Citations Titles Standardization Using Information Retrieval Techniques

Rogério Mugnaini, Esteban Fernandez Tuesta, Adalberto Otranto Tardelli

OMS/OPS/BIREME
Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde.
Rua Botucatu, 862 - 04023-901 São Paulo – Brasil
rogerio@bireme.br

Abstract

Citation analysis is an impossible method in many bibliographic databases. Few peripherical countries have worked with this type of data, using for this purpose, the information provided by ISI. The Brazilian database from SciELO Project presents near 475 thousands citations to scientific journals articles. Considering that citations demand several standardization work on titles, this paper presents a methodology, based on information retrieval, for standardization though ISSN database search. The result is the standardization of approximately 85% of the titles.

Keywords: Bibliometrics, Textual Statistics, Information Retrieval, Citation, Similarity, Brazilian, SciELO Project.

1. Introduction

Many questions related to the utilization of the ISI's (Institute for Scientific Information) databases for scientific production (? output) study of peripheral countries and also for European countries, above all the non-English-speaking, can be found in the literature, in the first place because the 'science' of the first ones cannot be considered mainstream.

Two postulates are in the basis of the bibliometrical analysis techniques (Rostaing, 1996): first, "the publication is, in a scientific context, a representation of the scientific activity of its author;" second, "exists relationship among the ideas of "the proper author and the ideas acquired from the work of its peers."

The citation analysis is a bibliometrical technique that makes sense, according to those postulates, within Merton's paradigm, in that the development of science is given according to own internal logic, objectifying to generate new knowledge; science whose product is reflected in written scientific communication, particularly in the scientific journals.

The impact factor has been quite questioned within the reasons raised above. The few journals indicated in the ISI's databases of countries as Brazil, have no sufficient visibility with regard to the Americans, for example. Certainly the impact of those journals, and of the large majority whose visibility is restricted to the national libraries' shelves are restricted to the local journals.

The citation analysis becomes increasingly important, as it is adopted as quality indicator of scientific journals. Being utilized for the quantification of the impact factor, measure created by the ISI since the 1960s (Garfield, 1994) is widespread among the science and technology's systems of many countries of the world.

Within the bibliographic databases owners, a minority has this type of data, whose analysis always requires a severe standardization treatment. This work aims to present a methodology to obtain a satisfactory level of correction and standardization, utilizing a information retrieval based process (Tardelli, 2003)[1], that is given through previous treatment of the cited titles and through a standardization database.

2. SciELO Project[2]

SciELO (Scientific Electronic Library Online) project is the result of the efforts of BIREME (Latin American and Caribbean Health Sciences Information Center) and of the FAPESP (State of São Paulo Research Foundation), and envisages the development of a common methodology for the preparation, storage, dissemination and evaluation of scientific literature in electronic format. It is contains currently a total of 114 journals, summing approximately 32 thousand articles, and around 475 thousand citations to journal articles, whose titles, tabulated, sum approximately 70 thousand variations.

It has their databases built in CDS/ISIS[3] (Computerized Information Service/ Integrated Scientific Information System) standard, which permits the establishment of relationships between register of a database or between records of different databases. Operations are done with CISIS[4] utilities, programs that permit the administration, manipulation and realization of operations (many of them nonexistent in the conventional versions of Microisis - CDS/ISIS 3.07 and WinISIS 1.31) between several CDS/ISIS databases.

With the ISSN database in the same standard, which can serve as reference for standardization of the cited titles, some procedures are sequentially adopted for the constitution of a titles standardization database. This standardization work becomes necessary due to the lack of uniformity in the writing of the titles of cited journals in the bibliographic reference set of the indexed articles in the database.

3. Process of standardization

A flow chart of the standardization process can be observed in the following figure.

Firstly, the title is sought in the form that was cited (V1), in the standardization database (if found, Y1). Not finding, this title will be sought directly in the ISSN basis (if found, Y2) and if it still was not found will be submitted to small substitutions of point and space (V1*) and sought in the basis ISSN (if found, Y3).

The errors of writing (digitations) are the leading cause of the non-uniformity of cited titles; in addition, abridgements are very common, since many titles are known by the abbreviated title; despite this, the most frequent citations have written titles (complete or abbreviated) correctly. The correction of titles keyed in incorrectly can be given utilizing a tool (Trigrams) that associates and orders expressions per the degree of similarity among them. Thus, the title cited incorrectly, with smaller frequency, can be replaced by a similar written form (V2), however more utilized (if it had), since this last tends to be more correct. Then this title will be submitted to the basis of standardization (if found, Y4) and if not found, continues the searches in the basis ISSN (Y5 and Y6), just like in stages Y2 and Y3.

