Categorizations and Annotations of Citation in Research Evaluation

Marc Bertin

LaLIC Laboratory , University of Paris-Sorbonne
Maison de la Recherche 28 rue Serpente 75006 Paris, France
marc.bertin@paris-sorbonne.fr

Abstract

Scientific evaluation is based for the most part on citation analysis. The phenomena of citation is not yet well studied. The use of the Contextual Exploration technique that allows the automatic semantic annotation of the relations between authors, gives some new possibilities in this area. New bibliosemantic indicators can be considered in order to provide a better perception of citations through the study of semantic categories. The computer implementation proposed in this article raises the question of the evaluation of this approach.

Introduction

We will propose in this article a framework of a new qualitative approach to citation analysis, which is at the core of research evaluation. The necessity of new approaches in research evaluation has also been underlined by (Moed 2005) in his book "Citation Analysis in Research Evaluation", where he suggests that the efforts in the development of science evaluation should be directed towards qualitative citation analysis "through contextual and cognitive-relational analysis". We agree with his point of view on the perspectives in citation analysis:

> Quantitative analysts of science could develop more 'qualitative' citation based indicators, along contextual or a cognitive-relational viewpoint, thus abandoning the principle underlying most citation analyses that all citations are equal. [...] It would be necessary to develop initially simple, and in a later phase more sophisticated, classifications of 'how' documents are cited from the perspective of research evaluation rather than from that of information retrieval.

Presently, according to (Moed 2005), there exist two major problems that block all developments in this direction.

The first one is the difficulty to find databases of scientific articles that give access to the text content, in a format that permits a full text analysis. Probably the on-line journals and the new services will meet this demand in the next several years, but for the moment such databases are still marginal for the needs of a large-scale implementation.

The second problem for the development of new refined approaches to citation analysis lies in the conceptual difficulty associated with the automatic classification of citations according to the context. As presented in (Moed 2005), such systems do not exist yet.

We will discuss here in detail the approach to citation analysis that we propose. It is based on the idea that the citation is not a simple unit, but rather the product of a complex phenomenon. We assume, as a starting hypothesis, that the bibliography presents information that is essential for the evaluation of a publication. In addiction to this, the different motivations for citation and relations between authors should be considered. That is why, it is necessary to develop new methods that will allow the automatic identification of the different relations between authors, namely how a given author is cited by others. We will therefore propose a methodology for the development of linguistic resources for the automatic annotation of the relations between authors.

Methodology

Apparently, the different approaches, such as research performance, impact factor or scolarly quality, do not take into consideration certain aspects of the citations. In fact, the act of citation is a complex phenomenon that is subject to various internal and external limitations. Citations could be used as a strategic element, for example be citing the editor of a journal or some reviewers (Case and Higgins 2000). However, the citation is also an act of persuasion: to cite somebody of authority permits to validate one's argumentation in order to convince. This leads us to consider the Mathieu effect. In 1995, Merton (Merton 1995) explains The Matthew effect in this way :

The Matthew effect is the accruing of large increments of peer recognition to scientists of great repute for particular contributions in contrast to the minimizing or withholding of such recognition for scientists who have not yet made their mark.

However, this effect is not a necessary condition to obtain a certain visibility. If a paper is significant, it will have, finally, an impact and find its place in the community. The question that interests us here is whether all citations in an article have the same value. We cannot agree with Luukkonen and Zuckerman (Zuckerman 1987; Luukkonen 1990) who state that:

The existence of various cognitive meanings of citations and motivations for citing does not necessarily invalidate the use of citations as (imperfect) performance measures.

To understand the citation act, we have to show the motivation of an author to cite their colleagues. Motivation plays an important role in the citation act. In 1964, and then in 1977 Garfield (Garfield 1977) proposes fifteen different reasons for citation. Similarly, in 1977, Small (Small, 1977) distinguishes five categories of relations between authors and each of these relations can be categorized.

We can consider the citation as a scientific tradition that has the purpose to identify a specific point located in the text of another author. The approach that we will propose takes into consideration the above-mentioned limitations of the current methods for citation analysis. It is based on the automatic annotation of the textual segments containing bibliographic citations according to a number of semantic categories. The automatic semantic annotation is realized by the Contextual Exploration method (Desclés 1991; 2006).

Bibliographic citations in texts can be presented in various ways. They are sometimes in numerical form and other times they contain the name of the author. To take into account these variations, we have created a classification of the different numerical and alphanumerical families of bibliographic citations. The localization of the bibliographic citations in the text is made by regular expressions.

