Predicting Citation Counts Using Text and Graph Mining

Avishay Livne
Computer Science and Engineering
University of Michigan, Ann Arbor
avishay@umich.edu

Eytan Adar
Computer Science and Engineering, School of Information
University of Michigan, Ann Arbor
eadar@umich.edu

Jaime Teevan
Microsoft Research
teevan@microsoft.com

Susan Dumais
Microsoft Research
sdumais@microsoft.com

Abstract

As the volume of scientific literature grows faster it becomes more difficult for researchers to identify promising papers that are likely to become influential in their field. We study the problem of predicting future citation counts of papers given information available at the time of publication (five years forward in our pilot study). We apply machine learning techniques on a dataset of millions of academic papers from several research domains to identify predictive features including venue reputation, authors and institutions, citation networks and content measures. We identify how these features are differentially predictive in various domains and identify possible reasons where citation behaviors might lead to these differences.

Keywords: influence prediction, bibliometrics, citation analysis, network mining, text mining

Introduction

The task of citation count prediction has gained a lot of attention in recent years as the usage of these counts for assessment of scholarly impact became more pronounced. For example the usage of citation counts in Google Scholar's ranking algorithm affects dramatically the exposure papers receive and as a consequence strengthens the Matthew Effect (Beel & Gipp, 2009). Additionally, as the volume of published works increases it becomes more difficult for researchers to keep up with the state of the art in their field by identifying key works as soon as they are published.

In this paper we tackle the task of citation count prediction using existing and new features. We apply the technique to multiple domains, identifying differences both in the ability to predict citation counts as well as the nature of features that contribute to the prediction. For example we find the rich get richer phenomenon of famous authors attracting more citations is more apparent in Biology and Medicine compared with other domains. Additionally, while the popularity of a paper's references is predictive of the paper's success in most domains it this is clearly not the case in Engineering and Physics. Finally, unlike most previous studies that evaluated their models using cross validation on a cohort of papers published in the same year our evaluation is done in a more realistic set up where the model is used to predict citations 5 years out (using data from 2000 and 2005 to predict out to 2005 and 2010). When cross validation is used to explain variation it is possible the model captures dynamics that are only true for the cohort of papers used.

Related Work

Early work on citation count prediction focused on a limited set of features and applied simple models such as linear regression and decision trees on relatively small datasets. For example, (Callaham, Wears, & Weber, 2002) studied 204 publications and using decision trees they were able to explain 0.14 in the variation of citation counts 3.5 years after publication, with the journal's impact factor being the most predictive feature. (Kulkarni, Busse, & Shams, 2007) used linear regression to study 328 articles published in 2000 and reported R^2 of 0.2 in predicting citation counts 5 years ahead. (Castillo, Donato, & Gionis, 2007) used linear regression and decision trees to predict citation counts 4.5 years ahead. They found that future citation counts were highly correlated with the citation counts accumulated within the first year after publication and that by adding features describing author's reputation they were able to improve their predictions. Two recent works (Fu & Aliferis, 2008; Ibáñez, Larrañaga, & Bielza, 2009) considered the usage of content to improve prediction, i.e. by identifying keywords in the text that are

associated with high citations. (Didegah & Thelwall, 2012) studied several features and found venue prestige to be the strongest feature, followed by the number of citations attracted by the references of a paper. (Yan, Huang, Tang, Zhang, & Li, 2012) studied features covering venue prestige, content novelty and diversity and authors' influence and activity. They also accounted for temporal dynamics by taking a recent version of each feature calculated on a limited time window. (Shi, Leskovec, & McFarland, 2010) analyzed properties of the citation networks projected by the references of a paper and found some patterns are more common in the projected networks of highly cited papers.

Data

In this work we experiment with a uniquely large and diverse dataset, extracted from Microsoft Academic Search. The dataset consists of 38 million papers and 19 million authors, and covers 15 academic domains. Due to the size of the corpus, the metadata associated with each publication (including author, venue, references, and citations) was automatically extracted. Because of variation in coverage of different domains, we focus most of our analysis on papers from 2000 to 2005 from seven key domains. This yields a corpus of 12.7 million papers and 3 million authors, published at over 17,000 venues (see Table 1 for a breakdown).

