Estimating Number of Citations Using Author Reputation

Carlos Castillo, Debora Donato, and Aristides Gionis

Yahoo! Research Barcelona
C/Ocata 1, 08003 Barcelona
Catalunya, SPAIN

Abstract. We study the problem of predicting the popularity of items in a dynamic environment in which authors post continuously new items and provide feedback on existing items. This problem can be applied to predict popularity of blog posts, rank photographs in a photo-sharing system, or predict the citations of a scientific article using author information and monitoring the items of interest for a short period of time after their creation. As a case study, we show how to estimate the number of citations for an academic paper using information about past articles written by the same author(s) of the paper. If we use only the citation information over a short period of time, we obtain a predicted value that has a correlation of $r = 0.57$ with the actual value. This is our baseline prediction. Our best-performing system can improve that prediction by adding features extracted from the past publishing history of its authors, increasing the correlation between the actual and the predicted values to $r = 0.81$.

1 INTRODUCTION

Editors in publishing houses (as well as producers for record labels and other industries) face often the following problem: given a work, or a promise of a work, what is a good method to predict if this work is going to be successful? Answering this question can be very useful in order to decide, for instance, whether to buy the rights over the work, or to pay in advance to the authors. The editor's prediction on the success of the work can, in principle, depend on the past publishing history or credentials of the author, and on the estimated quality of the item that is being examined. Of course, the estimation can be quite inaccurate, as the actual success of an item depends on many elements, including complex interactions among its audience plus external factors that cannot be determined in advance.

We are interested in the problem of estimating the *success* of a given item, understood as the impact in its community. In the case of books, for instance, success can be measured in terms of book sales. In the case of scholarly articles, success is typically measured as a function of the number of citations an article receives over time.

In this paper, we deal with the citation prediction task in the context of a large set of academic articles. Our main questions are:

- Can we characterize the evolution of the citations of a paper over time?
- Can we predict the number of citations of a paper, given information about its authors?
- Can we improve such a prediction, if we know the number of citations a paper has received over a short timespan?

The method we describe on this paper receives as input an article and the past publication history of the authors of that article. The output is an estimation of how many citations the article will accumulate over its first few years. Such prediction can be further improved over time as the system receives information about how many citations the article received over the first few months after its publication.

The next section relates our work with previous papers on this problem. In Section 3 we describe the dataset we are using, in Section 4 we describe the features we extract for the prediction task, and in Section 5 we discuss the experimental results we obtained. The last section outlines our main conclusions and describes future work.

2 Related work

The 2003 KDD Cup [7] included a citation prediction task resembling the one we undertake on this paper. The citation prediction task included estimating the change in the number of citations of papers between two different periods of time. Participants received data about the citation graphs and the contents of a set of about 30,000 papers from the e-print arXiv[1]. The training data covered a 3-months period (February to April 2003) and the testing data was the next 3-months period (May to July 2003). In contrast, in the current paper we do not use content attributes from the papers and the time period covered by the prediction task is in the order of years, not months.

The problem of predicting the ranking of scientists was studied recently by Feitelson and Yovel [5]. They show that a multiplicative process gives a good approximation of the number of citations that authors in a certain position in the ranking receive. In their paper, Feitelson and Yovel want to estimate the rank of each author in the list ordered by citations, not the citation counts. The main idea is that authors will move up in the rank until the rate of change of their citations is equal to the rate of change of the citations of the authors in similar positions in the ranking. In contrast, in our work we focus on the number of citations of particular papers (not authors), and mostly in the absolute number of citations, not on the ranking.

Popescul and Ungar [14] use machine learning to try to predict specific citations among papers (e.g.: if paper p_1 is going to cite paper p_2 or not), using features such as the authors, citation information, and the venues where papers appear. A related problem, predicting co-authorship relationships between authors, is studied by Liben-Nowell and Kleinberg [11]. In this research, we estimate aggregate counts in the citation network, not specific links.

[1] http://arxiv.org/

The Web has provided an enormous amount of data about dynamic networks, including general Web pages, blogs, Wikis, and other information systems. In general the dynamics of blogs [1, 9, 12], Wikipedia [3], authorship networks [11] and other networks [10] have attracted considerable attention in the last years. In the specific case of blogs, Fujimura et al. [6] observe that the number of in-links to individual blog entries is very small in general, and some time is needed to acquire them. To overcome these problems and be able to rank blog postings, they propose the *EigenRumor* algorithm, in which a variation of the hubs and authorities algorithms [8] is applied to authors and blog postings. In the case of Web links, Baeza et al. [2] observed that PageRank is biased towards older pages (as they have had more time to accumulate citations. Cho et al. [4] propose to reduce this bias by considering a different quality metric: a weighted sum of the derivative of PageRank over time and the actual PageRank value.

