# A new approach to bibliometrics based on semantic similarity of scientific papers

**Sorin Avram*, Dan Caragea**, Ioan Dumitrache*****

** University Politehnica of Bucharest, 060042 Romania*
*(e-mail: sorin.avram@uefiscdi.ro).*
*** Executive Agency for Higher Education, Research, Development and Innovation Funding, Bucharest, 010362, Romania (e-mail:dan.caragea@uefiscdi.ro)*
**** University Politehnica of Bucharest, 060042 Romania (e-mail: idumitrache@ics.pub.ro).*

**Abstract:** During the past decades, the deficiency of support tools for scientific research assessment policies led to applications development, focused on integrating citation complementary information: online usage statistics (webometrics/cybermetrics) and content analysis. However, due to lack of effective technical solutions, bibliometric data suppliers have refused to promote alternatives to the renowned and worn impact factor, with all its derivatives.

The present research unveils enhanced bibliometrical tools, meeting comunity's needs and proposing an automatic solution to evaluate the scientific relevance of a research article, in a particular research field, in relation with all the citing works. This process resorts to semantic processing and conceptual similarity analysis, proved to be superior to solutions already abandoned, based on lexical analysis.

*Keywords:* bibliometrics, semantic analysis, conceptual similarity, citation, semantic reference, article relevance factor, index comparison

## 1. INTRODUCTION

Nowadays, everybody agrees on the utility of measuring the prestige of scientific journals, whether it is paper printed or hosted in a famous publisher's online database. Initially generated by the publishers, the interest for such measurements has gradually occupied the authors perspective, who now strive to publish their articles in the most widely read journals, to achieve top visibility, scientific recognition and financial support for their activities.

Bibliometrics provided the answer, allowing us to recognize the most *popular* journal in a certain knowledge field. As "prestige" or "importance" are usually *subjective* conclusions, a first *objective* solution was based on citations, which led, of course, to a hierarchical list. In other words, the most cited journal is objectively, also the most prestigious. Therefore, the market principle was to extract the qualitative information, inferring the quantitative. Evolving from its initial purpose, bibliometrics has been focused on measuring articles and authors prestige. As everybody agreed, not all the articles published in a journal shared the same scientific interest, visibility and eventually, relevance. Therefore, using the same metrics, we can now recognize the most "important" researcher of all the authors who published in a research field or in the same journal.

Even though the metrics of the main bibliometric indices haven't changed in the last fifty years, the number of citations recorded annually has increased continuously, having the support of the bibliographic management tools that allowed a far greater number of references (Friedberg (2010)). This is explained by the authors' tendency to cite peer-reviewers whenever possible, thereby increasing the chances of publication for this article. Also, the current trend requires authors to use an extensive bibliography, in order to illustrate or simulate an exhaustive knowledge of the research field.

Perhaps the most relevant argument of the overwhelming increase of citations was the introduction of impact factors in the institutional or individual assessments, a process that stopped in 2010, when, United Kingdom, U.S.A and Germany excluded the metric methodologies due to the manipulation of impact factors (Deutsche Forschungsgemeinschaft (2010), Houses of Parliament (2004)).

In the latest years, the worn down impact factor (IF) has been even more criticized because of its incompleteness and shallowness, forcing the world's largest publishers to supplant the IF with other bibliometric indexes: *h-index (HI), SCImago Journal Rank (SJR), Source Normalized Impact per Paper (SNIP), Eigenfactor (EF), Article Influence Score (AIS), Immediacy Index (II)* and *Cited Half-Life (CHL)* (Friedberg, E.C. (2010)). All these metrics have proved ineffective and were soon exhausted by the scientific community due to a very simple reason: they were all developed on the same sixty year old footing - the citation.

Before presenting the method and implementation proposals, we have to admit that the actual technology is advanced enough to host the world's publications. For instance, the

entire archive of Elsevier, between 1823 and 2012, has about 7 TBytes of information (Elsevier (2002)), considering that the Dutch company has published more than 25% of the world's scientific literature (Elsevier (2010)).