Table 1
Domain Statistics

Domain	#Venues	#Papers 2000	#Papers 2005	Authors per paper 2000	Authors per paper 2005
CS	4,851	59,116	110,506	2.43	2.75
Biology	2,082	59,395	93,792	3.58	4.04
Chemistry	811	26,496	50,381	3.56	3.99
Medicine	5,524	125,113	214,854	3.52	3.67
Engineering	2,589	43,440	77,664	3.20	3.53
Mathematics	581	11,057	17,317	1.75	1.90
Physics	688	25,393	42,955	4.41	5.05

Methods

Features

We consider five groups of features, which we refer to as *Authors, Institutions, Venue, References Network* and *Content Similarity*. The first three—*Authors, Institutions* and *Venue*—describe the reputation of the paper's venue, of its authors and of its author's institutions. We start by calculating the following features for each venue, author and institution in the dataset: log of the sum of citation counts of papers published by the entity, log mean citations over papers published by the entity and log max citations, e.g. the citation count of the most cited work by the entity. We also calculate the *h*-index and *g*-index of these entities. The *h*-index is defined as the largest *h* such that at least *h* papers by the entity received at least *h* citations (Hirsch, 2005). The *g*-index is defined is the largest *g* such that the top *g* papers by the entity received together at least g^2 citations (Egghe, 2006). For each paper we aggregate the features of the entities (authors, institutions and venue) with which it is associated. For authors' features we take the mean over the paper's authors and so we do for institutions' features (we give equal weight to all institutions even if more authors are affiliated with one institution over another).
When assigning the features of entities to papers we calculate two versions for each feature; one is calculated over all the years strictly before the year in which the paper was published; and one is calculated based on data extracted in the one year preceding the paper's publication. Note that our granularity is of complete years so we will use data available by 1999 (inclusive) if a paper was published 2000 regardless of the month of publication.

A second group of features, based on the *References Network* (i.e., citations), attempt to measure how interdisciplinary the topic of the paper is as well as how interesting it is to the scientific community. We calculate the *h*-index, *g* -index, log mean citations and the log of the median of their

citations. We also measure the six network features described in (Shi et al., 2010) based on the network induced by projecting a paper's citations into the network, extending the network to include papers citing and cited by the paper's citations. The six metrics include graph density, clustering coefficient, connectivity, etc. to include the outgoing and incoming links of the references (considering only papers published before the paper in hand was published). As before, we consider two temporal versions for each feature—one calculated over all the years prior to publication and one using citations originated at papers published one year before the paper in hand.

The last group— *Content Similarity* —attempts so measure how hot is the topic of a paper using text analysis. We start by calculating two language models (LM) to extract popular terms. The first LM—denoted $LM_{published}$—is computed using the titles and abstracts of all the papers published in the same domain and year as the paper. The second LM—denoted $LM_{referenced}$ —is computed using the text snippets surrounding references made by papers published in the same domain and year as the paper. The weight of each term is the log of the number of its occurrences. We consider unigrams and bigrams after lowercasing the raw texts removing non-alphanumeric characters. We also ignore terms appearing in less than ten documents and stop words. Finally we represent each paper as a bag-of-words using its title and abstract and calculate its similarity to the two LMs.

Prediction Task and Model

For training our prediction model, we compute the training features based on papers published in 2000 and predict their citation counts as of 2005. For evaluating this learned model, we use papers published in 2005 and examine the extent to which our predictions of citations in 2010 correspond to actual citations in 2010. We use a support vector regression machine (SVR) model because they are well suited for handling large number of features. We use scikit-learn's implementation of SVR (Pedregosa et al., 2011) with the default parameters—RBF kernel of 3rd degree and penalty parameter of C=1.

Results

Table 2
Accuracy of Prediction (R^2) for Different Feature Groups

Features	Biology	Chemistry	Medicine	CS	Mathematics	Engineering	Physics
Authors	0.19	0.14	0.18	0.14	0.02	<0.01	0.02
Institutions	0.11	0.13	0.12	0.09	0.04	<0.01	<0.01
Venue	<0.01	0.05	0.10	0.14	0.09	**0.16**	**0.07**
References	**0.29**	**0.27**	**0.24**	**0.19**	**0.18**	0.07	0.04
Content	0.03	0.05	0.05	<0.01	<0.01	<0.01	<0.01
All Features	0.35	0.33	0.39	0.30	0.22	0.23	0.17

Table 2 presents the R^2 of our predictions with actual citations (five years from the date of publication) broken down by groups of features and each domain. All the values were found to be statistically significant ($p < 0.01$) due to the large size of the dataset. The first part of the part of the table shows the performance of each features group independently (the best group for each domain is highlighted). The bottom line lists the performance of the model combining all the features.

