p_i) = \frac{1-d}{N} + d \sum_{p_j \in I(p_i)} \frac{Pr(p_j)}{L(p_j)} \tag{1}$$

In the formula, $Pr(p_i)$ refers to the PageRank score of a node p_i, $I(p_i)$ is the set of nodes that link to p_i (i.e., the citations made to article p_i), $L(p_j)$ (i.e., the citations made in article p_j towards other articles) is the number of outbound links on node p_j, and N is the total number of nodes in the graph. The parameter d controls the random transitions to all nodes in the graph (i.e., the restarts), with a residual probability that is usually set to $d = 0.85$. In our experiments, the computation of PageRank relied on the implementation present in the WebGraph package[1], from the Laboratory of Web Algorithms of the University of Milan.

Let $x_p = (r_{p_1}, \ldots, r_{p_m})$ encode the *rank* values for a publication p at time points $t = (t_1, \ldots, t_m)$, and let $M = (N \times m)$ be a matrix storing all the observed *rank* values, so that each row corresponds to a publication and each column corresponds to a time point. Given these values, we wish to predict the *rank* value r_{p_*} for each publication at some time t_*, corresponding to a future time point. To do this, we propose to leverage on a regression model, using the k previous *rank* values as features, together with other features that capture (i) trends in the evolution of influence scores, and (ii) intrinsic properties of the publications, that can be used to group together similar sequences of impact scores (e.g., having the same authors or using the same textual terms).

Regarding the first group of features from the above enumeration, we propose to use the Rank Change Rate between consecutive snapshots of the citation graph, in order to capture trends in the evolution of the PageRank impact scores. The Rank Change Rate (Racer) between two instances t_{i-1} and t_i is given by the following equation:

$$\text{Racer}(p, t_i) = \frac{rank(p, t_i) - rank(p, t_i - 1)}{rank(p, t_i)} \tag{2}$$

[1] Available at `http://law.dsi.unimi.it/software.php#pagerank`

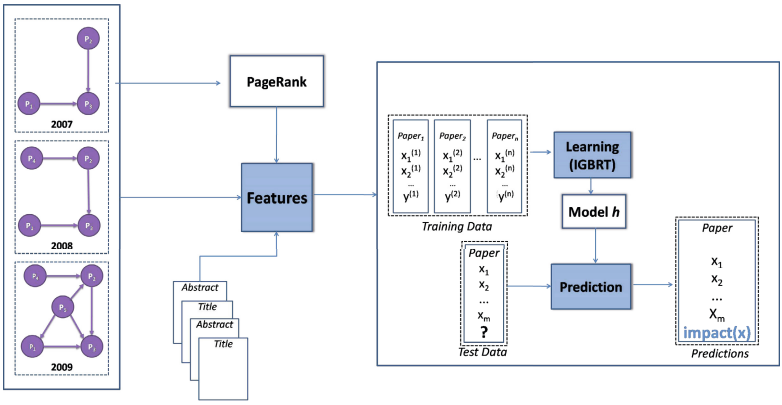

Fig. 1. The framework for predicting the future impact of academic publications

As for the second group of features, we used the assumption that publications from the same authors are likely to exhibit similar trends in the evolution of their impact scores. Thus, for each publication p, we computed the average and the maximum PageRank scores of all publications having an author in common with the set of authors of publication p.

To capture the intuition that impact metrics evolve differently for recent or older publications, we also used a feature that corresponds to the difference between the year for which we are making the predictions, and the year of publication. Finally, on what concerns the textual features, we used term frequency scores for the top 100 most frequent tokens in the abstracts and titles of the publications, not considering terms from a standard English stop-word list.

The above features were integrated into an ensemble regression approach based on the formalism of Gradient Boosting Regression Trees (GBRT).

We developed two different models, using the formulation of GBRT, for better understanding the impact of the combination of the aforementioned features, namely the (i) *Age Model*, and the (ii) *Text Model*. Both these regression models share a set of common features, which are:

- PageRank score of the publication in previous year(s) (Rank);
- Rank Change Rate (Racer) score towards the previous year(s);
- Average and maximum PageRank scores of all publications having an author in common with the set of authors of the publication (Auth);

In addition to the these common features, the *Age Model* also includes a feature that indicates the age of the publication, while the *Text Model* includes a feature indicating the age of the publication, as well as, features for the term frequency scores of the top 100 most frequent tokens in the abstracts and titles of the publications, up until the current date.

By experimenting with different groups of features, we can compare the impact of the amount of information that each model is given.