Recent results on predicting popularity in social networks point out that when users influence each other, the ranking of items may be less predictable than when they judge independently. Salganik et al. [15] experiment with an artificial cultural market showing that, while high-quality items rarely have low rankings and low-quality items rarely have high rankings, basically any ranking outcome is possible.

3 Dataset

CiteSeer[2] is a search engine for academic papers and citations. It has been in operation since 1997 and currently indexes over 750,000 bibliographic records. CiteSeer data is available through an Open Archives Initiative (OAI) interface that allows users to download records from the dataset. The set of records we used covers 581,866 papers published from 1995 to 2003, both years inclusive.

Given that we are interested in creating a model for the authors, we kept only papers for which at least 1 of the authors had 3 papers or more in the dataset. In this way, we obtained 519,542 papers, which is about 89% of the original set. The large temporal window available allow us to use part of the data available to build the reputation of each author. In particular, we select 1,500 papers created during 4 months in 1999 (i.e., right in the center of the temporal window) and use all the past data (papers, authors and citations) to extract the features related to the authors of each paper at the moment of its writing. We use the remaining 4.5 years in the future to monitor the popularity growth and to test our predictions.

Next we looked at the number of citations that papers received. We focused on two particular moments in time: first, 6 months after the publication of each paper. On average, papers in our dataset had 2 citations at that time. Second, we looked at 30 months (4.5 years) after publication. On average, papers in our dataset had 5.9 citations after that time. Figure 1 summarizes the *fraction* of citations papers receive between 6 months and 30 months. Overall, papers seem

[2] http://citeseer.ist.psu.edu/

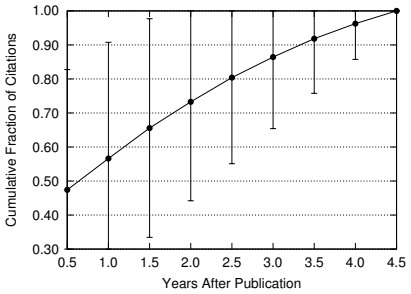

Fig. 1. Average profile of the number of citations over time in the `citeseer` dataset, taking as reference (=1.00) the events on the first 4.5 years. Error bars represent one standard deviation.

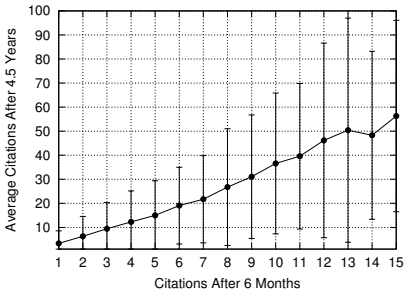

Fig. 2. Citations at the end of the first year versus average citations at the end of 5 years. Error bars represent one standard deviation. Std/avg ranges from 1.4 on the left to 0.7 on the right.

to accumulate citations steadily over their first few years. On average, a paper in our dataset receives roughly half of its citations on the first 6 months, and 57% of them on the first year. However, the variance on these statistics is very large. The error bars in the figure represent one standard deviation, and show that different papers accumulate citations following very different profiles.

The number of citations after 6 months and the number of citations after 30 months are correlated, and the correlation coefficient is about 0.57. This means that on average a paper that receives many citations shortly after it is published, will receive many citations over the following years. In Figure 2, we plot the number of citations after 6 months and the average number of citations after 30 months, including error bars. We can see that the correlation is not strong enough to make an accurate prediction. For instance, for papers with 10 citations in the first 6 months, one standard deviation means somewhere between 10 to 60 citations over 4.5 years, which is a rather large interval. Not surprisingly, the correlation improves as more data points are included, as we show later in Table 1. However, the goal of our task is to be able to estimate the number of citations of a paper shortly after it is published.

One remark about quality and citations is in order. The number of citations is not a perfect metric for the quality of a paper. The reasons behind the impact of a paper can vary, and it is not always true that quality is the key factor for the number of citations of a paper. Moreover, it is difficult to define an objective measure of quality itself. It is evident that surveys, methodological papers, or just papers addressing "hot topics", or in fields shared by large communities, are more likely to be read and cited than other papers, all other things being equal.