This paper addresses the reliability of bibliometric indexes in terms of academic public and policy makers' expectations, underlining the pros and cons of the existing solutions and their proved usability. Considering the opportunities and the importance of a straightforward automated process of article assessment, the research is focused on the actual techniques of citation weighting.

In the second section, it is presented a viable bibliometric solution based on the analysis of the scientific discourse, by measuring the scientific impact through semantic processing of research publications. Starting from the theory of Baeza-Yates and Ribeiro-Neto published in 1999, a new process model for scientific impact measurement was developed, based on conceptual structure of scientific papers instead of the lexical mass (Baeza-Yates and Ribeiro-Neto (1999)).

The result of this research is a new metric for article relevance scoring and journal ranking, which integrates the PageRank algorithm of Brin and Page with the conceptual similarity framework. Conclusively, a comparative analysis with the three most popular bibliometric indices presents the advantages and downsides of the new article based metric.

## 2. CITATION RELEVANCE

Unlike the quote, which is a takeover of a small piece of text, citing a source implies posting the reference, which normally is found in the bibliography section. Citing always meets the stylistic conventions of the publisher, being, formally, an alphanumeric expression. The quotation and the citation form the article's referential universe, the obvious manifestation of intertextuality.

Unfortunately, publishers and aggregators exclude all other information besides article bibliography when processing the journal or article metrics. The effective presence of the quote, citation size, frequency and relevance are not taken into account in the design of the newly promoted indices. Setting intertextuality between articles assumes full text, citations and quotations automatic analysis, altogether with references list.

If citation analysis had been a simple computation, similar to a peer-review process, bibliometrics development might have been narrowed to our actual status. However, the current stage of technological development imposes no such constraints, including citation and other complementary, more or less arbitrary elements in the construction of new bibliometric indicators. Up to now, bibliometric indicators have provided a metric representation of science, neglecting the scientific content of the text and its meanings. Instead, they used the number of citations as a measure of relevance, generating new confusion or disagreement in the scientific community (Woters (1999)).

### *2.1 Impact factor, publishers and open-access trends*

In 1955, Eugene Garfield, the world's most renowned researcher in information science, defined the "impact factor" as the number of citations received by a certain journal, in a certain time, divided by the number of articles published. The processing method has been perfected over time to include only "citable items", representing only a few types of documents that have proven to be frequently cited by the scientific community (Smith and Rivett (2009)).

Currently, most researchers refuse to publish in journals without impact factor, having enough arguments to support their decision. First of all, universities have started to distribute funds based on the number of publications of an author and their corresponding impact factors. In such circumstances, the decision to publish two complementary articles instead of just a comprehensive one has an obvious advantage for the research team, as well as the selection of the forthcoming publication journal. Another major repercussion of the impact factor influence is the dependence of all individual and institutional assessments on impact factors. If this dependence between scientists and journals' impact factors may have a positive impact on research performance, the question of a further free, non-constrained scientific advance is a difficult thing to prove (Feller (2010)). The current debate on the bibliometrics' usability is now focused on the relation between research quality and publication quantity: is this trend efficient in promoting science or in increasing papers production?

On the other hand, the open-access trend, which permanently increases its share in academic publishing, is not limited by any of the printed journal rigors. The limited space of each issue, scheduled publication and printing dependencies are no longer a constraint of modern publications. If information becomes available through open-access databases such as PubMed, J-Stage, Hispanic or CERN Document Server, the scientific article is no longer limited by the publication, but rather by the content and its writing.

As Stephan M. Feller asks, given the potential decrease of the IF influence, is there an option to replace it with an article impact factor? The solution of using online statistics (number of downloads / number of views) is not really an improved measure of quality and relevance of scientific papers. It is very unlikely that all articles published during a year, in the same journal, will have the same number of citations; usually, their distribution is Gaussian (Feller (2010)).

### *2.2 Article Influence Score, Eigen Factor*

The article influence has been developed in the recent years using new algorithms for weighting the relationship between articles published in magazines. Using the principle of Google page-rank, developed to evaluate websites, the algorithm uses the entire network of citations, weighting each network node according to the number of citations taken, but also the relevance of the articles quoted. It seems that the

*Eigen Factor* has taken the lead from the journal impact factor, with greater accuracy and the use of the entire universe of citations.

