Measuring the Centrality of the References in Scientific Papers
Anaïs Ollagnier, Patrice Bellot, Sébastien Fournier

To cite this version:

HAL Id: hal-01958737
https://hal.archives-ouvertes.fr/hal-01958737
Submitted on 25 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Measuring the Centrality of the References in Scientific Papers

Anaïs 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 to 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.

CCS CONCEPTS
• Information systems → Information retrieval; Retrieval tasks and goals; Recommender systems; Document representation; Content analysis and feature selection;

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 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” 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) and between references in the same paragraph. By that way we can see the differences in the importance of the references. The more frequent a reference is the more influential is it on the paper.

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

- The more citations of a reference occur throughout the paper the more the analyzed paper 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.

1http://journals.openedition.org/factsreports/831
2https://github.com/OpenEdition/bilbo
Granularity\textsubscript{fine}(cit\textsubscript{i}, cit\textsubscript{j}) = \begin{cases} 
1 & \text{if } a_j - b_i < \text{Avg}_{cit} \text{ref} \\
0 & \text{otherwise}
\end{cases} \quad (1)

Where \( \text{cit}_i \) and \( \text{cit}_j \) are citations extracted from an ordered subset referring to the same reference \( \text{ref} \) in the same paragraph \( P \). \( a_j \) is the \( \text{cit}_j \) start position and \( b_i \) is the \( \text{cit}_i \) end position in the paragraph \( P \). \( \text{Avg}_{cit} \text{ref} \) is the average of all the averages of distances in words between citations corresponding to the same reference \( \text{ref} \) 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 between 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:

\[
\text{Granularity}_{\text{coarse}}(\text{cit}_i, \text{cit}_j) = \begin{cases} 
1 & \text{if } idx(Q) - idx(P) < \text{Avg}_{cit} \text{ref} \\
0 & \text{otherwise}
\end{cases} \quad (2)
\]

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

Lastly, each reference receives a centrality indicator which represents a linear combination\(^\text{3}\) 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\(^4\), the iSearch collection\(^5\) and the CORE dataset\(^6\) 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\(^7\). 12 papers have been selected randomly in 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”\(^8\) written by D. Laborde. However, we experimented the approach on all the papers. This experiment can be reproduced online by means of \( \text{BIBLME RecSys} \)^\text{9}. \( \gamma \) 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\(^\text{10}\)). \( \alpha \) 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

---

\(^1\)The sum of the coefficients must be equal to 1.

---

\(^4\)http://www.cs.cornell.edu/projects/kddcup/datasets.html
\(^5\)http://atlab.dbt.dk/~isearch/
\(^6\)https://core.ac.uk/services#dataset
\(^7\)http://www.openedition.org/
\(^8\)http://journals.openedition.org/ateliers/8841
\(^9\)http://www.lax.org/ollagniera/demo/62/index.html
\(^10\)In order to have the sum of the coefficients equals to 1, 0.8 is the highest value here.
works. In the future, we plan to combine the centrality indicator we proposed is a way to re-

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.

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

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

(a)

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) 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.

ACKNOWLEDGMENT

This research was supported by ANR program “Investissements d’Avenir” EquipEx DILOH (ANR-11-EQPX-0013).

REFERENCES