4 Features

Notation. We consider a graph $G = (V_a \cup V_p, E_a \cup E_c)$ that summarizes all the authorship and citation information available. The vertices of this graph

are composed of V_a, which is the set of vertices representing authors and V_p, which is the set representing papers. The set of edges include $E_a \subseteq V_a \times V_p$, that captures the authoring relationship, so that $(a, p) \in E_a$ if author a has co-authored paper p. A paper can have more than one author and we denote by $k_p \triangleq |\{a/(a, p) \in E_a\}|$ the number of authors of a paper. The edges in graph G also include $E_c \subseteq V_p \times V_p$, that captures the citation relationship, so that $(p_1, p_2) \in E_c$ if paper p_1 cites paper p_2.

In any bibliography dataset, authors have a double role: from one side, they deliver original content and produces new items, from the other side, they provide an implicit evaluation of other authors. The two types of edges in E_a and E_c capture the authoring and citation relationships respectively. These are denoted in the framework of Fujimura and Tanimoto [6] *information provisioning* and *information evaluation*, respectively.

Each paper $p \in V_p$ has also a timestamp associated to its creation: time(p). It is assumed that all citations go from a newer paper to an older paper. In the graph G, we define the number of citations of a paper at time t, $\mathrm{C}_t(p)$ as:

$$\mathrm{C}_t(p) \triangleq |\{p'/(p', p) \in E_c \wedge \mathrm{time}(p') < t\}|$$

that is, $\mathrm{C}_t(p)$ is the number of papers citing p that were published in the first t units of time after p. We then extend this notation by defining the number of citations of an author a before time t as:

$$\mathrm{C}_t(a) \triangleq |\{p'/(p', p) \in E_c \wedge (a, p) \in E_a \wedge \mathrm{time}(p') < t\}| \ .$$

We are interested in determining whether the number of citations of a paper after a long time period, can be approximated by a function of some *a priori* features related to the authors of the paper, and/or to the number of citations of the paper after a shorter time period. More specifically, we use three different types of features: (1) *a priori* author-based features, (2) *a priori* link-based features, and (3) *a posteriori* features.

***A priori* author-based features** try to capture how well previous papers from the same authors have performed in the past. At time t, the past publication history of a given author a can be expressed in terms of:
(i) Total number of citations received $\mathrm{C}_t(a)$: the global number of citations received by the author i from all the papers published before time t.
(ii) Total number of papers (co)authored $\mathrm{M}_t(a)$: the total number of papers published by the author a before time t

$$\mathrm{M}_t(a) \triangleq |\{p/(a, p) \in E_a \wedge \mathrm{time}(p) < t\}|.$$

(iii) Total number of coauthors $\mathrm{A}_t(a)$: for papers published before time t

$$\mathrm{A}_t(a) \triangleq |\{a'/(a', p) \in E_a \wedge (a, p) \in E_a \wedge \mathrm{time}(p) < t \wedge a' \neq a\}|$$

Given that one paper can have multiple authors, we aggregate the values that capture the history of each of those authors. A detailed explanation on how

we aggregate these values is given in Appendix A. In total we obtain 12 *a priori* author-based features.

***A priori* link-based features** try to capture the intuition that good authors are probably aware of the best previous articles written in a certain field, and hence they tend to cite the most relevant of them. Mutual reinforcement characterizes the relation between citing and cited papers; and this relationship also translates to an implied relationship among citing and cited authors. The intuition is that authors cited by good authors should have a higher probability to be cited, and also that good authors usually cite significant papers. This type of implicit endorsement provided by links is the basis of link-based ranking algorithms like PageRank [13] and HITS [8].

If two papers p_1 and p_2 written by different authors a_1 and a_2 respectively cite each other; that is, $(a_1, p_1) \in E_a$, $(a_2, p_2) \in E_a$ and $(p_1, p_2) \in E_c$; we can infer an implicit relationship between authors a_1 and a_2. In the *EigenRumor* algorithm introduced by Fujimura and Tanimoto [6], the relationships implied by both provisioning and evaluation of information are used to address the problem of correctly ranking items produced by sources that have been proved to be authoritative, even if the items themselves have not still collected a high number of inlinks.

The exact implementation of EigenRumor we use is explained in the Appendix. Basically, we compute 7 EigenRumor-based features related to the hub and authority score of the authors of each paper p.

***A posteriori* features** simply monitor the evolution of the number of citations of a paper at the end of a few time intervals that are much shorter than the target time for the observation. We consider the number of citations that each paper receives in the first 6 monehts and in the first 12 months after its publication.

5 Experimental Results

Our goal is to produce an approximation of the number of citations of a paper p at time T, $\hat{\mathrm{C}}_T(p) \approx \mathrm{C}_T(p)$ using all the information available for items created before time $T' < T$. A first metric for the quality of such approximation is the correlation coefficient between the variables $\hat{\mathrm{C}}_T(p)$ and $\mathrm{C}_T(p)$.