Derived from the *Eigen Factor*, the *Article Influence Score (AIS)* is weighting its values with the relative frequency of publication (number of articles published in journals divided by the total number of articles published during the calculation) (West (2010)). We can easily observe that AIS is an average indicator of the journal, regardless of quality or relevance of the scientific articles and the quoted works.

### *2.3 Article content analysis*

In the past decades, research on the information relevance assessment (infometrics) was focused on three main directions:

- citation analysis - the introduction of new indicators (bibliometrics);
- online usage statistics analysis (webometrics/cybermetrics);
- full-text analysis, having a lower influence because of the lack of standardized tools and open full-text databases (content analysis / textual analysis).

In most cases, researches were carried out separately, bibliomining being one of the few solutions that integrate *cybermetrics* and *bibliometrics* to assess the scientific relevance using online visibility. Content analysis focused on the development of hybrids algorithms, that integrate tools such as the *terms similarity matrix*, *in-links* and *out-links*, with spectral or k-means clustering algorithms, research being in an early stage.

However, recent years have brought new opportunities by integrating semantic analysis in an increasing number of research areas. Its integration with bibliometrics started in 2000 and had a high dependence with the development of information processing technologies: *data mining*.

## 3. USING SEMANTIC ANALYSIS IN CITATION WEIGHTING

In 1991, Braam, Moed and Van RAAN have introduced the concept of combining *text analysis* with *bibliometrics* to improve clustering efficiency of scientific production (Braam et al. (1991)). Further developed in 1999, Baeza-Yates and Ribeiro-Neto have introduced the hypothesis that a high degree of similarity between a cited article and a citing one will induce an increased relevance (influence) between them (Baeza-Yates and Ribeiro-Neto (1999)). In other words, an article quoting another is particularly influenced in its research theme, as their degree of similarity is higher. Another five years later, Genisson, Glanzel and Janssens have investigated this concept, demonstrating the applicability through a pilot program (Glenisson et al. (2005)). They developed a lexical metric, using as input the whole set of words of the manuscript, and then developing a subsequent bibliometric representation: the name this technique was bag-of-words. Although this development was a milestone in the future of bibliometric tools, it was not immediately applied, as the major bibliometric and bibliographic providers were reluctant; there was also a lack a strong statistical argument and some fundamental scientific issues:

- the word set of each scientific work included the following morphological categories: articles, prepositions, conjunctions, etc., collectively named stop words – these were lexical units without semantic value;
- use these morphological classes outside discursive syntax cannot be taken into account when creating conceptual similarity metrics;
- the set of words that are semantically irrelevant are leading into error the similarity functions (metrics), used for bibliometric clustering and extraction of features;
- regardless the types of speech, scientific or not, the number of stop words and their number of occurrences prevails, in relation to all words of a manuscript.

This research proposes a different approach, using a semantic analysis application, Tropes, to achieve better, significant results, which can be integrated into an existing mining application. This tool is indexing and staging the word set using general and specialized dictionaries, which extract concepts from a scientific text and filters out *stop words*. In automatic text analysis, Tropes extracts references (class of equivalent terms and concepts), allowing grouping based on referential universes, themes and subthemes (Ghiglione et al. (1998)).

In order to prove the advantage of the previous hypothesis, that considered the use of conceptual references instead of the lexical set, a case study based on a representative article (from bibliometrics) has been developed.

The article data could be found in the top results of one of the most important scientific information databases: Science Direct (see Appendix A. List of analyzed articles). Another ten articles from the references were selected, all being available within the same database. These ten items form the comparison group, needed to measure the semantic distances between the citing and the cited articles.

In figure 1, text analysis results are presented: the number of references (nouns with semantic relevance), compared with the total number of lexical units, per each scientific article.