The actual learning approach that was used to build both the *Age* and *Text* models is named Initialized Gradient Boosting Regression Trees (IGBRT). This is an ensemble learning technique that, instead of being initialized with a constant function like in traditional Gradient Boosting approaches, it is initialized with the predictions obtained through the application of the Random Forests (RF) technique [1]. The IGBRT algorithm has, thus, a first step, in which the RF technique is applied, and then a final step, in which the traditional Gradient Boosting Regression Trees (GBRT) technique is applied [11]. The algorithm for GBRT evolved from the application of boosting methods to regression trees, through the idea that each step of the algorithm (i.e., the fitting of each of the regression trees for the final model) can involve the use of gradient descent to optimize any continuous, convex, and differentiable loss function (e.g., the squared loss). In the implementation that we have used, the individual regression trees that make up the boosting ensemble have a depth of 4, and they are are built through a version of the traditional CART algorithm, greedily building a regression tree that minimizes the squared-loss and that, at each split, uniformly samples k features and only evaluates those as candidates for splitting.

The general idea in GBRT is to compute a sequence of trees, where each successive tree is built over the prediction residuals of the preceding tree [5]. More specifically, the average y-value can be used as a first guess for predicting all observations (in the case of the IGBRT approach, the first guesses are instead given by the RF model). The residuals from the model are computed, and a regression tree is then fit to these residuals. The regression tree can then used to predict the residuals (i.e., in the first step, this means that a regression tree is fit to the difference between the observation and the average y-value, and the tree can then predict those differences), and these predictions are used to set the optimal step length (i.e., the weight of the current tree in the final model). The boosting regression model, consisting of the sum of all previous regression trees, is updated to reflect the current regression tree. The residuals are updated to reflect the changes in the boosting regression model, and a new tree is then fit to the new residuals, proceeding up to a given number of steps. Each term of the resulting regression model thus consists of a tree, and each tree fits the residuals of the prediction of all previous trees combined. The additive weighted expansions of trees can eventually produce an excellent fit of the predicted values to the observed values, even if the specific nature of the relationships between the predictor variables and the dependent variable of interest is very complex (i.e., nonlinear in nature).

4 Experimental Validation

To validate the proposed approach, we used a dataset[2] made available in the context of the ArnetMiner project, which encodes citation information between academic publications listed in the DBLP service [15]. The dataset contains information about 1,572,277 papers published until August 2011, having a total

[2] Available at `http://www.arnetminer.org/citation`.

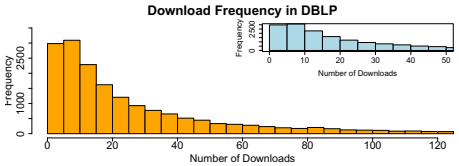

Fig. 2. Download frequency in the ACM DL for the papers in the dataset

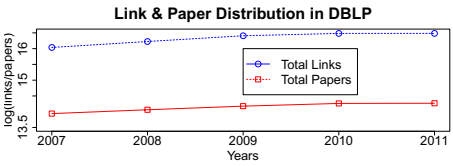

Fig. 3. Distribution for the links and the total number of papers

of 2,084,019 citation relationships. We also enriched the original ArnetMiner dataset with information about the number of downloads associated to each paper, by collecting this information in January 2012 from the ACM Digital Library service. We collected download information for a total of 17,973 papers. Figure 2 presents the distribution for the number of papers associated with a given number of downloads, while Figure 3 presents the number of papers and the number of citations, collected for each different year.

The dataset was split into five different groups, thus generating five graphs, corresponding to citations between papers published until the years of 2007, 2008, 2009, 2010 and 2011, respectively. PageRank scores were computed for each of these five graphs. We then evaluated our results on the tasks of (i) predicting the PageRank scores for the years of 2010 and 2011, based on the PageRank scores from the previous k years (with $1 \leq k \leq 3$), and (ii) predicting the number of downloads for papers in 2012, based on the PageRank scores from the most recent and from the previous k years. In terms of the prediction method, we used an implementation of the IGBRT algorithm that is provided by the RT-Rank project[3] [11]. Table 1 presents a brief statistical characterization of the DBLP dataset and of the five considered yearly graphs, showing the number of publications, citations, and authors per year.

To measure the quality of the results, we used Kendall's τ rank correlation coefficient [8], which consists of a non-parametric rank statistic that captures the strength of the association between two variables, with sample size n. Kendall's τ rank correlation coefficient is given by the formula below:

$$\tau = \frac{|\text{concordant pairs}| - |\text{discordant pairs}|}{\frac{1}{2}n(n-1)} \tag{3}$$

Kendall's τ rank correlation coefficient varies from $+1$ through -1, being $+1$ if the two rankings are equal and -1 if they are the exact opposite.

