Automatic Annotation in Text for Bibliometrics Use

Bertin M. and **Desclés J.P.** and **Djioua B.** and **Krushkov Y.**
LaLICC, UMR8139
Université Paris-Sorbonne/CNRS
28 rue Serpente,
75006 Paris, France

Abstract

The identification of the scientific production and the evaluation of the researchers is a problem we are facing at present times. Bibliometric methods use mainly statistical tools and their results are of quantitative nature. Consequently, this approach does not provide any tools of qualitative evaluation. It is for this reason we suggest to examine the specific details of the bibliographic references in the texts. This linguistic approach, which uses the contextual exploration method, allows us to annotate automatically the text and thus to propose a new way to address this problem. The computer application of this study will be integrated into the platform EXCOM (EXploration COntexuel Multilingue).

Introduction

Bibliometry is a quantitative evaluation of literature. Numerous studies contributed to the advance of this science and to the discovery of indicators allowing to estimate the productivity of a researcher, a country or an institution. Relational indicators help to determine the influence of the authors. They also allow to measure the relative rate of the scientific exchanges between two countries of comparable scientific mass, examples are present in (Esterle 2004). The theoretical approach of this tool will not be scrutinized, but we will show limitations of this model and will propose a new approach. At first, we will determine precisely how the quotations are used and propose a method capable to identify and extract them automatically. We consider a quotation to be an indicator which allows us to find linguistics clues. Subsequently we analyze these linguistic clues to propose a categorization. In the following, third phase, we formulate a proposition of annotation for this categorization and finally we implement this point of view in the platform EXCOM. The annotation of the corpus with the identified linguistic clues will allow us to apply the method of the contextual exploration. References are (Desclés 1997; 1991). We are capable to provide information about the nature of the citation and can therefore facilitate a qualitative rather than quantitative approach to this problem.

Bibliometry and scientific evaluation

Databases

A fashionable way to evaluate science is to use bibliometric studies. Therefore all studies can refer to bibliometric databases. Several databases were developed to provide access to bibliographical data. Others contain authors and citation data entries as Science Citation Index from Institute for Scientific Information or CiteSeer. ISI is the most cited and the most used for the following reasons:

- Homogeneity and coherence : databases are built on a logical and coherent structure of information.
- Multidisciplinary : they have a wide coverage allowing a panoramic study.
- Selectivity and complete perusal: databases contain the complete range of newspapers.
- Entry of the bibliographical references for every document. This is necessary for the study of the graphs and experts quotations.

The databases of the ISI, as well as the Science Citation Index, tend to offer to the community a beautiful tool. The evaluation of the researchers is calculated with the Impact Factor that we succinctly explain below.

Impact factor

It was invented by Eugene Garfield (Garfield 1955) from the profit-making Institute of Scientific Information and is a measure of the number of times a journal is quoted in references, for a limited period of time of two years. As a general rule, the journals with high impact factors are amongst the most prestigious. Garfield however said:

> The net result of all these variables is a conclusion that impact factors don't tell us as much as some people may think about the respective quality of the science that journals are publishing. Neither do most scientists judge journals using such statistics; they rely instead on their own assessment of what they actually read. None of this would really matter very much, were it not for the unhealthy reliance on impact factors by administrators and investigators' employers worldwide to assess the scientific quality of nations and institutions, and often even to judge individuals. There is no doubt that impact factors are here to stay.

The role of Impact Factor is important, but it does not measure the quality of the production of a scientist. It works only at an international level for country or institution, but should it be still used for an evaluation of an individual? As we said, this tool is the most used by the community but there are many limitations :

Limitations

We will concentrate on the ISI databases because they are the most often used and the most often quoted in studies and reports (Ricci 2003). The ISI databases contain the following limits.

- Negative citations: an author can be cited on a polemical subject or according to an error which he made.
- Auto-citation: Phenomena whose importance depends on the scientific domain.
- Linguistic impact: the English scientists are favored. It is important that the spelling of the names of authors are reported without errors. Polish names, for example, with their many consonants are incorrectly spelt.
- There is a delay between the publication of an article and a corresponding entry in the database.
- Many articles which are in the base are not cited.
- Coverage of newspapers: the selection of newspapers of the ISI covers the most important reviews, according to principles that some reviews can be representative of a domain.

Furthermore the base does not cover digital reviews. Beyond the technical limitations linked to databases, we shall emphasize that the practices of citations are questionable. For example, scientists in medical domain sign more papers than they can really contribute to, artificially increasing their citation record. The Impact factor model is imperfect because self-citation and lobbies are trained to cite only papers of colleagues and friends.

