An Iterative Approach for Citation Sentiment Analysis and Scholar Evaluation

Zheng Ma

ma@kdsl.informatik.tu-darmstadt.de

Research Training Group AIPHES

Computer Science Department, Technische Universität Darmstadt
Hochschulstrasse 10, 64283 Darmstadt, Germany

Keywords: Iterative machine learning approach, Sentiment analysis, Citation network, Author level metrics, P-index

Topic: Evaluation indicators, Data analysis and data mining

Abstract

Citation based metrics has been widely used to measure the impact of publications, journals and scholars. In this paper, we propose an iterative machine learning approach to improve the performance citation sentiment classification and the scholar modelling. This approach is based on the p-index, which is an extension of h-index, proposed by Ma et al. (2016).

1. Introduction

The total amount of knowledge acquired by mankind increases in an ever accelerating manner. Fuller (1982) has made an estimation of how quickly human knowledge doubles in size. He observed that the human knowledge doubles approximately every 100 years by the end of nineteenth century. In the middle of the twentieth century, the doubling time has decreased to 25 years. In academia, researchers advanced at an even higher pace. In merely 10 years, between 1901 and 1910, the amount of scientific publications has doubled (Hutton 1961). By that time, the catalogue of scientific papers contained already more than 380,000 author entries. This brought considerable challenges to the early bibliometricians, who were not equipped with modern computational tools. With few realistic alternatives available, the bibliometric analysis were often not further than citation counting (Smith 2012). As a result of its simplicity, the dominant role of citation counting in bibliometrics has been established and has remained until lately.

1.1. Importance of Citation Counting

Moreover, bibliometrics, especially the citation study, has now become a fundamental aspect of modern academia in almost all scientific disciplines (Smith and Hazelton 2011). “Bibliometric awareness” has become an indispensable criterion for research success (Smith and Hazelton 2011). In recent years, we have seen governments and other funding bodies increasingly seek to ensure that publicly-funded researchers are held accountable to disseminate their findings (Turale, 2010). Bibliometrics are very important in serving this role. For instance, United Kingdom established the Research Assessment Exercise system as well as Australia developed the Excellence in Research for Australia system (Anderson and Tressler 2009; Oppenheim 1995). Studies have shown that the journal ranking from these systems is highly correlated with citation count. Consequently, the important aspects of academia, e.g. research funding, research evaluation and scholar’s preference of publication are all affected by citation counting in some way.

1.2. Citation measures Impact

Much importance has been given to citation counting. But what does it really measure? When scientometricians are asked to choose what citation count measures between “quality” and “impact”, they are inclined to choose the latter. However, “quality” and “impact” are closely related. For instance, Cole et al. (1973) argued that “the data available indicate that straight citation counts are highly correlated with virtually every refined measure of quality” On the other hand, Martin et al. (1983) suggested that quality can be reliably indicated only when more than one indicators vote positively. They claimed “the indicators based on citations are seen as reflecting the impact, rather than the quality or importance, of the research work” (Martin and Irvine 1983).

While a spectrum of citation count based metrics are widely used, e.g. journal impact factor and h-index, a recent research warns the limitation of these traditional metrics, which also illustrates how “impact” relates to “quality” on an empirical scope. Radicchi et al. (2012) showed that “there is no bad publicity in science”. Criticized journal papers tend to have high scientific impact, which is reflected by citation count.

2. Methodology

2.1. Polarized Citation Network

A citation network demonstrates and reveals the collective wisdom that constructs a research area. The network structure provides material for an iterative expansion. Since each citation is made by the citing author’s free choice, it reflects the opinion of the citing author towards the cited paper. Apparently, these opinions are not always of the same nature, in which negative ones are also not negligible.

Teufel et al. (2006) and many other researchers have proposed various frameworks to categorize citation by their types and functions. Garfield (1962) published his seminal work, which provided a list of 15 possible reasons of citing. Among them, there are positive reasons, e.g. “Paying homage to pioneers or peers”, “Identifying methodology, equipment”. There are also 5 negative reasons, e.g. “Correcting previous work”, “Criticizing previous work”, “Disclaiming

work or ideas of others" and "Disputing priority claims of others". Similarly, Moravcsik et al. (1975) proposed another classification schema of citations, which also includes negative classes. In their quantitative study on the "Physical Review" articles, they found that one-seventh of the citations are "negational" and two-fifths are "perfunctory". This revealed the fact that considerable proportion of citations are negative. More recently, Teufel et al. (2006) proposed an annotation scheme for citations, which contains 2 clearly negative categories out of 12, namely "Unfavourable contrast/comparison" and "Weakness of cited approach". As a generalization of these various schemes, the most fundamental categorization is citation polarity categorization. In our recent work (Ma et al. 2016), we applied sentiment analysis on citation sentences, transformed the citation network into a polarized network and contributed to author modelling by introducing the p-index.

