

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/325314056>

# Measuring the Centrality of the References in Scientific Papers

Conference Paper · May 2018

DOI: 10.1145/3209280.3229104

CITATIONS

0

READS

435

3 authors:



**Ollagnier Anaïs**

Université Côte d'Azur

17 PUBLICATIONS 46 CITATIONS

[SEE PROFILE](#)



**Sébastien Fournier**

Aix-Marseille Université

88 PUBLICATIONS 360 CITATIONS

[SEE PROFILE](#)



**Patrice Bellot**

Aix-Marseille Université

225 PUBLICATIONS 1,327 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Tweet Contextualization [View project](#)



A Combination of Reduction and Expansion Approaches to Deal with Long Natural Language queries [View project](#)

# Measuring the Centrality of the References in Scientific Papers

Anais Ollagnier  
Aix Marseille Univ, Université de  
Toulon, CNRS, LIS, OpenEdition  
Marseille, France  
anais.ollagnier@univ-amu.fr

Sébastien Fournier  
Aix Marseille Univ, Université de  
Toulon, CNRS, LIS  
Marseille, France  
sebastien.fournier@univ-amu.fr

Patrice Bellot  
Aix Marseille Univ, Université de  
Toulon, CNRS, LIS  
Marseille, France  
patrice.bellot@univ-amu.fr

## ABSTRACT

Citation analysis is considered as major and one of the most popular branches of bibliometrics. Citation analysis is based on the assumption that all citations have similar values and weights each equally. Specific research fields like content-based citation analysis (CCA) seeks to explain the “how” and “why” of citation behavior. In this paper we tackle to explain the “how” from a centrality indicator based on factors which are built automatically according to the authors’ citation behavior. This indicator allows to evaluate bibliographical references’ importance for reading the paper with which user interacts. From objective quantitative measurements, factors are computed in order to characterize the level of granularity where citations are used. By the setting of the centrality indicator’s factors we can highlight citations which tend towards a partial or a global construction of the authors’ discourse. We carry out a pilot study in which we test our approach on some papers and discuss the challenges in carrying out the citation analysis in this context. Our results show interesting and consistent correlations between the level of granularity and the significance of citation influences.

## KEYWORDS

Content-based citation analysis, bibliometrics, references’ influence, centrality indicator

## 1 BACKGROUND

The analysis of academic writing from citation has been extensively used to measure the impact of publications. Nowadays these studies have a more flexible and comprising perspective as the detection of research trends, the cross-fertilization between research disciplines and more recently discovering the value of scientific research and forecasting future critical and key technology [3]. Despite the proposal of Alternative Metrics, bibliometric methods are firmly established and are an integral part of research evaluation methodology. Since the development of bibliographic database services such as *Web of Science* or *Scopus*, automatic citation analysis studies, which focus on whom researchers cite and address the “how” and “why” questions of citation behavior, has grown over the past years.

However, the use of citations to perform these analysis are established on some strong assumptions. Indeed, the persisting oversimplification of citation behavior involves that the citation analysis

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*DocEng’18, August 2018, Halifax, Nova Scotia, Canada*

© 2018 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06.

[https://doi.org/10.475/123\\_4](https://doi.org/10.475/123_4)

is based on the assumption that each reference makes equal contribution to the citing article [4]. In the recent years, studies have been conducted on how citations are used in researcher evaluations to avoid those based solely on quantity. This research field, known as content-based citation analysis (CCA), investigates motivations and purposes of citation usages. CCA can be divided into semantic (e.g., to understand researchers’ motivations and purposes of citation usages) and syntactic (e.g., to find the citation’s location within a standardized section) approaches. At the beginning of these studies, CCA were conducted manually on small paper sets. Later with recent developments in computing and information services, machine-learning techniques such as NLP have been implemented allowing partially to automate processes on large scale. Recently, CCA has been used for various applications such as citation recommendation systems [12], sentiment analysis applied to the context of the citations [13], citation categorization [9] and citation summarization [11]. Despite the endeavors research on CCA [6, 10], there is still no automatic content-based approach which allows to understand to what extent the writing of a paper is based on other papers.

