Predicting Interest in Information Sources on the Internet using Machine Learning

Martin Číž*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xcizm@fiit.stuba.sk

The most important goal of each content provider on the Internet is to capture reader's interest, so that a reader becomes a returning customer. Although it is useful to evaluate impact of previously published articles, there is an opportunity to determine article's potential of popularity even before it is published. There are many attributes that may indicate whether an article has potential or not, such as:

- title,
- content,
- author,
- source,
- topic,
- freshness and
- credibility.

We can also take into account "external" attributes of an article, for example number of references from other web sources. Many works get this value from retweets on popular microblogging site Twitter [1]. However, we have decided not to include external references and focus more on article itself – we should be able to evaluate its quality and future popularity in its content, not outside of it.

To predict popularity of an article we plan to use an approach based on regression in supervised machine learning. There are many different regression methods. For start we will use linear regression to see the potential of this approach. Then we will try more advanced regression methods.

We cannot discuss popularity of an article without taking into account selected time frame. Shape of popularity curve is different one hour after publishing an article when compared to the one which can be observed one day after it. Each article usually has a peak of maximum readings after which it quickly stagnates on an average

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 9–10.

number, so we have established that popularity of an online news article can be defined as number of times an article was visited in a short period of time after its publication, for example in one day. Our goal is to cover maximum number of readings in this period of time. We exclude special cases when article has periodical peaks of readings (for example one article about autumn clothes might be popular every autumn), as we are interested only in new articles that become popular after being published.

To make prediction as accurate as possible we need large amounts of training data, which we will need to edit specially for Slovak language: remove joining words and edit key words to their root form or to their lexical form.

We realize that when a visitor clicks on an article from the front page, she was guided only by its title and sometimes also by a short description. However, the content might give us more insight on what the article is about and why it might be popular. We should have this in mind when deciding what influence article's content has on its popularity. The content can be transformed by extracting keywords and by creating vector of values.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0752/14.

References

[1] Zaman, Tauhid; Fox, Emily B.; Bradlow, Eric T., A Bayesian Approach for Predicting the Popularity of Tweets, The Annals of Applied Statistics 2014