For the purpose of experimentation, we have based our analysis on a bilingual corpus of scientific articles in French and in English. After a first stage of automatic segmentation into sentences, paragraphs, etc., the segments in the text that contain bibliographic citations are identified by the regular expressions. Then, Contextual Exploration rules locate linguistic markers in the segments in order to carry out the semantic annotation.

We will use the semantic annotations obtained in this manner in order to make up a bibliographic database that will serve as a source for the construction of the bibliometric indicators. The semantic categories we identify during the annotation of segments are significant of the way the author has been cited in the text. The categorization we use is defined after the linguistic study of the discourse markers present in the textual segments that contain bibliographic citations. The semantic annotations make possible to access automatically the semantic content of the textual segments containing the bibliographic citations. The different categories for the annotation have been identified by Yordan Krushkov (Krushkov and Descles 2005). On Fig. 1 we present for information the semantic map that we use here. The linguistic resources are organized according to this semantic map.

		Quotation		
Point of view	**Comparison**	**Information**	**Definition**	**Appreciation**
Position	Resemblance Disparity	Analysis Citation Counter-example Hypothesis Method Result		Agreement Disagreement

Figure 1. Semantic map for Bibliosemantics

We consider that the phenomenon of auto-citation is especially important for the development of science evaluation methods. That is why, in our framework we make the distinction between the textual segments that contain citation of other authors, and the segments that contain auto-citations. Each of the categories in the semantic map can apply to both types of segments. We believe that this distinction is essential for the better understanding of the phenomenon of citation.

Implementation

The program implementation of our approach is based on two modules developed in the LaLIC laboratory: SegaTex and EXCOM. The figure 2 presents the general architecture of the system including the database and the web interface.

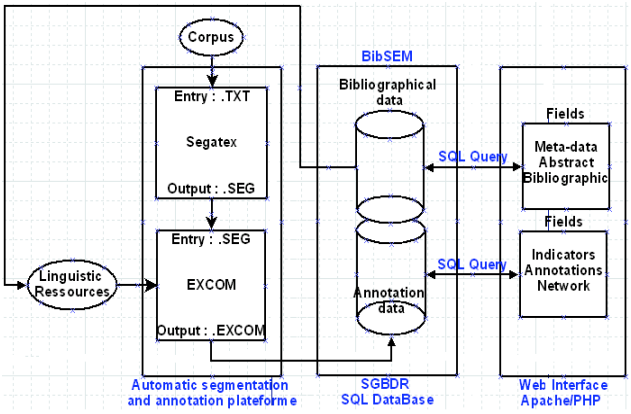

Figure 2. The architecture of our implementation

SegaTex

This is the module that carries out the automatic segmentation of texts into sentences, paragraphs, sections, etc. by using the work of (Mourad 1999; 2001). The segmentation rules of SegaTex, developed initially for French, apply successfully to texts in English. The result of the application of these rules is a file in the XML format. Fig. 3 presents a part of a segmented text.

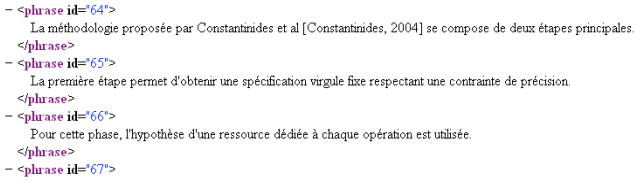

Figure 3. Segmentation of structured documents

EXCOM

The EXCOM system, developed by (Djioua 2006; Djioua and Desclés 2007), allows the automatic semantic annotation of the segmented texts. The annotation is added to the XML file in the form of new elements and attributes of the sentences. The semantic meaning of the annotations is related to the organization of each category recognized by the EXCOM system.

Figure 4. EXCOM Rules

The rules for the system are presented in the form of an XML file, structured as in the example on Fig. 4.

BibSEM

We have developed our proper database, called BibSEM, with controlled input and update. The information for the database is partially obtained by the attributes of the XML annotation output of EXCOM. In general there is a certain number of errors when working with external bibliographic databases, that can arise, for example, from errors in the input of some of the fields.

Interface

The graphical user interface that we have developed for the presentation and exploitation of the results is based on a classical Apache/php/Mysql system. Fig. 5 presents the organization of the page displaying the results. Here we will explain in detail the function of each of the elements of this page.