Leveraging information extracted from bibliographical reference analysis and from quantitative measurements, we propose a centrality indicator which allows to evaluate bibliographical references’ importance in the authors’ discourse. This indicator is integrated in the *BIBLME RecSys* system, the scholarly recommender system we developed in the context of a large digital library dedicated to Humanities and Social Sciences (HSS), allows to determine for each reference in a paper according to its citation occurrences whether this reference is used for the partial or the global writing of the paper. From this indicator, we don’t argue about how authors address a citation’s value according to its context at both the syntactic and semantic levels, we investigate how authors use citations in order to determine their influence on authors’ writing. The novelty of our approach is leveraged citations in order to determine automatically references’ importance.

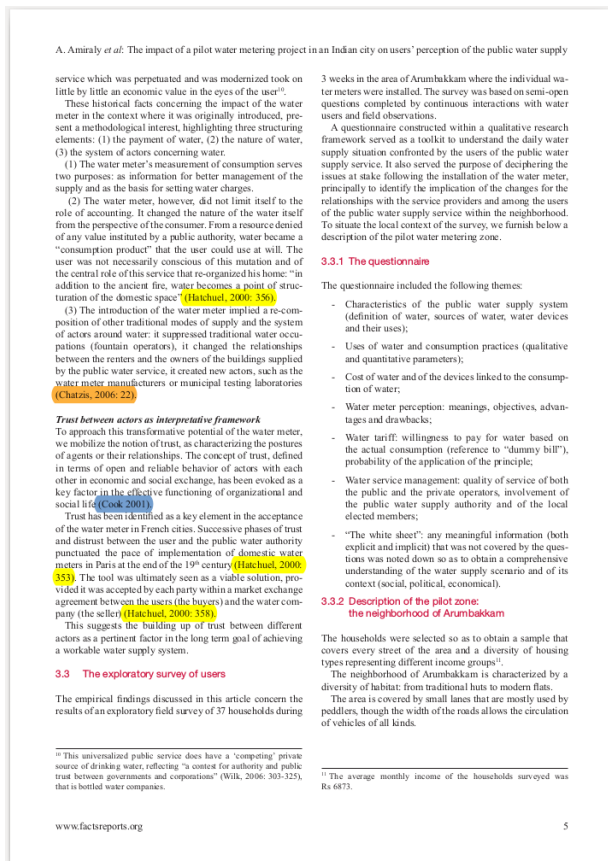
## 2 BIBLME RECSYS’ CENTRALITY FACTORS

In order to provide factors with the ability to highlight references according to their influence on the authors’ discourse, we focused on the characterization of the citation behaviors through quantitative measurements. To disambiguate, in this paper we use the term reference for the work that is cited and citation for the mention of it in the text.

### 2.1 Centrality Factors

Through citations, an author can promote a paper’s merit by including others works in order to further their own approaches, or by

contextualizing their works within the broader literature. Several kinds of citations have been identified through behavioral analysis of citations [1, 2]. Contrary to these works focusing on motivations and purposes of citation usages, we investigated on citations' distribution granularity within scholarly papers. By this way our purpose is to determine central references in a given paper. As we show in Figure 1, which corresponds to an extract of the paper "The impact of a pilot water metering project in an Indian city on users' perception of the public water supply"<sup>1</sup> written by A. Amiraly and A. Kanniganti, citations can be used according to different levels of distribution granularity.



**Figure 1: Example of citations' distribution granularity – yellow color: Hatchuel's citations, blue color: Cook's citation and orange color: Chatzis's citation**

By the term "distribution granularity", we refer to the textual density between each citation (occurrence in the paper) corresponding to the same reference. As we show in Figure 1 where each color corresponds to several citations referring to the same references, citations can appear in some specific zones (e.g., within a paragraph or within a section) or can be scattered throughout the paper. Our aim is to estimate the influence of references in a given paper by determining the level of distribution granularity of its citations and

then the centrality of each reference. To do that, we have established two factors which allow to identify different levels of distribution granularity, namely, the fine granularity and the coarse granularity. Based on the assumptions that references' importance increase proportionally with numbers of mentions and more detailed discussion of the cited document [8], we propose to construct the granularity factors according to the following assumptions:

- the more citations of a reference occur throughout the paper the more the analyzed papers is influenced by this reference;
- on the contrary, the concentration of citations within small and precise sections tends to strengthen the authors' arguments on specific aspects of the analyzed paper.