> Impact factor is not a perfect tool to measure the quality of articles but there is nothing better available. Impact factor has the advantage of being already in existence and is, therefore, a good technique for scientific evaluation. (Garfield 1994)

It is thus difficult to estimate the production of a researcher or a laboratory in view of the limitations that we enumerated. To propose a new approach, we wonder about the sense implied by the citations and their importance.

What is citation used for?

The purpose of the bibliography in science journals is to provide references to the imminent predecessors of the current work but it's also the most important factor to evaluate a scientist and a fashionable way to evaluate scientific studies. The number of times author's work has been cited by others remains a good indicator. It depends, of course, on the journals referenced in the databases. Citation analysis have been extensively discussed and many ways have been explored. Different fields of research are inter-citation which try to determine who cites whom. Of course, who is cited concurrently in the same bibliography is considered with co-citation. We have to understand that each approach uses measurement as normalized frequencies. We tried to reduce data into a two dimensional map or we used computationally intensive methods such as Pearson correlations methods such as Pearson correlations or Chi-sqares which are useful in the case of simultaneous occurrence of words or authors. All the studies use a statistical or mathematical approach based on data bases of the ISI or others. We can conclude that the existing bibliometric tools are not satisfactory. Citations are the best means to estimate the importance of articles but we are limited by a set of technical and human factors. Furthermore, we have no qualitative evaluation with regard to quotation. This can be achieved in part by stating that we need a new tool to evaluate the scientific production. The goal is to present a new method using Contextual Exploration. We want to demonstrate that a linguistic approach can provide relevant information.

New approach

Hypothesis

Our hypothesis is that by locating the indicators in corpus we will provide sentences localization with linguistics clues. These specific linguistic units, using the method of contextual exploration, give us opportunities to annotate articles with information about citation. This approach does not depend on the specific domain of articles. In fact, we dont need knowledge representation. This point is important because we can use this approach for all scientific domains.

Protocol

To begin with, we have to constitute a corpus of scientific publications of different length to identify the textual marks which imply a link in the bibliography. We call these marks indicators. They allow us to determine the place where the contextual exploration rules must be used. Then, we concentrated our effort on the study of some publications in French language. This corpus of scientific articles allowed us to identify the linguistics clues and elaborate a categorization. Finally, from this categorization, we can annotate the text which can then be processed with the EXCOM platform.

Bibliographic entries and indicators

Using by authors and editors in the formulation of citations within the text corresponding to the entries in that bibliography, ISO 690 does not apply to full bibliographic descriptions. Furthermore, the application of these rules is not always respected by authors. In fact, we found different ways to present the references depending on the newspapers as well as the authors. Following the protocol, we are looking at corpus to identify the type of quotation which is used. We have identified five categories but many combinations are possible. Below, we have enumerated the different types we encountered in corpus.

1. Numerical Type. [N] The Numerical Type is frequently used. There are two ways of using it: Quotation by order

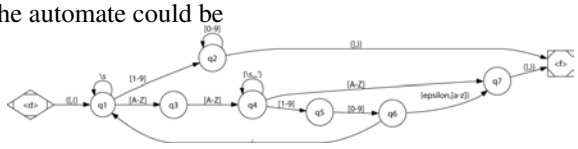
of appearance in the text or quotation by reference number quotation. The number is generally placed between square bracket like this: [1]

2. Condensed Type. [AUT99] The Condensed Type is a combination of the three first letters of the name and the last two numbers of the publications date of article. For example, we can have : [DES90], [WIN 97], [LER 94,BER 96a] or [BER 96b]. Note that apostrophe can be used, too[AUT99].
3. Et al. Type. [Author and al.] Et al. is an abbreviation of et alii, meaning and others. It is ordinarily used in lieu of listing all names of persons involved in a proceeding. If we have more than three names, we must use this form. For example, we have found:(Authoretal., 2003), (Author et al., 1997, Author et al., 1999), or author et al. (1998). Note for this sentence, that () are only used for date. Author is outside of the brackets.
4. Normal Type. [Author, 2000] This form has a big number of combinations. For example (Author1, 2000; Author2, 2000a) or for the same author, we can have Author (1968, 1976, 1982).We can notice, too, this form: Author (1958: 274-275).
5. Literal Type. [Author, Year, Page] Literal type includes the page number: (Author1, 1944, p.311 ; Author2, 1956, p.316).