We also used Spearman's Rank Correlation Coefficient to measure the quality of the obtained results [14]. In this case, if X_1 and X_2 are two variables with corresponding ranks $x_{1,i}$ and $x_{2,i}$, and if n is the sample size, then their Spearman Rank Correlation Coefficient is given by the following equation:

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n(n^2-1)} \tag{4}$$

[3] Available at `https://sites.google.com/site/rtranking/`

As in the case of Kendall's τ, the Spearman correlation corresponds to a value between $+1$ and -1. Notice that Kendall's τ penalizes dislocations in the ranked lists independently of the distance of the dislocation, whereas Spearman's ρ does this through the square of the distance. Thus, Kendall's τ penalizes two independent swaps as much as two sequential swaps, while Spearman's ρ gives a stronger penalty to the latter than to the former. Both Spearman's ρ and Kendall's τ measure the quality of the ranked lists independently of the actual impact scores that are produced as estimates (i.e., only the relative ordering is taken into account).

In order to measure the accuracy of the prediction models, we used the normalized root-mean-squared error (NRMSE) metric between our predictions and the actual ranks, which is given by the following formula:

$$\text{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(x_{1,i}-x_{2,i})^2}{n}}}{x_{\max} - x_{\min}} \tag{5}$$

In the formula, $x_{\min}$ and $x_{\max}$ correspond, respectively, to the minimum and maximum values observed in the sample of objects being predicted, and n corresponds to the sample size. We also used the Mean Absolute Error (MAE) metric, which in turn corresponds to the formula bellow, where $x_{1,i}$ corresponds to the prediction and $x_{2,i}$ to the real value:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_{1,i} - x_{2,i}| \tag{6}$$

Tables 2 and 3 present the obtained results, respectively for the prediction of the PageRank scores for the years of 2010 and 2011, and for the prediction of the number of downloads for each paper in the dataset.

Considering the prediction of the PageRank scores for the year of 2010, both models (i.e., the *Age Model* and the *Text Model*) have provided very similar results. Both models are also improved if we consider more input information (i.e., comparing the three groups of features and also within the same groups,

Table 1. Statistical characterization of the DBLP dataset

	Publications	Citations	Authors	Papers with Downloads	Papers with Abstract	Average Terms Per Paper
Overall	1,572,277	2,084,019	601,339	17,973	529,498	104
2007	135,277	1,150,195	330,001	15,516	343,837	95
2008	146,714	1,611,761	385,783	17,188	419,747	98
2009	155,299	1,958,352	448,951	17,973	504,900	101
2010	129,173	2,082,864	469,719	17,973	529,201	103
2011	8,418	2,083,947	469,917	17,973	529,498	104

the quality of the results improves consistently). Only for the set of features that combines the PageRank score of one previous year (Rank $k = 1$) with its respective PageRank Change Rate (Racer) and the average and maximum PageRank score of the author (Auth), the *Age Model* is outperformed by the *Text Model*. Comparing the error rate for the same year, one can acknowledge that, for both models, as we add more information, the error rate increases, causing a deviation in the results. Nevertheless, for the first two groups of features, the *Text Model* has a lower error rate than the *Age Model*, while the opposite happens for the third group of features. Having computed the absolute error for all the groups of features in both models, the results show that, on average, the *Text Model* has always a lower absolute error than the *Age Model*.

For the year of 2011, as we add more information to the models, the *Text Model* outperforms the *Age Model*, as shown in the last two sets of features from the third group, i.e., PageRank score, Rank Change Rate score and Average and Maximum PageRank scores of all publications having an author in common with the set of authors of the publication. Also, in the scenario in which the models only have information about the immediately previous PageRank score, the *Age Model* is again outperformed by the *Text Model*. Nevertheless, when considering the error rate for both models for the year of 2011, the *Text Model* has an overall higher error rate than the *Age Model*, showing that, even though the quality of the predicted results is lower in the *Age Model*, the rankings are more accurate.

As occurred in the computation of the absolute error for the year 2010, in all the groups of features in both models, the results for the year of 2011 show that, on average, the *Text Model* has a lower NRMSE than the *Age Model*.

Table 2. Results for the prediction of future impact scores. Highlighted in bold are the best results for each metric, according to group of features, and for each year.