The results show several interesting general patterns. First we note a great variability in performance across domains – while our models can predict citation count well in some domains (e.g., 0.39 in Medicine and 0.35 in Biology) they have more limited success in others (e.g., 0.17 in Physics). Note that previous studies focused on predicting citation counts for papers from these relatively easy domains. Second, we note the set of predictive features differ across domains. Reference features are the most valuable at predicting citation counts in most domains (except Physics and Engineering), while the Content features rarely provide useful predictive information. This result suggests that the interest of a research community may be better captured using graph mining techniques rather than text mining techniques, although other text analysis methods that those consider here may achieve better results.

The performance of individual features can provide insights to dynamics and social phenomena within each domain. We found the recent *h*-index of the references to be one of the most predictive features in CS, Biology, Chemistry, Medicine and Math. One possible interpretation is that this feature is a

proxy for the interest of the scientific community in the *area* the paper studies, therefore papers that score high in this metric are likely to be of interest to the community. The recent citation count of venues was found to be a prominent feature in CS, Medicine, Engineering and Physics, a result that goes in line with the findings of several studies that the prestige of a venue contributes to the paper's success. Another prominent feature in the Biology, Chemistry and Medicine domains was the network constraint of a paper calculated on its references-projected graph. However the relation of this feature with citation counts is not linear but as described in (Shi et al., 2010)—high impact and low impact papers have low network constraint while medium impact papers have high network constraint. As network constraint measures how interleaved the interaction of a node is in a single group of interconnected neighbors, a possible interpretation of this finding is that papers citing works from different contexts are likely to be ignored yet in rare cases such papers bridge between domains or subdomains becoming highly influential.

Conclusions

We found a great variability in the performance across domains – while in CS, Biology, Chemistry and Medicine our models perform well in Engineering, Math and Physics we observed a very limited success, which could explain the lack of works on this topic applied on these fields. More important, the differences in the relative importance of different features provide insight to the differences in dynamics across domains. For example the rich-get-richer phenomenon where the works of famous authors attract more attention is clearly more pronounced in Biology and Medicine compared with Engineering, Math or Physics. The *references* features we employ capture even more subtle patterns such as the one reflected in the network constraint feature of mainstream papers achieving moderate success compared with a few papers that bridge across different communities.

As each domain encompasses many sub-communities with different dynamics we plan to extend this work by looking into these subdomains. We also plan to explore new features that will utilize signals not captured by the features used in this study. We suspect the Content features failed to capture any meaningful signal because of the coarse representation of content that is of interest to the community. It is possible we could improve this by identifying sub-communities and their relative size. Finally we plan to investigate whether similar techniques could be used to predict the scholarly impact of higher level entities (e.g., researchers and universities).

References

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (An empirical study). *Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on* (pp. 439–446).

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA : the journal of the American Medical Association*, *287*(21), 2847–50.

Castillo, C., Donato, D., & Gionis, A. (2007). Estimating number of citations using author reputation. *String processing and information retrieval*, 1–10.

Didegah, F., & Thelwall, M. (2012). Determinants of Research Citation Impact in Nanoscience and Nanotechnology. *Journal of the American Society for Information Science and Technology*, 1–14.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Fu, L. D., & Aliferis, C. (2008). Models for predicting and explaining citation count of biomedical articles. *AMIA Annual Symposium Proceedings*, 222–6.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*, *102*(46), 16569.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics (Oxford, England)*, *25*(24), 3303–9.

Kulkarni, A. V, Busse, J. W., & Shams, I. (2007). Characteristics associated with citation rate of the medical literature. (P. Bacchetti, Ed.)*PloS one*, *2*(5), e403.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, *12*, 2825–2830.

Shi, X., Leskovec, J., & McFarland, D. a. (2010). Citing for high impact. *JCDL '10*, 49.

Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*.