The setting of these factors can reveal the mutual contact of the reference and its citations, but can also reflect their contact strength from the quantitative and the distributional point of view.

## 2.2 Centrality Indicator's Construction

Figure 2 shows the processes of the centrality indicator's construction. In step 1, we developed a bibliographical references detection system dedicated to scholarly papers [7] named BILBO and that is publicly available and deployed over the OpenEdition journals<sup>2</sup>. Thanks to it, the names of the authors, the titles, the year of publication and some meaningful elements of information are extracted in full-texts and in the reference sections at the end of the papers.

In Step 2, from references and their citations annotated and extracted, we build sets of citations and references. Then, we use these sets in order to link the citations to the bibliographical references by means of matching functions. These functions are both based on a strict matching and a fuzzy matching. They allow to compare citations with references but also citations between themselves. Then for each reference/citation whose matching functions are fulfilled, quantitative measures based on two factors can be computed for centrality.

The first factor corresponds to the frequency factor of usage counts citations which occur in the given paper. This factor aims to highlight the highest frequency. The second factor corresponds to the distribution granularity factors. Its aims is to discriminate the ways each reference appears in the paper. This factor has two levels of granularity :

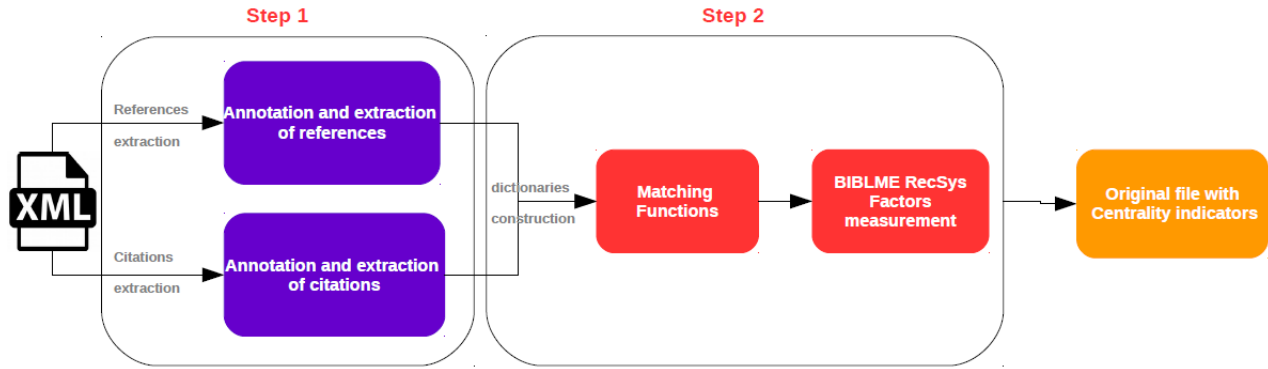
- **The fine granularity** is computed from citations referring to the same reference in the same paragraph and to the number of words between each one of these citations. A score is assigned if the number of words between these citations is less than the average of the distances between the citations corresponding to the same reference. The fine granularity function is as follows:

$$Granularity_{fine}(cit_i, cit_j) = \begin{cases} 1 & \text{if } a_j - b_i < Avg_{Cit} \\ 0 & \end{cases} \quad (1)$$

Where  $cit_i$  and  $cit_j$  are citations extracted from an ordered subset referring to the same reference  $ref$  in the same paragraph  $P$ .  $a_j$  is the  $cit_j$  start position and  $b_i$  is the  $cit_i$  end position in the paragraph  $P$ .  $Avg_{Cit}$  is the average of all the averages of distances

<sup>1</sup><http://journals.openedition.org/factsreports/831>

<sup>2</sup><https://github.com/OpenEdition/bilbo>



**Figure 2: Centrality indicators' Estimation.** Purple refers to annotation and extraction processes of references, red to the estimation of the indicators and orange to the output of the paper enriched by the centrality values.

in words between citations corresponding to the same reference  $ref$  in a paragraph  $P$ .