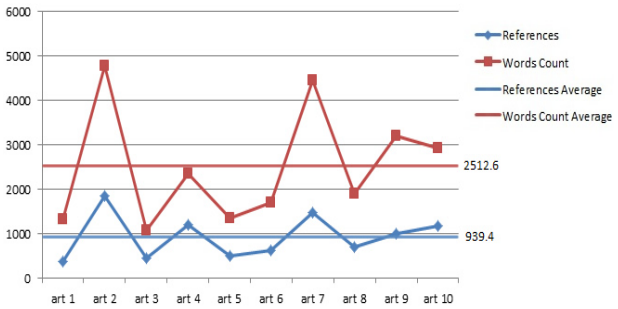


Fig. 1. References vs. Word Count Weight

The above graphic shows how the number of conceptual references keeps, in most cases, a constant proportion of the

total number of words within the article text. According to the results, conceptual references extracted from one article were 30 to 40 percent of the total lexical mass, having an average of 37.3%.

We can conclude in this respect, that only about a third of the words within a scientific article can be used in a semantic similarity measure, the rest of the remaining words, usually having just a role in structuring the discourse. From this point of view, is well known that any research field specific vocabulary is extremely limited, this fact being even more obvious for the research niche areas. Regarding the measurement of semantic similarity between items it is also clear that the input of the process is a far smaller number of items, compared to the already tested lexicometric solutions.

In order to prove this hypothesis, a case study of ten cited articles and one citing work was proposed. As all the articles were research papers in bibliometrics, a very high degree of lexical similarity could be noticed. On the other hand, any article related to bibliometrics will have bibliographic references within the same or complementary research field, leading to increased levels of overlapping references.

The lexical similarity analysis was performed with one of the most popular text mining techniques: Cosine similarity index, developed by Salton in 1983 (Salton and Macgill (1983)). This metric is better than the Jaccard index, especially because of the limited impact of the document size. As demonstrated by Sternitzke and Bergmann in 2009 (Sternitzke and Bergmann (2009)), Jaccard index is strongly influenced by the document length (number of words), with results up to 25% lower than the cosine index, even when comparing lexical subsets of the same text.

The lexical similarity index computation has been made using a Java implementation of the model developed by Salton, which used the word set of each document as input, for each of the tests.

$$similarity(d_1, d_2) = \frac{d_1 * d_2}{\| d_1 \| * \| d_2 \|} = \frac{\sum_{i=1}^{n} Ai * Bi}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

The similarity formula between documents $d_1$ and $d_2$ is computing the ratio between the product of the concept vectors and the product of their modules. $A_i$ is the concept vector of document $d_1$ and $B_i$ is the concept vector of $d_2$.

Analyzing the results in table 1, the similarity factors for the entire lexical sets are very high, with a maximum divergence of 5%. Returning to Baeza-Yates and Ribeiro-Neto theory (Baeza-Yates and Ribeiro-Neto (1999)), one can say that all cited articles have an extremely high degree of influence over the studied paper, between 84% and 90%. In such cases, results processing is extremely difficult, meaning that the assumption of using complete lexical sets for similarity measurements becomes useless.

**Table 1. Lexical Similarity Results based on Cosine Index**

| Lexical similarity between studied document and cited papers | Number of words (in studied doc: 3127) | Similarity |
|---|---|---|
| Cited Document 1 | 1325 | 85% |
| Cited Document 2 | 4784 | 87% |
| Cited Document 3 | 1079 | 84% |
| Cited Document 4 | 2359 | 84% |
| Cited Document 5 | 1362 | 84% |
| Cited Document 6 | 1709 | 89% |
| Cited Document 7 | 4450 | 86% |
| Cited Document 8 | 1912 | 90% |
| Cited Document 9 | 3204 | 85% |
| Cited Document 10 | 2942 | 90% |

In order to test the semantic analysis model, a conceptual references extractor had to be integrated with the similarity measuring algorithm. These items are concepts represented in the document, determining the scientific discourse referentiality. The exclusion of all other elements is justified by the similarity function properties, which uses the input data as a vector of words, ruling out the connections between them. From this perspective, using concepts instead of word sets is a more appropriate approach.