[1] Available at: < http://www.crics.info/reuniao_bvs3/program/docs/es/1 >. Access in: August, 5 2003.
[2] Available at: < http://www.scielo.org >. Access in: August, 5 2003.
[3] Available at: < http://www.unesco.org/webworld/isis >. Access in: August, 5 2003.
[4] Available at: < http://productos.bvsalud.org/html/pt/home.html >. Access in: August, 5 2003.

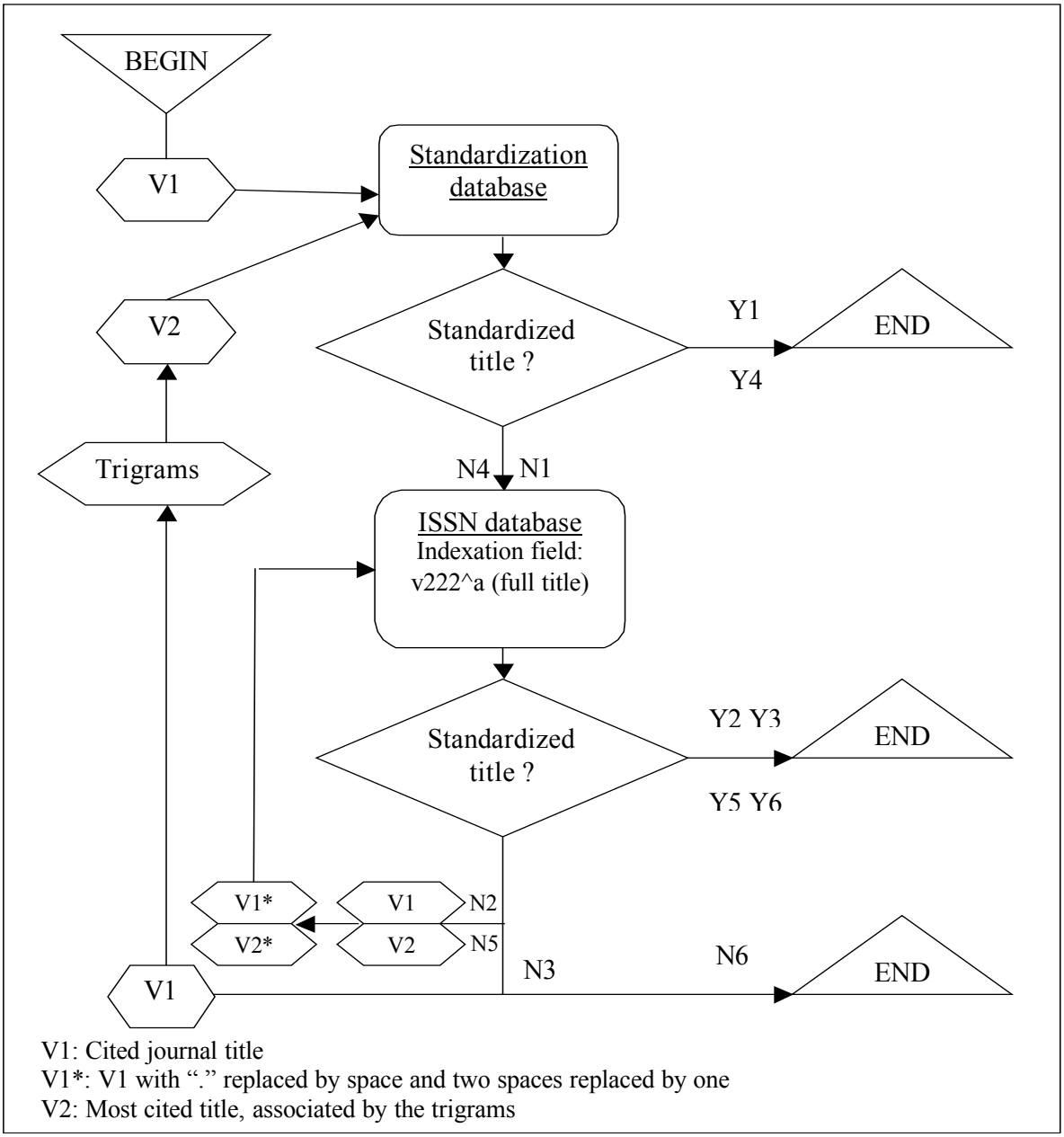

Figure 1. Flow chart of the standardization process

A description of the stages of the standardization process is presented in the above subsections.

3.1. *Standardization database*

3.1.1. *First stage*

The first stage consists of a search of full titles in the ISSN database. That search was defined in order to retrieve, utilizing the title of the journal cited in SciELO database, the corresponding title in the ISSN database. To this end, were defined firstly, the fields of the ISSN database that would turn "investigable", through a Fields Selection Table - FST.