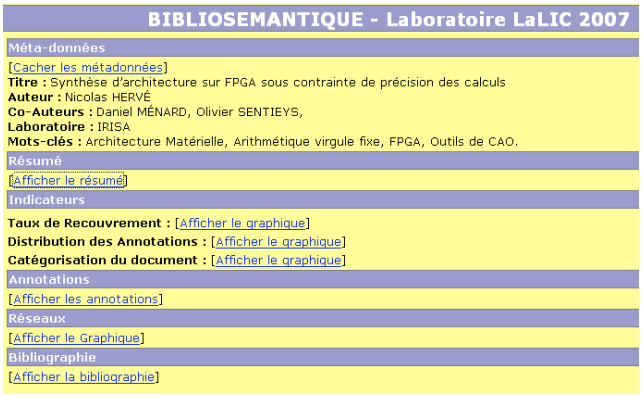

Figure 5. Graphical Interface

Metadata. This is the information that can traditionally be found in bibliographical databases, containing fields like title, authors, co-authors, keywords, etc.

Abstract. This is the abstract of the article that can be visualized by a dynamic link. Eventually, an automatic summary of the article could also be integrated in the interface. An appropriate summarization technique has been proposed and implemented by Antoine Blais (Blais 2007) whose summarization strategy is based on the automatic annotation of discourse categories by the Contextual Exploration method.

Indicators. The bibliometric indicators based on the semantic annotations carried out by the EXCOM system allow a refined analysis of the bibliographic citations. Three types of indicators are presented. The corresponding Indicators graphics will be presented further on.

Annotations. The zone of annotations allows the visualization of the annotated segments. In the context of our study the possibility to display the textual segment together with its annotation is of importance for the further analysis. In this zone, each of the bibliographic citations can be displayed in this way.

Networks. This part has not yet been implemented because the publications in our corpus have not yet been related to each other. We will discuss the approach in more detail in the final part of this paper.

Bibliography. The bibliography can be a source, for each new publication, of new linguistic markers for the bibliographic citations. In fact, sometimes in texts we can find citations that do not contain bibliographic references. However, we consider that these citations are also important for our analysis and have to be taken into consideration by the system. This approach is especially important for human and social sciences, where the publications in general have a very large number of references.

In the following part, we will discuss in more detail the bibliosemantic indicators.

Perspectives in using annotations

There is a great difference between the traditional approaches to citation analysis and the bibliosemantic approach issued by the linguistic analysis and the semantic categorization of the quotations. Here, we use the term quotation to designate all of the citations, references, and the bibliographic citations. As we have already mentioned, the automatic annotation by semantic categories requires an access to the document that permits full-text search. It should be noted that our aim is not the annotation of all textual segments, but rather the extraction and analysis of those segments that have been annotated. Here we propose some new bibliosemantic indicators, that will allows qualitative analysis.

Annotation indicator. This indicator gives two quantitative value : first one is bibliographical references identified in the article and next is the number of annotation corresponding to the number of textual segment identified by the bibliographical references which were annotated semantically.

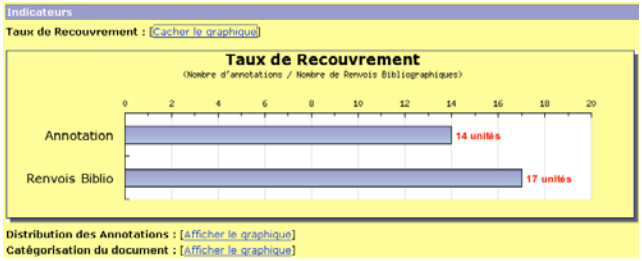

Figure 6. Annotation Indicator

This indicator is paramount in the evaluation of the system, because it shows the capacity of the system to annotate the bibliographic references. Figure 6 shows an example of the value of the coverage rate for one scientific article. It can be seen that from the total of 17 bibliographic references only 14 have been annotated.

Distribution of Annotations. According to the annotations, we can present graphically the distribution of the bibliographical references in the publication. If we take the example on the Fig.7, it is interesting to note that the segments annotated as definitions and results are very present in this article with a stronger concentration in the beginning and the middle of the publication. Moreover, in the middle of the article, almost at the same position, we see that there are segments annotated as Results and Information. This suggests that at this point in the article there is a discussion in which to position of the author is compared to other publications.

In this type of approach we conceive the act of citation as a non-uniform phenomenon, carrying a semantic meaning and a social and historical motivation. In this way of thought, the simple use of the bibliography as a measure unit is very restrictive. The distribution of annotations shows the structure of the citations in the text and the existence of non-homogeneous values of the citation according to their position in the text.