The correlation coefficient weights papers equally independent on their number of citations. This may be a disadvantage as there are some applications (such as search engines) in which it is more important to produce an accurate prediction for highly-cited papers than for the rest. For this reason, we also consider the following metric: we say that a paper is *successful* if it is among the top 10% of the papers published at the same time t (remember that t represents actually a certain period of time, e.g. one month). Next we evaluate $\hat{\mathrm{C}}_T(p)$ by measuring how accurately it can predict the "success" of a paper; for this classification task, we use the F-measure, defined as $\frac{2 \times precision \times recall}{precision + recall}$.

We used the freely-available machine-learning package Weka [16]. In particular we used the Weka implementation of linear regression (for prediction of the

number of citations) and C4.5 decision trees (for the prediction of success). In the case of decision trees, we applied an asymmetrical cost matrix to increase the recall, by making a misclassification of successful as unsuccessful 1.5 times more costly than a misclassification of unsuccessful as successful.

Tables 1 and 2 show the experimental results obtained over the 1,500 papers extracted as described in Section 3. The r and F values reported in this section are the average after 10-fold cross validation (in which 9/10 of the data are used to learn a model which is then tested in the remaining 1/10 of the data).

Table 1. Experimental results, using only *a posteriori* citation information. r is the correlation of the predicted value with the number of citations after 4.5 years. F is the F-Measure in the task of predicting "success" (defined as being in top 10% in citations).

A posteriori citations	Predicting Citations r	Predicting Success F
6 months	0.57	0.15
1.0 year	0.76	0.54
1.5 years	0.87	0.63
2.0 years	0.92	0.71
2.5 years	0.95	0.76
3.0 years	0.97	0.86
3.5 years	0.99	0.91
4.0 years	0.99	0.95

Table 2. Experimental results, using *a priori* and *a posteriori* features.

A priori features	*A posteriori* features: First 6 months r	First 6 months F	First 12 months r	First 12 months F
None	0.57	0.15	0.76	0.54
Author-based	0.78	0.47	0.84	0.54
Hubs/Auth	0.69	0.39	0.80	0.54
Host	0.62	0.46	0.77	0.57
EigenRumor	0.74	**0.55**	0.83	**0.64**
ALL	**0.81**	**0.55**	**0.86**	0.62

From Table 2 we can observe that by using *a priori* author information we obtains a clear improvement in the prediction, given that the correlation coefficient r of the predicted value goes from 0.57 to 0.81 in the prediction that uses the first 6 months of citations.

In the task of predicting success using 6 months of *a posteriori* data, the F-Measure increases significantly, from 0.15 to 0.55. The value F is a bit hard to interpret, but in this case, an F value of 0.55 reflects that about 57% of the

top-10% papers are detected, with about 5% false positives. Remember that we are predicting the impact of a paper after 30 months using 6 months of data.

6 Conclusions and Future Work

Our main conclusion is that, in the context of academic papers, information about the authors of a paper may help in predicting the number of citations it will receive in the future, even if we do not take into account other factors such as, for instance, the venue where the paper was published.

In the course of our experiments, we observed that *a priori* information about authors degrades quickly. When the features describing the reputation of an author are calculated at a certain time, and re-used without taking into account the last papers the author has published, the predictions tend to be much less accurate.

Wether the results shown on this paper are extrapolable to other communities or not remains to be seen. We have attempted the same prediction task over data from `Flickr`.[3] In Flickr, each "author" is a photographer, each "paper" is a photography, and "citations" are votes and/or comments a photographer places over somebody else's photography. This allows us to define information provisioning and information evaluation matrices similar to the ones we use for calculating the attributes we use in this prediction task.

We have used the same algorithms described in this paper, but so far we have not been able of improving the quality of a baseline prediction (using only *a posteriori* information) using *a priori* attributes. Of course, in Flickr the data is much more noisy and sparse, posting a photo is easier than publishing a paper, and in general the dynamics may be different from the dataset we have studied on this paper. However, one objective of our research is to devise algorithms for predicting the popularity of items in web communities, and we turn our future work there.

Acknowledgments: we thank Paul Chirita for helping us frame this problem at an early stage. We also thank Carles Rius from DAMA-UPC, who provided us a pre-processed version of CiteSeer data.

References

1. E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.
2. R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web structure, dynamics and page quality. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, 2002.
3. L. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal evolution of the wikigraph. In *Proceedings of Web Intelligence*, 2006.