The theory behind the semantic extractor software was defined by Ghiglione (Ghiglione et al. (1998)): conceptual references are classes of equivalent terms, mostly nouns, with *synonymic* or *hyponymic* relationships between them, that can be reduced to one master class based on semantic principles widely distributed and included in general dictionaries of all languages. E.g. various names of plants, with regional or dialect forms, including scientific description are recognized as a single reference.

Although the semantic extractor software proposed references aggregation in a more general universes (e.g. "medicine", “science”, “tools”, etc.) our setup focused on primary references, which insured an acceptable level of details. This way, the case study included even all terms that could not be found in the application dictionary: names of people, specific technical terms, etc.

The semantic similarity analysis results can be found below.

The results are significantly different from those shown in Table 1, where the whole lexical set has been used. On the one hand, as can be seen in Table 2, result presented greater semantic distances between the studied documents (with a maximum of 63%) and, on the other hand, a larger spectrum of similarity grades (22% - 63%).

In light of these results, our proposal of processing and filtering the lexical set through semantic analysis is justified both by the obtained results, as by its ergonomics and possible future applications.

**Table 2. Semantic Similarity Results based on Cosine Index**

| Semantic similarity between studied document and cited papers | No. of semantic references (studied document: 1267) | Similarity |
|---|---|---|
| Cited Document 1 | 367 | 24% |
| Cited Document 2 | 1862 | 31% |
| Cited Document 3 | 450 | 35% |
| Cited Document 4 | 1201 | 43% |
| Cited Document 5 | 505 | 35% |
| Cited Document 6 | 636 | 36% |
| Cited Document 7 | 1486 | 40% |
| Cited Document 8 | 714 | 63% |
| Cited Document 9 | 991 | 22% |
| Cited Document 10 | 1182 | 55% |

## 4. ARTICLE RELEVANCE FACTOR

Up until now, the only way bibliometrics measured the impact of scientific papers was the number of citations. In most cases, the authors of highly cited papers had a major impact in their research field; recently however, we can find more obvious cases when the number of citations doesn't back up a relevant author. Because citation engineering practices spread in publishing, we can now find co-authorship, citation clubs and other techniques to obtain inflated results, based on citations. Therefore, a solution capable to correct the errors of the current system had to be provided, by analysing and developing a new measure of scientific papers impact: article relevance factor.

As presented in section 2.2, eigen factor and article influence score are the indicators that dominate publication assessment space, mainly due to their ability to interpret the entire graph of citations, not just adjacent (directly cited) nodes. Based on this hypothesis, the article relevance factor was developed, based on the formula for calculating the relevance of Web pages, PageRank, developed by Sergey Brin and Larry Page in 1998 (Brin and Page (1998)):

$$PR(p) = \frac{(1-d)}{N} + d\sum_{i=1}^{k}\frac{PR(p_i)}{C(p_i)} \tag{2}$$

The formula uses weighted transfer of web-page relevance, *PR(p)*, where *N* is the total number of web pages, *d* is the damping factor (empirically calculated, for solving the cases of circular references - RankSink, with the value of 0.85 (Brin and Page (1998))) and $C(p_i)$ is the total number of out-links of $p_i$.

Correlating PageRank formula with the article relevance factor, the number of out-links $C(p_i)$ is the equivalent of the total number of citations of article *i*, and *N* (total number of web pages) corresponds to the total number of articles. In the original formula, *(1-d)/N*, corresponds to the probability of a particular page to be absolutely randomly open. However, citations networks have distinct elements from the virtual space, as influential authors have higher chances of being cited than others do; considering this, the first term of this formula must meet proportionality between the citations of a paper and the total of registered citations (Liu et al. (2005)):

$$ARF(a) = (1-d)\frac{CC(a)}{\sum_{j=1}^{N} CC(a_j)} + d\sum_{i=1}^{k}\frac{ARF(a_i)}{CC(a_i)} \tag{3}$$

Considering this, article relevance factor *(ARF*) accounts for the total number of citations received by that paper *CC(a) (citation count (article))* and the value of works that cite it ($\sum$ *ARF*($a_i$)).