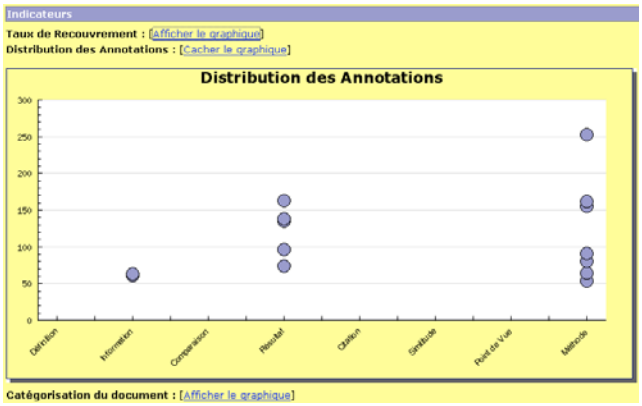

Figure 7. Distribution of annotations

Categorization of publication. The scientific publications vary in nature. For example we can consider articles of synthesis or methodology. Using the semantic map, we can categorize a document according to the relative number of bibliographic references for each of the categories on the highest level in the semantic map.

This approach makes it thus possible to consider bibliosemantic indicators based on the treatment of a single text and however remains compatible with statistical approaches. It also offers a greater facility for the development of new indicators. A normalization of these indicators will make possible their use for an evaluation or a larger scale which will allow comparison between publications.

Evaluation and discussion

It is difficult to measure the quality of scientific production by the existing methods because in the calculation of the bibliometrics indicators citations are considered as measure units, and not as a set of complex phenomena. Presently, it is essential to give some new impulsion.

By looking inside an article, we can understand that the distribution of citations is non-homogeneous. Citations have different meanings according to their position. Hargens 2000 et Voos and Dagaev 1976 (Hargens 2000; Voos and Dagaev 1976) have shown that citations occur more frequently in the beginning of publications. Indeed, in the introduction we can often find a brief state of art of the domain.

We have worked on a corpus of 16 scientific articles as a basis for the evaluation. The corpus contains 386 bibliographic links that have been identified automatically using the regular expressions presented in (Bertin 2006). The number of annotated segments is 67. Each of the

article in the corpus contains at least one annotation. At the present state of the linguistic resources, 17, 35% (Fig. 8) of the textual segments containing bibliographic links contain enough linguistic markers to be annotated.

Category	Segments
Comparison	1
Dissimilarity	1
Similarity	3
Definition	3
Hypothesis	1
Citation	4
Result	20
Analysis	2
Information	12
Example	3
Method	17

Figure 8. Liste of corpus annotation's using our categorization

This evaluation suggests several issues that must be taken into consideration for the optimization of the system. The first one is the fact that we work on the sentences as the basic textual segments. For the segmentation we use the Segatex program. An annotation of the sentence clauses would be more relevant for our needs. In fact, a segmentation into clauses would allow the identification of more than one annotation in the same sentence.

The second issue that we want to discuss is the larger size of some of the categories, such as the Methods and the Results, that need some particular attention. We have to preview other sub-categories, for example synthetic results, theoretical results, practical results, etc.

Thirdly, the reasons why the rest of the bibliographic references are not annotated are the following:

- For the annotation we consider only the relation between authors among all the possible relations
- The basic segments for the annotations are the sentences. However, an author can include several citations in the same sentence.
- Segmentation into sentence clauses have not yet been implemented into the system.
- The linguistic markers for the semantic annotation can be in another segment than the bibliographic citation.
- The last and most important reason is that a large number of citations are tacitly introduced in the text and therefore not possible to categorize semantically.

At last, in a single publication we cannot show a good visualization of the results (see fig. 6 and 7). It is interesting to note that we can present the annotations distribution on a graph where the horizontal axis represents the different categories and the vertical axis represents the number of the sentence in the text. Therefore, we can see (example on fig. 7) that there is a concentration of citations in the beginning of the article, around the first 50-100 sentences, and then another concentration around the sentence 150. At the end of the article, around the sentence 250, we have only one citation. We must have in mind that annotated citations do not represent all the citations in the document and the fig. 6 shows that for this article there are 17 bibliographic citations for 14 annotations.

Conclusion

The last few years, the attitude "publish or perish" appeared, leading to practices that could be unfavorable for the quality of the publications. One of the short term risks of this practice is undoubtedly an increased scientific production but of a lower quality, which obliges a reader to go through a large number of publications in order to study an idea or a concept. In a longer term, there is a risk of uniformity of scientific research, resulting in a decline of the diversity and a tendency of homogeneous research.