A database in the CDS/ISIS standard contains a set of files that composes it, within which, the Inverted File. This file contains all the terms that can be used as points of access during the retriever of the data, and for each term, a reference list of the records of a Master File from

which the term was extracted. The collection of all the access points for the database is called of "Dictionary". The Inverted File can be thought as a remissive index of the Master File, where are defined the investigable elements of the database, through the Fields Selection Table, that contains the fields to be inverted, and the technique of indexing to be used for every each. There are 5 indexing techniques identified by numerical codes: ø to 4. In this stage it was utilized the technique ø, that constructs an element of search based on every line extracted from the format (information inside one or more fields, with optional adding of any string). Used to index entire fields or sub fields with up to 60 characters.

The fields of the ISSN database selected for realization of the search were:

210		**Abbreviated key title**
	&a	Abbreviated key title
	&b	Abbreviated qualifying information distinguishing otherwise identical key titles
	&c	Abbreviated qualifying information distinguishing otherwise identical abbreviated key titles
222		**Key title**
	&a	Key title
	&b	Added qualifying information distinguishing otherwise identical key titles
245		**Title proper**
	&a	Title proper or common title part of the title proper
	&s	Section, sub series, supplement designation
	&u	Section, sub series, supplement title
246		**Variant title(s)**
		Access to portions of titles and developed forms of key title
		Parallel titles
		Other forms of title not specified
		Cover title
	&a	Title

Source: http://www.issn.org:8080/English/pub/tools/format

The content of each of those fields was utilized combining the contents of its sub fields and making substitutions of point by space, and two spaces by one. In that way, a record of the ISSN basis whose search fields are presented as follows:

```
210  "  ^aTherapia^bHels.^cSuom. p."
222  " 0^aTherapia^bHelsinki"
245  " 0^aTherapia"
246  " 3^aAstra therapia"
246  " 3^aAstraZeneca therapia"
```

It will be represented by the terms of access as follows:

```
THERAPIA
THERAPIA HELS
THERAPIA SUOM P
THERAPIA HELS SUOM P
THERAPIA HELSINKI
ASTRA THERAPIA
ASTRAZENECA THERAPIA
```

The search expression, at this stage, utilizes the cited title, complete, that can be associated to more than one record of the ISSN database. At this time single relationships between cited title and title of ISSN database are barely retrieved.

The citation with journal title "THERAPIA", matches with the string "Therapia" (ISSN database) of Helsinki, finds, in the basis ISSN also the editions, with equal title, of Buenos Aires, Barcelona and Bratislava. It is concluded that an incomplete citation is with difficulty identified, and as much fewer words in the title, greater is the chance of ambiguity.

The resulting database presents a total of 17,175 single relations, between cited title and title in the ISSN, representing approximately 24.3% of the different cited titles, which covers a total of 194,866 citations (41.1%).

3.1.2. *Second stage*

In the second stage the search is carried out utilizing the words of the cited title, instead of the complete expression. This type of indexing can be obtained with technique 4, that constructs an search element based on each sub field, considering as search element any sequence of alphanumeric characters. The Fields Selection Table of ISSN database in that stage, barely utilizes the sub field "a" of field 222, indexed word-to-word.

The citation made to the abbreviated title "CAN J ANIM SCI", will be used as expression of search, having each one of the four words any termination ("$"), and separated by Boolean operator "and", turning out: "CAN$ * J$ * ANIM$ * SCI$", that retrieves the title "Canadian journal of animal science" of the ISSN database.

The resulting database from this stage presents a total of 18,697 single relations, between cited title and title in the ISSN. However, of those relations 10,033 are exclusive with the relations of the first stage, representing then, around 14.2% of the different cited titles, which covers a total of 24,203 citations (5.1%).

3.2. *Correction Process*

3.2.1. *Similarity indexes*

The search for relative similarity between documents, was utilized the article written by G. Salton and C. Buckley (1988). The words that compose a document, are extracted, and placed in order to form a vector,

$$D = (t_0, w(d,0); t_1, w(d,1); \ldots; t_t, w(d,t))$$

where each t_i represents a word that composes the D document and each $w(d,i)$ represents the weight of the word t_i in the D document. Similarly, the required information, or "Query" is represented in vector form as

$$Q = (q_0, w(q,0); q_1, w(q,1); \ldots; q_t, w(q,t))$$

where each q_i it represents a word of the "query" Q and each $w(q,i)$ represents the weight of that word in the "query".

