Prediction of Citation Counts of Journal Articles

Adam BACHO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
adam.bacho@gmail.com

Nowadays there are dozens of new scientific articles released each day. On one hand this gives authors or researchers many more opportunities than they had ever before, but on the other hand, it becomes harder and harder for them to distinguish articles of high quality for their further work. One of the top characteristics determining the quality of an article are citation counts, i.e., how many times was the given article cited by other articles [1].

Thus, in this work we designed a predictive model based on article and journal characteristics which tend to be good predictors of the citation counts. In the first place we had to choose the dataset and then we pre-processed it. From several available options we chose PubMed Central dataset in which more than 900,000 of open access articles are stored.

After the cleaning process we obtained 9,582 articles with 2005 as their year of publication containing all of information necessary for building our predictive model. All the features that we retrieved from the articles originally formatted as XML files are shown in Table 1. The reason we chose articles from the year 2005 is that we want to try and predict citation counts exactly N years (in this case eight) after the publication of the article. We included number of citations an article received one, two and five years after its publication in our model, as we expect its better performance with them.

We also included some novel features that were not examined in the previous works, the most prominent of them being *Eigenfactor Score* that can be obtained from ISI Web of Knowledge. There are some advantages of Eigenfactor when compared with better known Journal Impact Factor:

- Eigenfactor is not influenced by journal self-citation
- highly cited journals will influence the citation network more than less cited ones

* Supervisor: Róbert Móro, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 3–4.

Firstly, we performed univariate and multivariate linear regression just with features listed in Table 1 on the articles published in 2005.

While in univariate model just one feature (special char in the title) was marked as statistically insignificant (p-value = 0.75), in multivariate model there were six statistically insignificant features (special char in the title, abstract chars, abstract words, keywords, citations in release year and one year after publication). The most significant features were pages, authors, citations after two and five years after publication (all with p-value < 0.001). The only feature that had positive regression coefficient more than 1 was citations five years after the publication of the article. It is obvious, since the final citation counts were counted in year 2013, thus just 3 years later. Finally, we expect even better results after adding features mentioned before. We will perform additional experiments with articles and then compare the results of Eigenfactor and Journal Impact Factor.

Table 1. The features of our predictive model.

#	Feature	Median (Mean)	Range	Hypothesis
1	special char in a title	-	-	title contains special character
	Length of title:			
2	title (chars)	93 (96)	4-275	more title characters
3	title (words)	13 (13)	1-37	more title words
4	abstract (chars)	1379 (1355)	0-4661	more abstract characters
5	abstract (words)	199 (196)	0-706	more abstract words
	Number of:			
6	references	30 (34)	0-497	more references
7	pages	3 (5)	1-405	more pages
8	authors	4 (5)	1-42	more authors
9	affiliations	0 (0)	0-18	more affiliations
10	keywords	0 (2)	0-22	more keywords
	Citations received after N year(s):			
11	0	0 (0)	0-7	more release year citations
12	1	0 (1)	0-19	more after-1-year citations
13	2	0 (1)	0-42	more after-2-years citations
14	5	1 (3)	0-172	more after-5-years citations

Extended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 3-8.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Cheek, J., Garnham, B., Quan, J.: What's in a number? Issues in providing evidence of impact and quality of research(ers). *Qualitative Health Research*, (2006), vol. 16, no. 3, pp. 423-435.