We must be able to identify, annotate and extract automatically all these forms to identify the sentences that are likely to contain linguistics clues. So, one way is to use deterministic state automata. We considered an automate to be able to identify any type of entry, including unusual and original entry.

Deterministic state automata

To extract these indicators, the best way is to use deterministic state automata. The computer application will be in Perl with Regular expressions. For each regular expression there is at least one finite automat, which accepts all the words suitable for the regular expression and ends in its accept end state. We denote an FA by the 5 tuple (S, Q, d, qo, F), where S is an alphabet, Q is a set of states, q0 is the starting state, d is a Transition function, and F is the set of accepting states. The automate could be



From this automate, most entries, called indicators can be written with a deterministic state automate. The extraction of indicators can be done with regular expression. Deterministic state automata will allow to evaluate the number of bibliographical references. We denote RefB the number of bibliographic index and RenB the number of entries. The relation

$$RefB > RenB$$

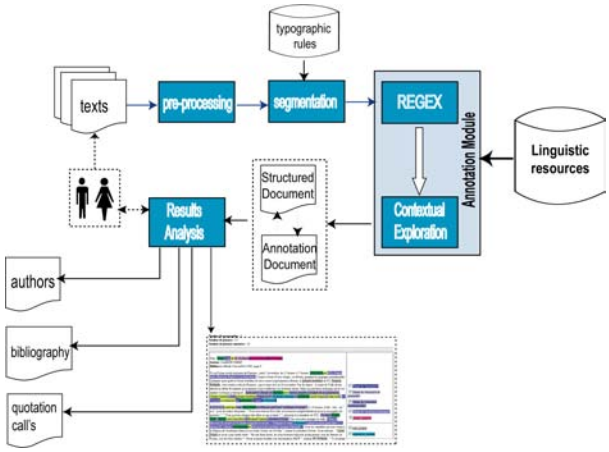
shows us that the document presents a gap and the relation

$$RenB \geq RefB$$

shows us that the document should be well formed. For the last case, we will research, for example, the distribution of entries through the corpus. An author with a large number of entries will be more significant than an author with few entries. But only categorization, possibly with help of some sort of qualifying coefficient, can define the nature of relation between authors, according to our purposes.

Corpus

For this study, we constituted a corpus based on articles of the laboratory LaLICC, the scientific articles stemming from databases of the INRIA, as well as from articles of the review INTELLECTICA. First of all, it should be noted that this is a small heterogenous corpus with potential to deploy a substantial variety of methods to process bibliographic references. Furthermore the corpus covers various domains, to include Cognitive science, Databases, XML, Linguistics, in order to demonstrate its capacity in the multi-domain context. At present the corpus is exclusively constituted by text in French. This corpus has been used as the base to develop our automata. For the practical implementation we decided to deploy the language Perl to process our corpus. The use of the regular Expessions is going to allow us to process the textual data. In fact, with regular expression, we retrieve entry from corpus and annotate citation in the corpus. Bibliography could be extracted by the same method, using regular expression. Authors were identified according to their types between the bibliography and entry. So matching between entry and bibliography gives us the possibilities to extract Author. This information will be used later by Contextual Exploration Method. As a result, we obtain an annotated corpus with entry and bibliography data. The annotation of the corpus is a simple tag allowing us to localize the part of the text where we can apply the rules of the contextual exploration. This is a visual explanation of informatics application for extraction and annotation.



Linguistic approach

Contextual Exploration Method

Contextual Exploration, proposed and developed by Jean-Pierre Desclés and LaLICC group, is based upon the observation that it is possible to identify specific semantic information contained in certain parts of text.

1. Indicator. It must be taken into account to determine an analyzed textual unit. This study will show how the indicators reflect the bibliographic referencies. In this example, the indicator is 'Clarke et al.'.

   > A similar way for defining texts was proposed by Clarke et al. in their paper An algebra for Structured Text Search

   Indicator is not always placed into () or []. It can also be retrieved by our state automate. 'et al' that is giving us the precise information. Sometimes, we only have the name of the author, therefore, we need to extract author's name from the corpus. The author's name will always be present in the bibliography. It is thus not necessary to maintain the list of named entities. We will annotate in the initial phase the name as an indicator in order to avoid this ambiguity.

2. Linguistic clues.

   The linguistics clues must be taken into account in order to determine specific semantic information. They are the only knowledge we need to create categories. Linguistics clues can be found around the indicator, in the same segment of text.