The report of Moed comes at a critical moment when the Impact Factor is subject to numerous criticisms. The difficulty in the categorization of quotations can be raised by the Contextual Exploration method. The platform EXCOM allows the automatic semantic annotation, and opens new possibilities for the citation analysis and the creation of new indicators. The conception of networks using this methodology will lead to a thorough analysis of the current indicators using the impact factor of Garfield and the evaluation of the science more widely.

We have proposed here a method that is innovative compared with the actual paradigm of citation analysis. We have discussed also the problem of evaluation. Our implementation shows this new method that uses the semantic annotation and structured information. This work in progress cannot be evaluated by using the traditional measures of precision and recall. According to the indicator Coverage rate, we have to develop new protocols for evaluation of the qualitative annotation. Our approach meets problems of different nature, such as sociological, psychological or historical, that have to be considered. For the purpose of the evaluation, we have to specify how human annotations can be used in an evaluation protocol. Finally, the bibliosemantic indicators must be defined with caution as they are the means for science evaluation.

Acknowledgments

I would like to thank professor Jean-Pierre Desclés for his advice, the fruitful discussions and his help in writing this article.

References

Bertin M., Desclés J.P., Djioua B.,Krushkov Y. 2006. "*Automatic Annotation in Text for Bibliometrics Use.* FLAIRS 2006, Melbourne, Florida

Blais A., Atanassova, I. , Desclés J-P., Zighem, L. and Zhang, M. 2007, Discourse Automatic annotation of Texts : an application to summarization.FLAIRS 2007, Key West, Florida.

Case D. O. and Higgins G. M. 2000. How Can We Investigate Citation Behavior? A Study of Reasons for Citing Literature in Communication. in *Journal of the American Society for Information Science, vol.51, n°7*, pp.635-645.

Desclés J-P., Jouis C., Oh H-G., Reppert D., 1991. Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte . In Knowledge modeling andexpertise transfer, pp.371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam.

Desclés J-P. 2006. Contextual Exploration Processing for Discourse Automatic Annotations of Texts. FLAIRS 2006, Melbourne, Florida. Invited Speakers

Djioua B.., Garcia Flores G., Blais A., Jean Pierre Desclés J-P., Gael G., Jackiewicz A., Le Priol F., Nait-Baha L., Sauzay B. 2006. *EXCOM : an automatic annotation engine for semantic information*. FLAIRS 2006, Melbourne, Florida.

Djioua B.. and Desclés J-P. 2007. Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts. FLAIRS 2007, Special Track "Automatic Annotation and Information Retrieval : New Perspectives", Key West, Florida, 9-11 Mai.

Garfield, E. 1977. Can Citation Indexing Be Automated? In Essay of an Information Scientist, vol. 1, Philadelphia: ISI Press.

Hargens. L. L. 2000. Graphing Micro-Regions in the Web of Knowledge: A Comparative Reference-Network Analysis. In B. Cronin & H. B. Atkins (Eds.), The Web of Knowledge. A Festschrift in Honor of Eugene Garfield Medford: ASIS, pp. 497-516.

Krushkov, Y. 2005. *L'exploration contextuelle des appariements entre les références bibliographiques et les passages textuels dans un corpus de textes linguistiques*. Under direction of Descles J-P. Master's thesis, University of Paris-Sorbonne.

Luukkonen, T. 1990. *Citations in the rhetorical, reward, and communication systems of science*. Unpublished PhD thesis, University of Tampere, Tampere.

Merton, R. K., 1995. The Thomas Theorem and The Matthew Effect. *Social Forces*, 74(2) 379-424, December.

Moed H.F. Eds. 2005. *Citation Analysis in Research Evaluation.* Springer.

Mourad G. 1999. La segmentation de textes par l'étude de la ponctuation. CIDE'99 (2e Colloque International sur le Document Électronique), p.155-171.

Mourad G. 2001. *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE*, Ph. D. Thesis, Univ, Paris-Sorbonne.

Small H.G. 1977. A Co-Citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research. *Social Studies of Science*, Vol. 7, 139-66

Voos, H., and Dagaev, K. S. 1976. Are All Citations Equal? Or, Did We Op. Cit. Your Idem? in *The Journal of Academic Librarianship, vol. 1, n°6*, 1976, pp. 19-21.

Zuckerman, H., 1987. Citation analysis and the complex problem of intellectual influence. in *Scientometrics, vol. 12 n°5-6*, pp.329-338.