Improve Sentiment Analysis of Citations with Author Modelling

Zheng Ma[1], Jinseok Nam[1] and Karsten Weihe[2]

[1] {ma,nam}@kdsl.informatik.tu-darmstadt.de
[1]Information Center for Education, German Institute for Educational Research (DIPF)
Schloßstraße 29, 60486 Frankfurt, Germany

[2] weihe@cs.tu-darmstadt.de
[1,2] Computer Science Department, Technische Universität Darmstadt
Hochschulstrasse 10, 64283 Darmstadt, Germany

Abstract

In this paper, we introduce a novel approach to sentiment polarity classification of citations, which integrates data about the authors' reputation. More specifically, our method extends the h-index with citation polarities and utilizes it in sentiment classification of citation sentences. Our computational results show that our method yields significant improvement in terms of classification performance.

1 Introduction

1.1 Background

Citation count between scientific publications have been the primary metric to measure importance and impact of articles or authors for decades. The advantage is simplicity and effectiveness. However, with the progress of machine learning, NLP and other disciplines, researchers developed various techniques to improve the quality of citation analysis and hence the quality of scholarly importance measurement. One of the efforts was to apply PageRank in citation network, which introduces weight on citation links for more accurate measurement (Ding, 2011). Although, this type of weighting is widely criticized (Alvarez and Soriano, 2014). Recent bibliometric studies showed that "there is no bad publicity in science", because criticized and controversial papers tend to be highly cited too (Radicch, 2012; Perc, 2014). Consequently, these controversial papers are positively estimated according to citation-count-based metrics, for example, impact factor and h-index. As Bonzi (1982) argued that if a cited work is criticized, it should consequently carry lower or even negative weight for bibliometric measures (Athar and Teufel, 2011).

1.2 Our contribution

Sentiment analysis of citation sentence makes this kind of fine-grained bibliometric measures possible. Augmented with polarized citation links, the citation network can be weighted more accurately by using negative weights. We introduce the p-index, which is the h-index extended by citation polarities.

Our assumption is that papers from prominent researchers are more probable to be cited in a positive manner than the papers from controversial researchers. Generally, if we know more about a researcher, it should be easier to determine the polarity of citations his/her paper receives. Our research question is whether or not the performance of citation sentiment analysis can be improved with better author modelling, in particular modeling with citation polarity.

In this paper, we report our on-going work on the citation sentiment analysis task. The rest of the paper is structured as follows: the next section briefly reviews some important work in this field. Then we introduce our method. In section 4, we present our experiment details. Preliminary experiment results and discussions about the results are structured in section 5 and 6. Finally, we summarize our work and discuss possible directions of future work.

2 Related Work

Teufel et al. (2006) was one of the pioneers in the field of citation classification. She proposed a classification scheme of 12 citation functions and used supervised technique for the classification task.

Proceedings of NAACL-HLT 2016, pages 122–127,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

Athar (2011) continued this thread of research and focused in polarity classification of citations. They experimented the SVM classifier with rich features, such as negation, sentence splitting and dependency features. Athar (2012) published the Citation Sentiment Corpus, which contains 8736 annotated sentences.

Jochim (2012) used the following facet scheme to describe the nature of a citation: conceptual vs. operational; organic vs. perfunctory; evolutionary vs. juxtapositional; based on vs. alternative work; confirmative vs. negational.

Dong (2011) performed semi-supervised classification over categories: background, fundamental idea, technical basis and performance comparison. They have also developed the ACL Anthology Searchbench, which provides very practical paper search functionality with graphical presentation of local citation network of searched paper.[1]

Abu-jbara et al. (2013) has also contributed to this field by working on citation context identification, citation purpose classification and citation polarity identification. Their method achieved good performance by using SVM with linear kernel on a rich feature set.

3 Methodology

3.1 The Basis: The H-Index

There are several metrics to measure the impact of researchers. H-index (Hirsch, 2005) and g-index (Egghe, 2006) are two popular ones. These metrics and their variations are all based purely on citation counts. The h-index is defined as follows:

"A scientist has index h if h of his or her Np papers have at least h citations each and the other (Np – h) papers have ≤h citations each." (Hirsch, 2005)

Mathematically, it can be represented as formula (1):

$$h\_index(f) = \max_i \min(f(i), i) \qquad (1)$$

where f is the function that corresponds to the number of citations for each publication, sorted in descending order.[2]

Obviously, h-index finds a balance point between publication amount and the citation count of each publication. It is however not able to model the different polarities of citations.