In this paper, we propose a method to iteratively improve the performance of sentiment analysis and thus the quality of author modelling by expanding the coverage of the polarity classification in the citation network.

2.2. Sentiment analysis

Sentiment analysis (SA) is an active field of research in natural language processing. The goal is to classify the sentiment polarity of sentences. While plausible performance of SA systems is achieved in its classic fields of applications, e.g. product review, SA in scientific writing remains a challenging task. One obvious reason is the objective, conservative, even implicit style of writing in scientific publications. Authors are especially cautious when they mention other work in a negative manner. Teufel et al. (2006b) and Athar (2014) have used machine learning techniques to automatically classify citation sentences.

2.3. Author modelling

Traditionally, the journal-level and article-level metrics are the main areas of research in bibliometrics and scientometrics. Author-level metrics have attracted interest in recent years. H-index (Hirsch 2005), author-level impact factor (Pan and Fortunato 2014) and author-level eigenfactor (West et al. 2013) are among the most popular ones. These metrics also utilize citation counts in some manner. As scholars play the central role in research/publication activities, author modelling will benefit a more comprehensive modelling of academia. Their work formed a constructive step towards this direction.

The widely used author-level metric h-index can be defined as below (Hirsch 2005; Torra and Narukawa 2008):

"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."
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.

We augment the h-index with polarities of citation links and developed the "p-index" (Ma et al. 2016):

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

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.

As an exponential coefficient, α is defined to be greater than 1. Similarly, the negative citation exponential is defined in the range: $0 < \beta < 1$. Thus, P-index is defined as an indicator positively correlated with an author's academic reputation. Higher value corresponds to better reputation and vice versa. As an extension to h-index, p-Index serves as a sentiment-sensitive author-level citation metric. (Ma et al. 2016) demonstrated that the performance of citation sentiment analysis is significantly improved with the help of p-index as a key feature in the machine learning algorithm.
In contrary to direct linking "quality" with "impact", we assume that the author's performance better conforms to his/her academic reputation. In technical terms, we assume that an author with more positive citations and less negative citations is likely to have better academic performance and thus higher publication quality.

2.4. An Iterative Machine Learning Approach

In this paper, we report our progress in the machine learning algorithm development for author modelling and citation sentiment analysis. To overcome the data sparsity problem, we use an iterative approach to incrementally exploit the vast amount of unlabeled citation links. The main contribution of this paper is the iterative machine learning algorithm that significantly improves the classification performance.

Firstly, we train a classifier on the train set in the first iteration, which provides a baseline performance of citation sentiment classification. In each further iteration, we use the classifier trained in the previous iteration to classify an unlabeled auxiliary dataset. Then we calculate the p-index of all authors using all labeled data available and replace the h-index with p-index. Now we have more training data instances (from auxiliary set) and updated knowledge of authors (p-index), we can train a new classifier of the current iteration and have it tested. It is expected to be the best classifier so far, as it is trained on the best knowledge available at this step.

In the next iteration, we calculate new p-index based on the train set and test set, which is labeled in the test phase of the last iteration. We also use the latest classifier to predict and label a new auxiliary set. And again, we can train a new classifier with the best knowledge at this point. This procedure repeats until the stopping-criteria is met.

The algorithm can be briefly summarized in the following listing:

1. Prepare dataset
2. First Iteration (Absolute Baseline)
    a. Train citation sentiment classifier on TrainSet
    b. Test on TestSet to obtain the baseline
3. Second Iteration
    a. Calculate p-index
    b. Prepare AuxSet = Aux-Deg2

c. Use the classifier trained in step 2a. to predict and label AuxSet
d. Update h-index with p-index in TrainSet, TestSet, AuxSet
e. Train classifier on TrainSet+AuxSet
f. Test on TestSet >> 2nd result

4. Third Iteration
    a. Calculate new p-index and update TrainD, TestD and AuxSet = Aux-Deg3
    b. Use the classifier trained in step 3e. to predict on the updated AuxSet
    c. Train classifier on TrainSet+AuxSet
    d. Test on TestSet >> 3rd result
5. Further Iteration N

    Repeat "Third Iteration" with AuxSet = Aux-DegN. Stop when the stopping-criteria is met.

