Survey in sentiment, polarity and function analysis of citation

Myriam Hernández A
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Quito, Ecuador
myriam.hernandez@epn.edu.ec

José M. Gómez
Universidad de Alicante
Dpto de Lenguajes y Sistemas Informáticos
Alicante, España
jmgomez@ua.es

Abstract

In this paper we proposed a survey in sentiment, polarity and function analysis of citations. This is an interesting area that has had an increased development in recent years but still has plenty of room for growth and further research. The amount of scientific information in the web makes it necessary innovate the analysis of the influence of the work of peers and leaders in the scientific community. We present an overview of general concepts, review contributions to the solution of related problems such as context identification, function and polarity classification, identify some trends and suggest possible future research directions.

1 Extended abstract

The number of publications in science grows exponentially each passing year. To understand the evolution of several topics, researchers and scientist require locating and accessing available contributions from among large amounts of available electronic material that can only be navigated through citations. Citation analysis is a way of evaluating the impact of an author, a published work or a scientific media.

Sugiyama (2010) established that there are two types of research in the field of citation analysis of research papers: citation count to evaluate the impact (Garfield, 1972) and citation content analysis (Councill et al., 2008).

The advantages of citation count are the simplicity and the experience accumulated in scientometric applications, but many authors have pointed out its weakness. One of the limitations is that the count does not difference between the weights of high and low impact citing papers. PageRank (Page et al., 1998) partially solved this problem with a rating algorithm. Small (1973) proposed co-citation analysis to supplement the qualitative method with a similarity measure between works A and B, counting the number of documents that cite them.

Recently, this type researchers' impact measure has been widely criticized. Bibliometric studies (Radicchi, 2012) show that incomplete, erroneous or controversial papers are most cited. This can generate perverse incentives for new researchers who may be tempted to publish although its investigation is wrong or not yet complete because this way they will receive higher number of citations (Marder et al., 2010). In fact, it also affects the quality of very prestigious journals such as Nature, Science or Cell because they know that accepting controversial articles is very profitable to increase citation numbers. Moreover, as claimed by Siegel and Baveye (2010), it is more influential the quantity of articles than their quality or than the relationship between papers with a higher number of citations and the number of citations that, in turn, they receive (Webster et al., 2009).

Other limitation of this method is that a citation is interpreted as an author being influenced by the work of another, without specifying type of influence (Zhang et al., 2013) which can be misleading concerning the true impact of a citation (Young et al., 2008). To better understand the influence of a scientific work it is advisable to broaden the range of indicators to take into account factors like the author's disposition towards the reference, because, for instance, a criticized quoted work cannot have the same weight than other that is used as starting point of a research.

Proceedings of the First Workshop on Argumentation Mining, pages 102–103,
Baltimore, Maryland USA, June 26, 2014.

These problems are added to the growing importance of impact indexes for the researchers' career. It is becoming more important to correct these issues and look for more complete metrics to evaluate researchers' relevance taking into account many other "quality" factors, one of them being the intention of the researcher when citing the work of others.

Automatic analysis of subjective criteria present in a text is known as Sentiment Analysis. It is part of citation content analysis and is a current research topic in the area of natural language processing in the field of opinion mining and its scope includes monitoring emotions in fields as diverse as marketing, political science and economics. It is proposed that this type of analysis be applied in the study of bibliographic citations, as part of citation content analysis, to detect the intention and disposition of the citing author to the cited work, and to give additional information to complement the calculation of the estimated impact of a publication to enhance its bibliometric analysis (Jbara and Radev, 2012). This analysis includes syntactic and semantic language relationships through speech and natural language processing and the explicit and implicit linguistic choices in the text to infer citation function and feelings of the author regarding the cited work (Zhang et al., 2013).

A combination of a quantitative and qualitative/subjective analysis would give a more complete perspective of the impact of publications in the scientific community (Jbara et al., 2013). Some methods for subjective citation analysis have been proposed by different authors, but they call for more work to achieve better results in detection, extraction and handling of citations content and to characterize in a more accurate way the profile of scientists and the criticism or acceptance of their work.

Although work in this specific area has increased in recent years, there are still open problems that have not been solved and they need to be investigated. There are not enough open corpus that can be worked in shared form by researchers, there is not a common work frame to facilitate achieving results that are comparable with each other in order to reach conclusions about the efficiency of different techniques. In this field it is necessary to develop conditions that allow and motivate collaborative work.

Acknowledgments

This research work has been partially funded by the Spanish Government and the European Commission through the project, ATTOS (TIN2012-38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7-611312) and FIRST (FP7-287607).

Reference

Councill, I. G., Giles, C. L., & Kan, M. Y. (2008, May). ParsCit: an Open-source CRF Reference String Parsing Package. In LREC.

Garfield, E. (1972, November). Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science.

Jbara, A., & Radev, D. (2012, June). Reference scope identification in citing sentences. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 80-90). Association for Computational Linguistics.

Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In Proceedings of NAACL-HLT (pp. 596-606).

Marder, E., Kettenmann, H., & Grillner, S. (2010). Impacting our young. Proceedings of the National Academy of Sciences, 107(50), 21233-21233.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Radicchi, F. (2012). In science "there is no bad publicity": Papers criticized in comments have high scientific impact. Scientific reports, 2.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265-269.

Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010). Identifying citing sentences in research papers using supervised learning. In Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on (pp. 67-72). IEEE.

Webster, G. D., Jonason, P. K., & Schember, T. O. (2009). Hot Topics and Popular Papers in Evolutionary Psychology: Analyses of Title Words and Citation Counts in Evolution and Human Behavior, 1979-2008. Evolutionary Psychology, 7(3).

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. PLoS medicine, 5(10), e201.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology, 64(7), 1490-1503.