Model	Features	PageRank 2010			PageRank 2011		
		ρ	τ	NRMSE	ρ	τ	NRMSE
Age	Rank k = 1	0.97251	0.91640	**0.00032**	0.99299	0.98371	0.00011
	Rank k = 2	0.98365	0.93819	0.00062	**0.99991**	**0.99948**	**0.00001**
	Rank k = 3	**0.98907**	**0.95064**	0.00064	0.99990	0.99938	0.00048
	Racer + Rank k = 1	0.97245	0.91736	**0.00035**	0.99989	**0.99940**	0.00023
	Racer + Rank k = 2	0.98371	0.93876	0.00065	**0.99990**	0.99930	**0.00016**
	Racer + Rank k = 3	**0.98887**	**0.94937**	0.00066	0.99524	0.98662	0.00055
	Auth + Racer + Rank k = 1	0.96752	0.90985	**0.00054**	**0.99985**	**0.99945**	**0.00025**
	Auth + Racer + Rank k = 2	0.98405	0.93555	0.00083	0.99984	0.99934	0.00030
	Auth + Racer + Rank k = 3	**0.98925**	**0.94687**	0.00070	0.99380	0.98285	0.00053
Text	Rank k = 1	0.97087	0.91017	**0.00036**	0.99921	0.99797	**0.00025**
	Rank k = 2	0.98310	0.93104	0.00063	**0.99980**	**0.99924**	0.00045
	Rank k = 3	**0.98869**	**0.94515**	0.00063	0.99950	0.99834	0.00058
	Racer + Rank k = 1	0.97112	0.90989	**0.00055**	0.99943	0.99845	**0.00016**
	Racer + Rank k = 2	0.98320	0.93144	0.00067	**0.99973**	**0.99907**	0.00019
	Racer + Rank k = 3	**0.98880**	**0.94701**	0.00067	0.99941	0.99807	0.00064
	Auth + Racer + Rank k = 1	0.97052	0.99845	**0.00016**	0.99970	0.99906	**0.00025**
	Auth + Racer + Rank k = 2	0.98370	**0.99907**	0.00019	0.99986	0.99934	0.00028
	Auth + Racer + Rank k = 3	**0.98884**	0.99807	0.00064	**0.99988**	**0.99939**	0.00070

Regarding the prediction of download numbers, one can acknowledge that the *Text Model* shows evidence of better results. Moreover, in the *Age Model*, we can verify that adding information about PageRank Change Rates to the previous PageRank scores affects the results negatively, while combining previous PageRank scores with PageRank Change Rates, and average and maximum PageRank score of the author, provides better results, as well as, a lower error rate. From this fact, we can conclude that the *Age Model* provides a more accurate prediction as it includes more information, while the opposite happens in all groups of the *Text Model* (i.e., within the same group, as we add more information to the model, the quality of the results decreases, even though they are far better than the corresponding results in the *Age Model*).

We can also verify that the *Age Model*, for the groups of features that only include previous PageRank scores, and for the ones that combine previous PageRank scores with PageRank Change Rates and average and maximum PageRank score of the author, have a lower error rate than the corresponding groups in the *Text Model*. Even though with better overall results, the *Text Model* has a greater error rate than the *Age Model* for the prediction of download numbers. In what concerns the MAE, the results showed that, overall, the *Text Model* has a lower absolute error rate than the *Age Model*, in all groups except for the third.

From the results in Tables 2 and 3, we can see that predicting the number of downloads is a harder task than predicting the future PageRank scores. Also, predicting the future PageRank scores for 2011 turned out to be easier than making the same prediction for the year of 2010, which may be due to the combination of some aspects, namely the fact that we naturally took more papers into account while predicting the future PageRank scores for 2011 (i.e., we used

Table 3. Results for the prediction of future download numbers. Highlighted in bold are the best results for each metric, according to group of features, and for each year.

Model	Features	ρ	τ	NRMSE	MAE
Age	Rank k = 1	0.38648	0.27430	0.00806	31.78320
	Rank k = 2	0.42215	0.30015	0.00310	28.29388
	Rank k = 3	**0.43232**	**0.30810**	**0.00292**	**25.6090**
	Racer + Rank k = 1	**0.43966**	0.30766	**0.00767**	30.14214
	Racer + Rank k = 2	0.33702	0.47472	0.00784	29.38590
	Racer + Rank k = 3	0.33134	**0.46124**	0.00883	**27.01311**
	Auth + Racer + Rank k = 1	0.33776	0.25584	0.01547	37.21956
	Auth + Racer + Rank k = 2	0.53355	0.38949	0.00881	28.10313
	Auth + Racer + Rank k = 3	**0.54069**	**0.39625**	**0.00786**	**25.62784**
Text	Rank k = 1	0.52502	0.38370	**0.00912**	39.33091
	Rank k = 2	**0.52612**	**0.38496**	0.00913	39.30055
	Rank k = 3	0.50600	0.36748	0.00932	**36.39041**
	Racer + Rank k = 1	**0.53254**	**0.38880**	**0.00912**	39.33073
	Racer + Rank k = 2	0.52240	0.38230	0.00913	39.30457
	Racer + Rank k = 3	0.50874	0.37034	0.00932	**36.39390**
	Auth + Racer + Rank k = 1	**0.57098**	**0.42348**	**0.00912**	39.30848
	Auth + Racer + Rank k = 2	0.56513	0.41801	0.00913	39.29639
	Auth + Racer + Rank k = 3	0.56090	0.41486	0.00932	**36.36546**

more 20,768 papers than in 2010) providing, therefore, more information to the models, for training and for testing.