3. Experiments

3.1. Dataset

In our experiments, we use the ACL Anthology Network (AAN) (Radev et al. 2009) as the paper/author corpus and Citation Sentiment Corpus (CSC) (Athar and Teufel 2012). CSC has the sentiment classification label on a fraction of the citations in the AAN corpus. The size of CSC is around eight thousand citation sentences. In previous experiments, we observed that the machine learning algorithm suffers from data sparsity with the dataset. In this paper, the main contribution is utilizing the unlabeled data to improve machine learning performance. In particular, we prepare the auxiliary datasets, as described in the following.

In a large citation network like AAN, there often exist more than one citation paths connecting two sub-sets of nodes. These paths may have various length. We use these alternative citation paths as the source of the auxiliary dataset. In this experiment, Aux-Deg2 is the collection of citation links which are on the 2-edge citation paths between the source paper set and target paper set. Aux-Deg3 is the collection of citation links on the 3-edge citation paths from the source papers to the target papers and so forth. To avoid including too distant citation links, the auxiliary dataset stops expansion after the 5th iteration.

3.2. Results

Support Vector Machines (SVM) is regarded as a good choice as a sentiment classification algorithm (Cortes and Vapnik 1995; Maas et al. 2011; Pang, Lee, and Vaithyanathan 2002; Wilson, Wiebe, and Hoffmann 2009). We use it to classify the sentiment polarity of citation sentences. The feature set comprises of author level features i.e. p-index, author-ID and affiliation-ID as well as basic linguistic features. The stopping criteria is when the Macro F1 score is not improving for 2 times in a row or there is no new auxiliary dataset available.

We evaluate the performance of the classifiers in each iteration according to the strict Macro-F1 measure. Each evaluation is performed using 10-fold cross validation.

To demonstrate the advantage of the iterative approach, instead of using the first iteration, we set the baseline at the iteration when p-index is involved, namely the 2nd iteration. An example of our preliminary results is listed in **Table 1.**

Since we use 10-fold cross validation, there are 10 combination of train set and test set split. Each one of the first 10 rows in **Table 1** lists the Macro F1 score of all the iterations performed on one data split.

Iter.1	**Iter.2**	**Iter.3**	**Iter.4**	**Iter.5**	**Iter.6**	**Baseline**	**Test**
0.514	0.541	0.509	0.508	0.504		0.541	0.541
0.52	0.54	0.551	0.548	0.571	0.571	0.54	0.571
0.535	0.569	0.553	0.547	0.563	0.563	0.569	0.569
0.536	0.555	0.587	0.6	0.6	0.6	0.555	0.6
0.501	0.521	0.536	0.529	0.533	0.533	0.521	0.536
0.505	0.522	0.548	0.549	0.565	0.565	0.522	0.565
0.5	0.489	0.518	0.511	0.512	0.512	0.5	0.518
0.631	0.601	0.578	0.578			0.631	0.631
0.539	0.539	0.542	0.524	0.523		0.539	0.542
0.506	0.491	0.506	0.542	0.521		0.506	0.542
					Avg.	**0.5424**	**0.5615**
*Each row represents a data split in the 10-fold cross validation.					**Rel. Improv.**	**2.59%**	**6.2%**

Table 1 Macro-F1 scores of iterative sentiment classification - preliminary result example

For comparison, the baseline result (non-iterative approach) is defined as the best Macro-F1 achieved until the 2nd iteration. The test result (iterative approach) is defined as the best Macro-F1 achieved until the last iteration. The result of first iteration, where h-index instead of p-index is used, is defined as the "absolute baseline".

As shown in **Table 1**, the Macro F1 score is further improved after the second iteration, by 8 data splits out of 10. The average Macro F1 of non-iterative method is 0.5424; relative improvement compared to the "absolute baseline" is 2.59%. In contrary, the average Macro F1 of the iterative method is 0.5615 and the relative improvement compared to the "absolute baseline" is 6.2%.

Along with these iterations, the h-index of authors have also incrementally evolved into an optimized p-index. Since p-index is calculated based on the count of positive and negative citations. Every improvement of the citation sentiment classifier will result in an improvement of the accuracy of the p-index. The quality of author modelling is improved consequently.