3. Localization of linguistics clues.

   It is necessary to know the location of the linguistics clues in the sentence in relation to the indicator because localization is a very important factor for contextual exploration. There are five possibilities to designate different localization in phrases. It can be the first word, before the middle, in the middle, after the middle, and at the end of the phrase. A simple example how linguistic clues can be used:

   > A similar way for defining texts was proposed by Clarke et al. in their paper An algebra for Structured Text Search

   In the above sentence 'Clarke et al.' is the Indicator and 'was proposed by'is the linguistic clue. Its position is in front of the Indicator. The position has a key importance. We will see at the end of this article how this information is treated during annotation process. The following attributes can be used for annotation:

   Beginning |before |middle |after |end

Categories

Quotation is divided in five categories.

1. Point of view : The first category is the point of view. It is used extensively in the corpus. Using the point of view, we can assert our opinion on the question. The following linguistics clues allow us to classify a quotation in the category of point of view:

   Selon |d'après |pour |considérer que |nous y voyons |comme le dit |...

   They are easy to track down and are located in front of the indicator which activates the extraction of the phrases.

2. Comparison : The second category which we are interested in is the comparison. We often compare the works of the researchers. If the comparison is neutral, we have to use the contextual exploration to determine if there is a case of resemblance or disparity. For resemblance, we have the following linguistics clues :

   ressembler |comme dans les travaux de |le rapport avec |...

   For disparity, we have the following linguistics clues :

   différer de |contraire l'approche de |contrairement ce qu'affirme |...

3. Information : The category of information is immense. It includes subcategories such as the hypothesis, the analysis and the result. For analysis, we have the following linguistics clues : **a été analyse dans |l'analyse de |lors de son analyse |...**

   For result, we have these linguistics clues :

   nous avons démontré |donner de nombreux exemples de |a publié ses résultats |a dégagé |...

4. Definition : The sentences which contain definitions are important. For result, we have the following linguistics clues:

   ils caractérisent |la notion ... introduite dans |...

5. Appreciation : In this category, the author gives judgement in a positive or negative way about another author. For Appreciation, we have the following linguistics clues:

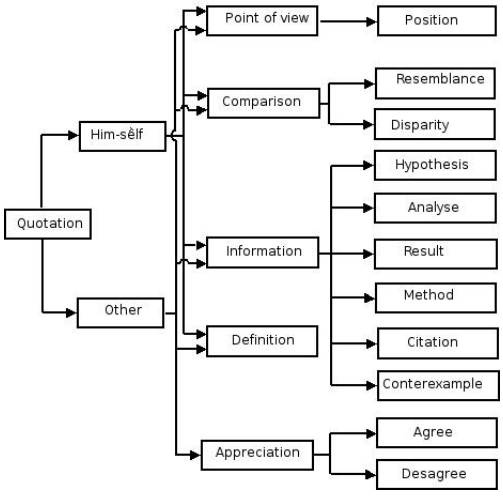
   ont rejeté |n'as pas répondu |en trahissant |...

   sérieusement notre proposition.

After this first step of identification, it will be necessary to widen the contents of the categories through the study of new publications and synonyms.

Representation of Categories

For the tree Structure of Quotation categories see diagram below. We have to consider who cites whom (him-self or others) and the use of the right rules for contextual exploration. (Krushkov 2004 2005)



We can notice that there are no subcategories appreciation for "himself".

Annotation

Formalism

Annotation can be used to describe this approach. It is an example of syntaxes.

ANNOTATION : PARAMETERS

We obtain the following annotation and parameters for categories :

1. Annotation

   ANNOTATION::= Point of view |Comparison |Information |Definition |Appreciation |Position |Resemblance |Disparity |Hypothesis |Citation |Analyse |Result |Method |Counterexample |Agree |Disagree.

2. Parameter.

   There are three different parameters. PARAMETERS::=(enonciator = (himself |other;) @ ((localization = (Beginning |Before |Middle |After |End;))

EXCOM platform

A major objective for EXCOM system is to explore the semantics of text for enhancing information extraction and retrieval through automatic annotation of semantic relations. Most annotation systems are based on linguistic focus on morphological analysis, part-of-speech tagging, chunking, and dependency structure analysis. The methodology used by EXCOM, called Contextual Exploration, describes the discursive organisation for the text with calling upon exclusively linguistic knowledge present in texts. Linguistic knowledge is structured in form of lists of linguistic marks and declarative rules. The constitution of this linguistic knowledge is independent of a particular domain. Linguistic rules for identifying and annotation of semantic segments are based on different organisations of the text. Some of them use lists of simple patterns expressed by regular expressions, others need to identify structures like titles, section, paragraphs and sentences for extraction strategies. The most relevant rules for EXCOM are those called "Contextual Exploration" rules. A process of such rule is a complex algorithm which leans on a prime textual mark (called indicator) and secondary phrases whose function is resumed in confirm or invalidate the semantics carried by the indicator.