However, as it is shown in the first part of the research, not all cited works have same relevance in the context of new research, and the semantic similarity degree is the relevancy factor. Based on the above, *Citation Relevance Weight (CRW)* is defined as the influence ratio of a cited paper, over a new research.

$$CRW(a_i, a_j) = \frac{sem\_sim(a_i, a_j)}{\sum_{z} sem\_sim(a_i, a_z)} * CC(a_i) = SSW(a_i, a_j) * CC(a_i) \tag{4}$$

The numerator *sem_sim* ($a_i$, $a_j$) is the semantic similarity degree between the citing paper $a_i$, and cited paper $a_j$, calculated by the method presented in chapter *3. Using semantic analysis to citation weighting.* The denominator is the sum of all similarity degrees between the works cited $a_z$ and the one citing them $a_i$. $CC(a_i)$ is the citations count from $a_i$, so *CRW* is the weighting factor, smaller or greater than one, of the standard citation. For further calculations, the first term was named *Semantic Similarity Weight (SSW)*, being defined as the ratio of conceptual references imported from a specific paper related to the total conceptual references from the reference list papers.

Because of its advantage over the mathematical alternatives for citation engineering practices censuring (such co-authorship or citation clubs (Yan et al. (2011)), the Citation Relevance Weight is suitable to be applied in the *ARF* expression:

$$ARF(a) = (1-d)\frac{CC(a)}{\sum_{j=1}^{N} CC(a_j)} + d\sum_{i=1}^{k}\frac{ARF(a_i)}{CC(a_i)} * CRW(a_i, a) \tag{5}$$

Therefore the final formula:

$$ARF(a) = (1-d)\frac{CC(a)}{\sum_{j=1}^{N} CC(a_j)} + d\sum_{i=1}^{k} ARF(a_i) * SSW(a_i, a) \tag{6}$$

Computing the relevance factors of an entire article database is a superior effort than the one required for *Journal Impact Factor, AIS* or *Scimago-JR*, focused on journal data collections. Just as calculating *PageRank*, *ARF* has to be applied iteratively, until relatively constant results are obtained. Yet, due to the small number of citations per article (compared to the journal total), a lower number of iterations is required (42 according to PageRank). *Semantic similarity degree – SSW* computation is performed once for each pair

"cited paper - citing paper", so the computational cost is linearly proportional to the number of items. In fact, the computational time increase of applying this metric to a full-text database is justified by the refined method of calculation, switching from journal to article.

The advantage of this solution is the improved article based metric, which can be aggregated into a relevance journal metric:

$$JRF = \frac{\sum_{i=1}^{N} ARF(a_i)}{N} \tag{7}$$

In the *Journal Relevance Factor* formula, *N* represents the number of journal articles, and the numerator is the sum of the published articles relevance degrees.

Calculating this factor over a specified period of time is a sensitive decision: an article in chemistry or biology can receive up to 30 citations in the first two years of life, but one representative paper in mathematics will usually get 5. The ability to interpret the entire network of citations is a great advantage of the method, capable to differentiate between specific research fields citations patterns, being research field independent. Because of these advantages and similarities with other PageRank implementations (Eigenfactor, SCImago Journal Rank), a five year timeframe can be considered for ranking purposes, in order to insure superior accuracy in journals and articles rankings.

Other opportunities may be available by using the *ARF* metric, besides the *Journal Relevance Factor*: building an author relevance factor or institution scientific relevance ranking can be just as simple, as long as we have a proved metric for the base of any opera: the scientific paper.

## 5. BIBLIOMETRIC INDEXES COMPARISON

Bibliometrics still plays an important role in scientific publication assessment, but its importance is obviously decreasing in the context of latest debates and criticisms. Rankings provided by Thomson Reuters in the Journal Citation Reports have enjoyed a position of monopoly for many years, but Scimago Journal Rank (SJR) has won a lot of market share, given its larger computational area: almost all JCR journals and nearly 50% more are included in Scopus. Yet, almost all current assessments are moving towards peer-review, especially because of the large manipulation of citation data and the current metrics incapacity to correct it. Another reason would be the insufficient cover of certain fields like social sciences and recently discovered research niches.