3.2 Our Extension: The P-Index

In order to embed the polarity information in author modelling, we introduce the p-index, where p stands for polarity:

$$p\_index(f) = h\_index(f) \cdot p^{\alpha} \cdot n^{\beta} \qquad (2)$$

where p is the amount of positive citations the author receives. and n is the amount of negative citations, with positive citation coefficient α and negative citation coefficient β. Since P-index is an indicator positively correlated with an author's academic performance and reputation, higher value corresponds to better performance and reputation. Thus, α as an exponential coefficient is defined to be greater than 1. Similarly, the negative citation exponential is defined in the range: $0 < \beta < 1$.

The polarity of citation reflects the opinion of peers. The controversial authors that receive a lot of negative citations can be distinguished with the p-index. Thus the p-index combines scholar's performance information measured by h-index and the reputation information measured by citation polarity distribution. In general, the author is better modeled with p-index.

4 Experiment

4.1 Corpus

In the experiments, we used the Citation Sentiment Corpus (Athar, 2012) with ACL Anthology Network (Abu-jbara et al., 2013). ACL Anthology Network (AAN) is a widely used corpus containing computational linguistic publications. It contains more than 21,212 papers, 17,792 authors and 110,975 citation links. The Citation Sentiment Corpus contains 8,736 citation sentences extracted from AAN. Theses citation sentences are manually annotated into three classes: positive, negative, objective (neutral).

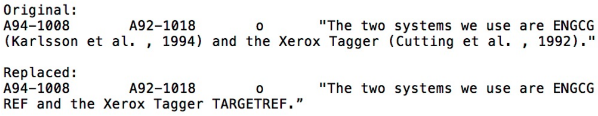
4.2 Preprocessing

Basically, we only performed two preprocessing steps, which we call text cleaning and citation mention replacement. In the text cleaning step, we solely removed erroneous characters and corrected the

[1] http://aclasb.dfki.de/

[2] https://en.wikipedia.org/wiki/H-index

spacing. In the citation mention replacement step, we use regular expressions to replace the target citation mention in the sentence with the label "TARGETREF", and other citation mentions with the label "REF", as shown in the following example:



```
Original:
A94-1008      A92-1018      o      "The two systems we use are ENGCG
(Karlsson et al. , 1994) and the Xerox Tagger (Cutting et al. , 1992)."

Replaced:
A94-1008      A92-1018      o      "The two systems we use are ENGCG
REF and the Xerox Tagger TARGETREF."
```

Figure 1 An Example of Citation Mention Replacement[3]

This way, the machine learning algorithm will be able to identify the position of the target reference and learn the semantic pattern between the target reference and the other tokens of the sentence. However, a few citation mentions cannot be automatically recognized. In our experiment, we ignore those citation sentences, in which the citation mentions cannot be replaced by TARGETREF or REF.

In order to exclude possible bias, we randomized the ordering of the data instances in the Citation Sentiment Corpus.

4.3 Features

In our experiments, each citation sentence is taken as one data instance, along with a set of features. We used the following feature set:

- Tf-IDF weighted unigram: straightforward feature for text classification tasks
- Polarity distribution of the target paper: the numbers of positive, negative and objective citations that the target paper receives. This is a kind of polarity modelling on paper level.
- Author ID: the author ID of the target paper as provided in AAN. We use the string form of author ID
- Affiliation ID: the author's affiliation ID of the target paper, also string form. To the best of our knowledge, our approach is the first attempt to utilize affiliation information in a citation sentiment analysis task.
- H-index/ P-index: the author's h-index or p-index of the target paper. The original h-index is provided in AAN and used by us as a baseline. P-index is calculated from h-index using formula (2) and we use it to improve the performance of citation sentiment classification.

4.4 Experimental Setup

In our experimental implementation, we used the Python module SciKit-Learn[4] for machine learning, feature extraction and evaluation. Support Vector Machine (SVM) is chosen as the classification algorithm. SVM is a widely used algorithm for text classification (Athar, 2012 and Abu-jbara, 2013).

Since our main contribution is improving classification performance using p-index, we mainly focus on the effects caused by replacing the h-index by the p-index. The SVM hyper-parameter is one of the settings that stay identical in all experiments. Therefore, in our on-going work, we have not broadly explored the parameter grid of penalty parameter C and kernel coefficient γ, which are the main parameters of SVM with Radial Basis Function (RBF) kernel (Hsu et al., 2006).