4. Summary

Citation analysis has always been an important area of scientometric research. In this paper, we proposed an iterative machine learning approach to improve the performance of citation sentiment analysis. The main contribution is providing an approach to utilize the unlabeled data in the citation network. Preliminary experiment results showed that the iterations significantly improved the sentiment classification performance. As a result, the quality of author modelling by means of p-index is also improved.

With citation sentiment analysis, the measuring power of citations is enhanced. By considering citing author's opinion, which is expressed in the citation sentence, it is more capable to gauge the "quality" besides the "impact".

Acknowledgement

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1."

Bibliography

Anderson, David L. and John Tressler. 2009. "The 'Excellence in Research for Australia' Scheme: A Test Drive of Draft Journal Weights with New Zealand Data." *Agenda: A Journal of Policy Analysis and Reform* 16(4):7–24.

Athar, Awais. 2014. *Sentiment Analysis of Scientific Citation*. *Technical Report Number 856, University of Cambridge, Computer Laboratory*

Athar, Awais and Simone Teufel. 2012. "Context-Enhanced Citation Sentiment Detection." Pp. 597–601 in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cole, J. R. and S. Cole. 1973. "Social Stratification in Science." *American Journal of Physics* 42:923.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20(3):273–97.

Fuller, R.Buckminster. 1982. *Critical Path*. St. Martin's Griffin.

Garfield, Eugene. 1964. "Can Citation Indexing Be Automated?" *Statistical Assoc . Methods for Mechanized Documentation* 269:84–90.

Hirsch, J. E. 2005. "An Index to Quantify an Individual's S Scientific Research Output." *Proceedings of the National Academy of Sciences, U.S.A.* 102(46):16569–72.

Hutton, R. S. 1961. *Journal of Documentation* 17(1):3–14.

Ma, Zheng, Jinseok Nam, and Karsten Weihe. 2016. "Improve Sentiment Analysis of Citations with Author Modelling." Pp. 122–27 in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Maas, Andrew L. et al. 2011. "Learning Word Vectors for Sentiment Analysis." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 142–50.

Martin, Ben R. and John Irvine. 1983. "Assessing Basic Research. Some Partial Indicators of Scientific Progress in Radio Astronomy." *Research Policy* 12(2):61–90.

Moravcsik, Michael J. and Poovanalingam Murugesan. 1975. "Some Results on the Quality and Function of Citations." *Social Studies of Science* 5(1):86–92.

Oppenheim, C. 1995. "The Correlation between Citation Counts and the 1992 Research Assessment Exercise Ratings for British-Library and Information-Science University Departments." *Journal of Documentation* 51(1):18–27.

Pan, Raj Kumar and Santo Fortunato. 2014. "Author Impact Factor: Tracking the Dynamics of Individual Scientific Impact." *Scientific Reports* 4:4880.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up?: Sentiment Classification Using Machine Learning Techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 79–86.

Radev, Dragomir R., Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. "The ACL

Anthology Network Corpus." Pp. 54–61 in *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, *NLPIR4DL '09*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Radicchi, Filippo et al. 2012. "In Science 'there Is No Bad Publicity': Papers Criticized in Comments Have High Scientific Impact." *Scientific Reports* 2:25–27.

Smith, Derek R. 2012. "Impact Factors, Scientometrics and the History of Citation-Based Research." *Scientometrics* 92(2):419–27.

Smith, Derek R. and Michael Hazelton. 2011. "Bibliometric Awareness in Nursing Scholarship: Can We Afford to Ignore It Any Longer?" *Nursing and Health Sciences* 13(4):384–87.

Teufel, Simone, A. Siddharthan, and Dan Tidhar. 2006a. "An Annotation Scheme for Citation Function." *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (July):80–87.

Teufel, Simone, A. Siddharthan, and Dan Tidhar. 2006b. "Automatic Classification of Citation Function." *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (July):103–10.

Torra, V. and Y. Narukawa. 2008. "The H-Index and the Number of Citations: Two Fuzzy Integrals." *IEEE Transactions on Fuzzy Systems* 16(3):795–97.

West, Jevin D., Michael C. Jensen, Ralph J. Dandrea, Gregory J. Gordon, and Carl T. Bergstrom. 2013. "Author-Level Eigenfactor Metrics: Evaluating the Influence of Authors, Institutions, and Countries within the Social Science Research Network Community." *Journal of the American Society for Information Science and Technology* 64(4):787–801.

Wilson, Theresa A., Janyce Wiebe, and Paul Hoffmann. 2009. "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis." *Computational Linguistics* 35(3):399–433.