In sum, we have shown that the proposed framework based on ensemble regression models, offers accurate predictions, providing an effective mechanism to support the ranking of papers in academic digital libraries.

5 Conclusions and Future Work

In this paper, we proposed and evaluated methods for predicting the future impact of academic publications, based on ensemble regression models. Using a set of successive past top-k impact rankings, obtained through the PageRank graph-ranking algorithm, we studied the evolution of publications in terms of their impact trend sequences, effectively learning models to predict the future PageRank scores, or to predict the future number of downloads for the publications. Results obtained over a DBLP citation dataset, covering papers published in years up to 2011, show that the predictions are accurate for all experimental setups, with a model that uses features relative to PageRank scores, PageRank change rates, and author PageRank scores from the three previous impact rankings, alongside with the term frequency of the top 100 most frequent tokens in the abstracts and titles of the publications, obtaining the best results.

Despite the interesting results, there are also many ideas for future work. Our currently ongoing work is focusing on the application of the prediction mechanism to other impact metrics, perhaps better suited to academic citation networks. A particular example is the CiteRank method, which is essentially a modified version of PageRank that explicitly takes paper's age into account, in order to address the bias in PageRank towards older papers, which accumulate more citations [18]. Another interesting example of an impact metric for publications would be the Affinity Index Ranking mechanism proposed by Kaul et al., which models graphs as electrical circuits and tries to find the electrical potential of each node in order to estimate its importance [7]. We also plan on experimenting with the application of the proposed method in the context of networks encoding information from other domains, particularly on the case of online social networks (i.e., predicting the future impact of blog postings or twitter users) and location-based online social networks (i.e., predicting the future impact of spots and/or users in services such as FourSquare).

References

[1] Breiman, L.: Random Forests. Machine Learning 45(1) (2001)
[2] Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics 1(1) (2007)
[3] Chien, S., Dwork, C., Kumar, R., Simon, D.R., Sivakumar, D.: Link evolution: Analysis and algorithms. Internet Mathematics 1(3) (2003)
[4] Davis, J.V., Dhillon, I.S.: Estimating the global PageRank of web communities. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)

[5] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5) (2000)
[6] Kan, M.-Y., Thi, H.O.N.: Fast webpage classiffication using URL features. In: Proceedings of the ACM International Conference on Information and Knowledge Management (2005)
[7] Kaul, R., Yun, Y., Kim, S.-G.: Ranking billions of web pages using diodes. Communications of ACM 52(8) (2009)
[8] Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1/2) (1938)
[9] Langville, A., Meyer, C.D.: Survey: Deeper inside PageRank. Internet Mathematics 1(3) (2003)
[10] Lerman, K., Ghosh, R., Kang, J.H.: Centrality metric for dynamic networks. In: Proceedings of the Workshop on Mining and Learning with Graphs (2010)
[11] Mohan, A., Chen, Z., Weinberger, K.Q.: Web-search ranking with initialized gradient boosted regression trees. Journal of Machine Learning Research 14 (2011)
[12] Radicchi, F., Fortunato, S., Markines, B., Vespignani, A.: Diffusion of scientific credits and the ranking of scientists. Physical Review (2009)
[13] Sayyadi, H., Getoor, L.: Future rank: Ranking scientific articles by predicting their future PageRank. In: Proceedings of the SIAM International Conference on Data Mining (2009)
[14] Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology 15 (1904)
[15] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and mining of academic social networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)
[16] Vazirgiannis, M., Drosos, D., Senellart, P., Vlachou, A.: Web page rank prediction with Markov models. In: Proceedings of the International Conference on World Wide Web (2008)
[17] Voudigari, E., Pavlopoulos, J., Vazirgiannis, M.: A framework for web Page Rank prediction. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) EANN/AIAI 2011, Part II. IFIP AICT, vol. 364, pp. 240–249. Springer, Heidelberg (2011)
[18] Walker, D., Xie, H., Yan, K.-K., Maslov, S.: Ranking scientific publications using a simple model of network traffic. Technical Report CoRR, abs/physics/0612122 (2006)