- **The coarse granularity** is measured from citations corresponding to the same reference throughout a given paper. We count the number of paragraphs which separate each of these citations. Then, a score is assigned if the number of paragraphs between these citations is less than the average of all the averages of distances. The coarse granularity function can be calculated as follows:

$$Granularity_{coarse}(cit_i, cit_j) = \begin{cases} 1 & \text{if } idx(Q) - idx(P) < Avg'_{Cit} \\ 0 & \end{cases} \quad (2)$$

Where  $idx()$  is a function which gives the index of a given paragraph.  $P$  and  $Q$  two paragraphs and  $cit_i$  and  $cit_j$  are the citations extracted from an ordered subset referring to the same reference  $ref$ .  $Avg'_{Cit}$  is the average of all the distance (number of tokens) averages between paragraphs that separate two citations corresponding to the same reference  $ref$ .

Lastly, each reference receives a centrality indicator which represents a linear combination<sup>3</sup> of the above factors. Users can set the weight of each factor. For example, if the coarse granularity factor is set at a high value, the references which occur throughout the paper are emphasized.

### 3 EXPERIMENTS

Some datasets are available such as the KDD dataset<sup>4</sup>, the iSearch collection<sup>5</sup> and the CORE dataset<sup>6</sup> which contain full-texts and citations. However, they do not provide information about citations' positions within the full-texts which is a prerequisite to compute the granularity factors. In order to test our proposal, the "Centre pour L'édition Électronique Ouverte" (Cléo) provided us data from its OpenEdition portal<sup>7</sup>. 12 papers have been selected randomly in

<sup>3</sup>The sum of the coefficients must be equal to 1.

<sup>4</sup><http://www.cs.cornell.edu/projects/kddcup/datasets.html>

<sup>5</sup><http://itlab.dbit.dk/~isearch/>

<sup>6</sup><https://core.ac.uk/services#dataset>

<sup>7</sup><http://www.openedition.org/>

various scientific fields such as languages, anthropology, ethnology, communication, law and culture, health, economy and development, education, agriculture, and environment. In the following, we explicitly give the five first references ranked for the paper "The hermit and the virtuoso"<sup>8</sup> written by D. Laborde. However, we experimented the approach on all the papers. This experiment can be reproduced online by means of *BIBLME RecSys*<sup>9</sup>.  $\gamma$  and  $\beta$  which correspond respectively to the value attributed to the coarse granularity factor and the fine granularity factor have been set alternatively at the highest value ( $0.8^{10}$ ).  $\alpha$  which refers to the frequency factor has been set at 0.1 for all experiments. Table 1 shows the rankings obtained from weight settings.

We can observe ranking changes according to the weights attributed to the granularity factors. We have identified 41 matches between citations concerning the coarse granularity and 5 matches for the fine granularity. Thanks to the manual study of the content of this paper, we have determined that the granularity factors allow to highlight different levels of centrality. For instance, concerning Table 1(a) we have observed the following citation usage: the Böhm's paper (indicator: 22) has been cited 7 times within 4 sections, the Cziffra's paper (indicator: 10.3) has been cited 6 times within 4 sections and the Sapiro's paper (indicator: 1.9) has been cited 3 times within 2 sections. Concerning Table 1(b), the Cziffra's paper (indicator: 1.9) has been cited 3 times within the same section, the Neuhaus' paper (indicator: 0.7) has been cited 4 times within the same section and the Sapiro's paper (indicator: 0.5) has been cited 2 times within the same section. Concerning the two first references of each Table, they refer to the paper's main topic which focuses on the pianist G. Cziffra. Their occurrences are both throughout the paper and are more condensed on some parts of the paper: it is a very central reference for this paper.

Finally, following the setting of the weights of the granularity factors, 63.6% of papers show a lot of changes with regard to the ranking of their references. For 18,2% of the papers the ranking is completely modified. Conversely, 36.4% of papers show no change. The aim of the papers plays a key role on how the author(s) frame(s)

<sup>8</sup><http://journals.openedition.org/ateliers/8841>

<sup>9</sup><http://www.lsis.org/ollagniera/demoV2/index.html>

<sup>10</sup>In order to have the sum of the coefficients equals to 1, 0.8 is the highest value here.