When a standardized size is used, we obtain the well-known similarity formulae

$$sim(Q,D) = \frac{\sum_{k=1}^{t} w(q,k) * w(d,k)}{\sqrt{\sum_{k=1}^{t} w(q,k)^2} * \sqrt{\sum_{k=1}^{t} w(d,k)^2}}$$

Within the objectives of the "National Library of Medicine", of automatic methods of indexing "indexing initiative", the work written by Aronson *et al.* (1999) presents a section devoted to "trigrams". This method consists of the construction of vectors, not barely as being words, but, sets of three anagrams of each word, establishing weights for each "trigram" depending on its location in the word.

The results of the similarity between Q and D will remain sorted descendant, in such a way that among the first ones, will be those "queries" more similar to the documents.

4. Results obtained

An amount of approximately 85% of the titles was standardized, being distributed as the following table. The different moments of the process of standardization share individually for the conversion of the titles, being able to observe which of them are the more significant.

Stages	Frequency	Percentage
Y1	214,888	*46.6*
Y2	171,803	*37.3*
Y3	2,403	*0.5*
Y4	7,798	*1.7*
Y5	6,434	*1.4*
Y6	68	*0.0*
No standardized	57,586	*12.5*
	460,980	***100.0***

Y1 and Y4b are the stages where the titles found are perfectly identified in the ISSN database, amounting to a total of approximately 50% of the citations. In the other four stages they are already retrieved more than one title, not being able to be aimed with precision the title cited for lack of information of the citing author.

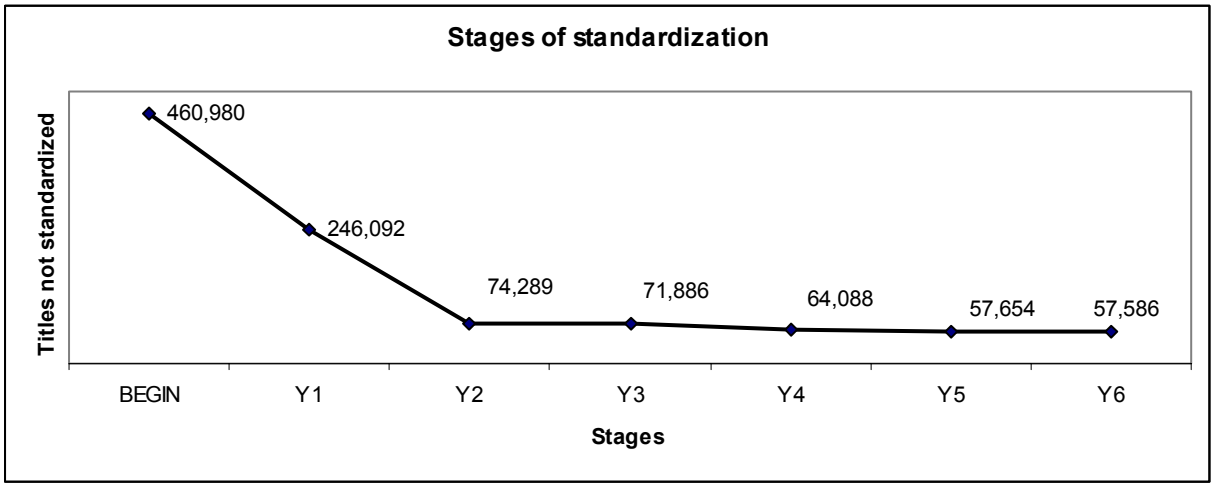

Figure 2. Stages of standardization

The first two stages have principal role in all the process of standardization, since they represent almost 96% of the standardized total.

References

Aronson A.R., Bodenreider O., Chang H.F, Humphrey S.M., Mork J.G, Nelson S.J., Rindflesch T.C. and Wilbur W.J. (1999). *The indexing initiative*. A report to the Borrad of Scientific Counselors of the Lister Hill Nacional Center for Biomedical Communications.

Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindflesch T.C. and Wilbur W.J. (2000). The NLM indexing initiative. In *2000 AMIA Annual Fall Symposium*: 17-21.

Garfield E. (1995). Quantitative analysis of the scientific literature and its implications for science policymaking in Latin America and the Caribbean. *Bulletin of PAHO*, vol. (29): 87-95.

Humphrey S.M., Rindflesch T.C. and Aronson A.R. (2000). *Automatic Indexing by discipline and high-level categories: methodology and potential applications.*

Rostaing H. (1996). *La bibliométrie et ses techniques*. Sciences de la Société.

Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Procesing & Management*, vol. (24): 513-523.

Tardelli A.O. (2003). Indización automática y "vector mining": herramientas para recuperación y vinculación de información en las fuentes de información de la BVS. In *3a Reunión de Coordinación Regional de la BVS*.