In order to answer our research question in section 1, we performed the experiment in a comparative manner. Firstly, we partition the corpus into 10 subsets and prepare 10 pairs of train and test sets. Then, for each pair, we perform the following procedure:

1. Run the classifier with h-index to obtain the baseline performance, denoted by $F1_b$ in the following.
2. Calculate the p-indices of authors as defined in Formula (2).
3. Update data instances of both train and test set: replace h-index value with p-index
4. Run the classifier on the new train and test set. It yields the test result $F1_t$.
5. Calculate the relative improvement from baseline to test result as:

$$\delta = \frac{F1_t - F1_b}{F1_b} \qquad (3)$$

[3] In this instance, "A94-1008" is the AAN paper ID of the citing paper and "A92-1018" is the ID of the cited paper. "O" is the sentiment annotation, meaning "objective" (neutral).

[4] http://www.scikit-learn.org

This procedure is applied to all the train and test set pairs. Next we calculate the average of all of these relative improvement. This is the final result of one experiment; it measures the effectiveness of our method under one specific setting.

Since the p-index depends on the positive and negative citation coefficients, α and β, the final relative improvement depends significantly on the choice of α and β. Systematic results are presented in the following section.

5 Result

5.1 Evaluation metric

In the field of text classification, Macro-F1 is widely used to evaluate the performance. It is especially suitable for our work, because the Citation Sentiment Corpus is highly imbalanced.

We did not compare our results with others. As a consequence of different data preparation, algorithm setting and many other detailed factors, it is nontrivial to reproduce the results reported in other work.

Moreover, in principle, our work is to examine the effectiveness of a novel feature on author level, which is independent to the utilization of other features on sentence level, etc. (Athar, 2011 and Abujbara et al., 2013).

5.2 Search the Setting Space

We performed numerous experiments with various settings. The dimensions of setting space are:

- Whether to enable feature: polarity distribution of the target paper
- Whether to enable feature: affiliation ID
- Whether to enable feature: author ID
- Positive citation coefficient α. We have sparsely explored the range of $1 < \alpha < 100$.
- Negative citation coefficient β. It has a relatively narrower definition range, $0 < \beta < 1$, which allows us to search more thoroughly.

It is a considerably large setting space to explore. We treated the first 3 dimensions as a group and the last 2 dimensions as another. In this way, we reduced the dimensionality of the setting space and made the exploration more practical.

Pol P	**Af. ID**	**Au. ID**	**Imprv.**
-	-	-	**9.3%**
Enabled	-	-	**2.6%**
-	Enabled	-	**6.2%**
-	-	Enabled	**9.6%**
Enabled	Enabled	-	**6.2%**
Enabled	-	Enabled	**9.6%**
-	Enabled	Enabled	**11.2%**
Enabled	Enabled	Enabled	**11.2%**

Table 1: Relative Improvements under different settings

Initially, we fixed the first setting group and search for the best values of the second setting group. After a coarse search, we found that the best settings of exponential coefficients are around the point ($\alpha = 1.1, \beta = 0.7$). Then a more intense search is performed in a small space surrounding this point. The best result achieved so far is at ($\alpha = 1.085, \beta = 0.727$).

Subsequently, we fixed α and β at these values and test the 8 possible combinations of the 3 boolean settings in the first group. The best results are obtained with both affiliation ID and author ID features enabled.

Thus we have spotted one local maxima in the settings space: {Polarity of Paper Feature: Disabled; Affiliation ID Feature: Enabled; Author ID Feature: Enabled; α = 1.085; β = 0.727}. This produces the most significant result so far: 11.2% relative improvement of Macro-F1 in the citation sentiment classification over the baseline: avg. Macro-F1=0.535. The baseline is obtained with the same system using h-index instead of the p-index.

5.3 Observation

Table 1 illustrates the relative improvement achieved under different feature selections. Each row represents an experiment. The first three columns indicate which features are enabled in an experiment and the last column reveals the relative improvement under this setting.

The first observation is that the polarity distribution of a paper (first column in the table) is not a helpful feature. When combined with other features, it hardly contributes to the system. When used alone, it decreases the performance.

In comparison, the affiliation ID is a better feature. Although it also causes some performance decline when used alone. But when combined with author ID, it makes positive contribution.

The best feature among these three is the author ID. The presence of author ID always boosts the performance of the system.

It is also worth mentioning that under this pair of exponential coefficients, all these possible feature selections deliver positive improvements.