The next step is to include this approach in EXCOM platform, developed by LaLICC laboratory, which is an engine for automatic annotation of textual corpus. This new system uses essentially new technologies around XML and a programming language Perl. It is centered around the search for semantic information on the closed contents - the documentary bases and on the opened contents as Web. We use rules for filtering textual segments of several semantic categorizations relevant to the identification of the relations of matching entities names with references to temporal and spatial expressions as a matter of fact to provide the answer to the question Which is in connection with Whom? Where? And When? In every task of semantic annotation there is an associated set of linguistic tags (lists of indicators and clues) and a set of rules. The conditions of release of these rules are expressed in specific ways which activate certain levels of the engine of annotation. Every level corresponds to a general algorithm of functioning. These algorithms are expressed in specific ways which activate certain levels. They are going to activate different mechanisms. The engine is designed so that rules of EC can use more elementary levels as the regular expressions or the structures. It is recommended that designers of rules order them in a way so that rules are activated at the lowest and at the highest level. Several levels of annotation are foreseen for the engine of semantic annotation. We find the following levels there:

1. Regex: regular expressions with lists of tags with algebraic operators. This first module is a module of annotation of low level which allows the recognition of some named entities, spatial and temporal expressions, finally more complex textual expressions which can be described by a finite state machine. In this first layer, we add the possibility of using lists of tags and regular expressions as well as algebraic operators*,+and? In this context we developed a language of expression of the rules of this level. Each of the modules of annotation thus has a language of representation of rules and an associated compiler.
2. Structure: use of annotations as linguistic markers. The Hierarchy in levels of the annotation modules gives us the possibility to use the conditions of release of the rules of the textual segments already annotated by information which can be of various nature. This module thus imposes on the compiler of the alterations to be strictly linked to the structural segmentation of documents. The annotation engine has to have the possibility to reach any segment of the processed document. The engine should not be triggered either by the number of annotations or by the identity of these annotations.
3. Contextual Exploration: CE comes to conclude the third and the most important module of annotation. It is in this module that our proposition shows how to extend the XML technology and how it can be used in a system of semantic annotation of texts. The target language of the associated compiler is XSLT and the annotation engine is a processor XSLT. For a better flexibility of processing and an easy integration with the other modules, it is recommended to link the processor XSLT with a language host such as Perl or Java.

Conclusion

We have thus demonstrated that it was possible to identify and to annotate the textual segments from bibliography. Furthermore, through this linguistic study of the textual segments, we identified and categorized linguistics clues. The annotation of these linguistics clues facilitates automated processing within the frame of the contextual exploration method implemented in the platform EXCOM. This phase allows us to qualify the relations between the author, the coauthors and also the bibliography. It is possible to classify citation according to a qualitative approach. The categorization leans on linguistic elements and also the evaluation which is more suggestive, we need linguistic approach to determine a protocol for evaluation. In the future phase we will increase the number of articles in order to have more

consequent corpus, this will for example enable a statistic approach to the distribution of different categories in texts. Furthermore, we will resolve the limitation of statistical method, related to a possible distortion resulting from the negative citation of authors. This method requires neither access to databases nor knowledge representation and can be applied on any scale whether it is a considerable volume of articles, a single publication, or documentation set of a laboratory. It is offering a better use of the bibliography.

References

Desclés, J. P. 1991. Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. *Knowledge modeling and expertise transfert* 371–400.

Desclés, J. P. 1997. Système d'exploration contextuelle. *Co-texte et calcul du sens* 215–232.

Esterle, Laurence, G. F. 2004. Indicateur des sciences et techniques : Rapport de l'observatoire des sciences et des techniques. Technical report, OST.

Garfield, E. 1955. Citation indexes to science: a new dimension in documentation through association of ideas. *Science* 122:108–11.

Garfield, E. 1994. The impact factor. *Current Contents*.

Krushkov, Y. 2004-2005. L'exploration contextuelle des appariements entre les références bibliographiques et les passages textuels dans un corpus de textes linguistiques. Master's thesis, Université Paris IV Sorbonne.

Ricci, J. F. 2003. Indicateurs bibliométriques. Technical report, Système d'information des Hautes Ecoles suisses.