[3] http://www.flickr.com/

4. J. Cho, S. Roy, and R. E. Adams. Page quality: in search of an unbiased web ranking. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005.
5. D. G. Feitelson and U. Yovel. Predictive ranking of computer scientists using citeseer data. *Journal of Documentation*, 60(1):44–61, 2004.
6. K. Fujimura and N. Tanimoto. The eigenrumor algorithm for calculating contributions in cyberspace communities. *Trusting Agents for Trusting Electronic Societies*, 2005.
7. J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *SIGKDD Explor. Newsl.*, 5(2):149–151, 2003.
8. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
9. R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, December 2004.
10. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.
11. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, 2003.
12. Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, 2006.
13. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
14. A. Popescul and L. H. Ungar. Statistical relational learning for link prediction. In *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003.
15. M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.
16. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

APPENDIX A.

For repeatability of our experiments, we include here the specific list of features we used for the prediction task.

***A priori* author-based features**. Remembering that k_p is the total number of the authors of the paper p, let $t = time(p)$, $\mathrm{C}_t(a)$, $\mathrm{M}_t(a)$, $\mathrm{A}_t(a)$ the global number of citations, papers and coauthors of the author a at time t.

We computed for each paper p the following features at time t :

- Features based on the number of citations $\mathrm{C}_t(a)$
  1. Sum of all citations collected by all the authors: $\sum_a^{k_p} \mathrm{C}_t(a)$
  2. Average citations per author: $\frac{\sum_a^{k_p} \mathrm{C}_t(a)}{k_p}$

3. Maximum number of citations: $max_a \, \mathrm{C}_t(a)$
4. Sum of all citations collected by all the authors per paper: $\sum_a^{k_p} \frac{\mathrm{C}_t(a)}{\mathrm{M}_t(a)}$
5. Average citations per author per paper: $\frac{\sum_a^{k_p} \frac{\mathrm{C}_t(a)}{\mathrm{M}_t(a)}}{k_p}$
6. Maximum number of citations per paper: $max_a \frac{\mathrm{C}_t(a)}{\mathrm{M}_t(a)}$

– Features based on the number of papers $\mathrm{M}_t(a)$

7. Sum of all papers published by all the authors: $\sum_a^{k_p} \mathrm{M}_t(a)$
8. Average number of papers per author: $\frac{\sum_a^{k_p} \mathrm{M}_t(a)}{k_p}$
9. Maximum number of papers: $max_a \, \mathrm{M}_t(a)$

– Features based on the number of coauthors $\mathrm{A}_t(a)$

10. Sum of all coauthors of each authors: $\sum_a^{k_p} \mathrm{A}_t(a)$
11. Average number of coauthors per author: $\frac{\sum_a^{k_p} \mathrm{A}_t(a)}{k_p}$
12. Maximum number of coauthors: $max_a \, \mathrm{A}_t(a)$

***A priori* link-based features**. Following [6], consider:

– the *provisioning matrix* $P_{a,t}$ is the matrix induced by the authoring relationship $E_{a,t}$, defined as:

$$E_{a,t} \subseteq V_a \times V_p \triangleq \{(a, p*) \in E_a \wedge \mathrm{time}(p*) < t\}$$

– the *evaluation matrix* $P_{e,t}$ is the matrix induced by the evaluation relationship $E_{e,t}$, defined as:

$$E_{e,t} \subseteq V_a \times V_p \triangleq \{(e, \hat{p}) : (e, p*) \in E_{a,t} \wedge (p*, \hat{p}) \in E_c \wedge \mathrm{time}(p*) < t\}$$

We compute the score $R(p)$ of the paper p and the authority $A_{t'}(a)$ and hub $H_{t'}(a)$ of the author a, where:

$$R = \alpha P_{a,t}^T A_t + (1 - \alpha) P_{e,t}^T H_t$$

$$A_t = P_{a,t} R \qquad H_t = P_{e,t} R$$

Aggregating the authority and hub scores for all the coauthors of each paper, we obtain 7 features:

– Relevance of the paper

1. EigenRumor $R(p)$;

– Authority features

2. Sum of the authority scores of all the coauthors: $\sum_a^{k_p} A_t(a)$
3. Average authority per author: $\frac{\sum_a^{k_p} A_t(a)}{k_p}$
4. Maximum authority: $max_a A_t(a)$

– Hub features

5. Sum of the hub scores of all the coauthors: $\sum_a^{k_p} H_t(a)$
6. Average hub per author: $\frac{\sum_a^{k_p} H_t(a)}{k_p}$
7. Maximum hub: $max_a H_t(a)$