**Table 1: Citation ranking from weight settings. (a)  $\gamma = 0.8, \beta = 0.1, \alpha = 0.1$  (b)  $\gamma = 0.1, \beta = 0.8, \alpha = 0.1$** 

Citation	Centrality indicator
Böhm, 1995: Tribute to Cziffra	22.0
Cziffra, 1977: Cannons and flowers	10.3
Sapiro, 2007: The artistic vocation between donation and self-donation	1.9
Veyne, 1983: Did the Greeks believe their myths?	1.0
Neuhaus, 1971: The art of piano	0.7

(a)

Citation	Centrality indicator
Böhm, 1995: Tribute to Cziffra	4.5
Cziffra, 1977: Cannons and flowers	1.9
Neuhaus, 1971: The art of piano	0.7
Sapiro, 2007: The artistic vocation between donation and self-donation	0.5
Suchman, 1990: Action Plans, Practical Reasons	0.4

(b)

the papers and can change the usage of references. There are different types of scientific literature [5], for this article we experimented our approach on papers corresponding to an original research, a review of one another scientific paper or a case study. Considering that, we obtained various levels of centrality. Indeed, in a review article the author(s) give(s) an overview of existing literature in a field with a balanced perspective. At the opposite, in an original research the author(s) focus(es) more on precise previous studies that help to contextualize the proposal.

## 4 CONCLUSION

Beyond counting citations to a set of papers – by a single author, institution, or even an entire country – operated by current bibliometric indicators, we proposed a new bibliometric measurement to reflect references' centrality on a given paper. From the assumption that a reference's importance can be highlighted by its occurrences and how it is discussed within an academic writing, we have created two factors : the frequency factor and the granularity factors. From the setting of *BIBLME RecSys* factors, we observed that references' importance can not be reduced to the number of their mentions. Indeed, according to the value attributed to the granularity factors, different levels of centrality have been observed. The content study of selected papers allowed us to confirm that the centrality indicator we proposed is a way to reflect how authors frame their works. In the future, we plan to combine the centrality indicator with the current bibliometric indicators and to evaluate them on a large scale in the context of a recommender system dedicated to scientific papers.

## ACKNOWLEDGMENT

This research was supported by ANR program "Investissements d'Avenir" EquipEx DILOH (ANR-11-EQPX-0013).

## REFERENCES

- [1] Donald O. Case and Georgeann M. Higgins. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the Association for Information Science and Technology* 51, 7 (2000).
- [2] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *arXiv preprint arXiv:1609.00435* (2016).
- [3] Riu Li, Tamy Chambers, Ying Ding, Guo Zhang, and Liansheng Meng. 2014. Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology* 65, 5 (2014), 1007–1017.

- [4] Chi-Shiou Lin, Yi-Fan Chen, and Chieh-Yu Chang. 2013. Citation functions in social sciences and humanities: Preliminary results from a citation context analysis of Taiwan's history research journals. *Proceedings of the Association for Information Science and Technology* 50, 1 (2013), 1–5.
- [5] Rowena Murray and Sarah Moore. 2006. *The handbook of academic writing: A fresh approach*. McGraw-Hill Education (UK).
- [6] Jeppe Nicolaisen. 2007. Citation analysis. *Annual review of information science and technology* 41, 1 (2007), 609–641.
- [7] Anaïs Ollagnier, Sébastien Fournier, and Patrice Bellot. 2016. A supervised Approach for detecting allusive bibliographical references in scholarly publications. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. ACM, 36.
- [8] Rong Tang and Martin A. Safer. 2008. Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation* 64, 2 (2008), 246–272.
- [9] Zehra Taşkan and Umur Al. 2018. A content-based citation analysis study based on text categorization. *Scientometrics* 114, 1 (2018), 335–357.
- [10] Ludo Waltman. 2016. A review of the literature on citation impact indicators. *Journal of Informetrics* 10, 2 (2016), 365–391.
- [11] Jie Wang, Shutian Ma, and Chengzhi Zhang. 2017. Citationas: A summary generation tool based on clustering of retrieved citation content. *Framework* 7 (2017), 8.
- [12] Rui Yan and Hongfei Yan. 2013. Guess what you will cite: Personalized citation recommendation based on users' preference. In *Asia Information Retrieval Symposium*. Springer, 428–439.
- [13] Abdallah Yousif, Zhendong Niu, John K. Tarus, and Arshad Ahmad. 2017. A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review* (2017), 1–34.