5.4 Negative results

In the preceding sub section, we have reported the results from successful experiments. We have certainly tried other settings too, which are not as fruitful as the ones above.

Before we improved the sentiment classification by modelling the authors, we initially tried to improve it by modeling the papers. We applied python-igraph[5] implementation of the PageRank algorithm (Page et al., 1999) on the paper citation network. We used the PageRank value as a feature to model the target paper. However, in our experiments, this feature barely improved classification.

As for the p-index, we also tried different definitions. Instead of the exponential calculation, we tried linear and logarithm calculations too. The exponential version is proved to be the best.

Another variation of the p-index was to use $\log(p - index)$ in the place of p-index. Because when an author has a high positive citation count with relatively low negative citation count, her/his p-index will be considerably high. We presumed that suppressing it might be beneficial. However, experiment results support the p-index without logarithm.

Besides, we also tried out different settings of the machine learning classifier. We briefly tested SVM with linear kernel. Unlike reported by other work (Athar, 2011 and Abu-jbara et al., 2013), SVM with linear kernel performs worse than with RBF kernel in our experiments.

6 Discussion

As described in section 5, we answered our research question by comparing the Macro-F1 score between the baseline algorithm with h-index and the test algorithm with p-index. Significant improvement of classification performance verifies that the performance of citation sentiment analysis can be improved with better author modelling, in particular modeling with citation polarity.

Moreover, both good features observed in section 5.3 are the ones that describe some aspect of the author – namely, the affiliation of author or merely the ID of author. This straightforwardly supports our hypothesis that the better modeling of author is advantageous in citation sentiment analysis.

On the other hand, the feature “polarity distribution of paper” also models reputation, but on the paper level, which makes it useless for classification.

In the experiments, SVM with RBF kernel works better than SVM with linear kernel. The reason could be that the new features we introduced in this work needs non-linear separation in the hyper space. Another possible reason is that our parameter search is not adequately complete.

7 Conclusion and Future work

In this work, we answered our research question that polarity modeling of author significantly improves the citation sentiment analysis performance.

We believe that this approach can also contribute to sentiment analysis on social media or other domains. Generally, if the author’s reputation can be modeled using polarized links, the sentiment analysis should benefit from utilizing this model.

For example, if some online forum facilitates rating functions, e.g. “thumb-up” and “thumb-down” button, its users’ reputation can be modeled with polarity distribution. This model could assist the sentiment analysis of replies on the forum. Under the assumption that respected forum users are more likely to be replied in a positive sense.

With the current result as a proof of concept, we plan to further test our method by modeling the author with other methods, like PageRank on author citation network (Ding, 2011). We will also consider utilizing some popular semantic features to make our result more comparable to other systems.

Acknowledgments

This work has been supported by the German Institute for Educational Research (DIPF) as part of the

[5] http://igraph.org/python

graduate program "Knowledge Discovery in Scientific Literature (KDSL).

References

Amjad Abu-jbara, Jefferson Ezra. 2013 and Dragomir Radev. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of NAACL-HLT*, pages 596–606.

Myriam H. Alvarez. and Jose M. Soriano. 2014. Sentiment,Polarity and Function Analysis in Bibiometrics: A Review, In *Proceedings of the First Workshop on Argumentation Mining, pages 102-3*, ACL

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*. Pages 81-87. ACL.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Pages 597-601. ACL.

Susan Bonzi. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science,* 33(4):208–216.

Ying Ding. 2011. Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology* 62.2 (2011): 236-245.

Cailing Dong, and Ulrich Schäfer. 2011. Ensemble-style Self-training on Citation Classification. *IJCNLP*. Page 623-631.

Leo Egghe. 2006 "Theory and practise of the g-index." *Scientometrics* 69.1 (2006): 131-152.

Jorge E. Hirsch. 2005. An index to quantify an individual's scientific research output. In *Proceedings of the National academy of Sciences of the United States of America*, *102*(46), 16569-16572.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2003. A practical guide to support vector classification. Technical Report, National Taiwan University.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of the 2012 International Conference on Computational Linguistics*, pages 1343–1358

Lawrence Page, Sergey Brin, Rajeev Motwani & Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web. Technical Report, Stanford University.

Matjaz Perc. 2014.. The Matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98):20140378.

Filippo Radicchi. 2012. In science “there is no bad publicity”: Papers criticized in comments have high scientific impact. *Scientific Reports 2*

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. ACL.