Given the lack of confidence in bibliometric measures, researchers pushed towards a large variety of metrics, causing an *index inflation*. Many derivatives have been developed in an attempt to solve the obvious shortcomings of any known index: *immediacy index, 5 year impact factor, cited half-life, m-quotient, a-index* and *hw-index*. However, all of the above are based on different mathematical models, but applied to the same set of citation data.

In a short and undesirable conclusion, we might believe that journals business related interests outweigh the academic ones, leading to an unqualified set of rankings, more and more repelled by the scientific community. In order to have an overview of each metric, six qualitative criteria were considered:

*a. Capability to employ the whole citation network*

AIS, SCImago and ARF exploit the whole citation network. The PageRank algorithm underlying these metrics uses the structure of the entire citation: the score/rank/factor is recursively defined in terms of the scores of the citing journals and its computation involves the propagation of the journal results over the entire citation graph. Being developed in 1960s, the computation of Garfield's Impact Factor is based on the citations of just a local part of the network, consisting of the journal adjacencies in the citation graph.

*b. Journal size independence*

As many statistics and studies reported, the journal size has a key influence on the impact factor. In October 2000, M. Amin and Mabe published a case-study report on Thomson's IF: small journals (publishing less than 35 papers per year) had a +/-40% fluctuations from one year to another. This effect can be easily reduced by publishing a larger number of articles, being confined to +/-10% margin for large journals, with monthly issues and more than 600 articles per annum (Amin and Mabe (2000)). Also, the Eigenfactor score, which is a measure of the journal's importance to the scientific community, has journal size as component, by definition (eigenfactor.org).

AIS, SJR and the proposed ARF make use of the entire citation network, using data collected during a longer period of time (3 and 5 years), so they are not determined by short article count fluctuations. This is also the case of the 5-year impact factor, which is a much stable indicator than the standard IF (Amin and Mabe (2000)).

*c. Sensitivity to recent journals*

From this perspective, the Impact Factor has an important advantage. After only two years after being indexed in the Thomson's database, the journal receives its first ranking. New journals are somehow disadvantaged by the SCImago Journal Rank, since its citation information is selected on a 3-year time frame. Eigenfactor and AIS are the most demanding, accounting citation behavior between fields by taking a full 5-year measurement window.

*d. Sensitivity to research field differences*

The impact factor exhibits a significant variation according to subject fields. In general, fundamental and pure subject areas have higher average impact factors than specialized or applied ones. The differences are so significant that the top journal in one field may have an IF lower than the bottom journal in another area. The reason for this pronounced subject variation in IFs has to do with the number of researchers in a certain field. Understandably, the more active

authors are active in a field, the more citations will accumulate.

Making use of the entire citations network and calculating the mean influence over a specified research field, AIS and SJR allow for a direct comparison of journals, independent of their subject classification (Ortner (2010)).

*e. Capability to use/calculate citation weight*

A potentially controversial issue related to the calculation of citation based indexes is the use of weighted citation counts. AIS and SJR weight citations with the importance of the citing journals. Citations from highly-ranked journals, like Nature or Science are considered more important than citations from lower-tier journals. Nevertheless, the validity of citation weighting based on the citation profile of the citing journal has been extremely disputed. Some argue that it incorporates a measure of journal prestige, while others argue that it is arbitrary: a citation from an ordinary paper in a prominent journal will be weighted higher than a citation from an excellent paper published in an unknown third-tier journal.

ARS is approaching the issue from another perspective: the citation is weighted based on the value of the citing paper (relevancy), but also with a comprehensive factor of the cited paper contribution to the new research.

By contrast, the Impact Factor simply counts citations without weighting them. As a result, the Impact Factor has been classified as a bibliometric measure of popularity, while the Eigenfactor score captures the bibliometric notion of prestige.

*f. Proof to Citation Engineering Practices*

While the impact factor has been proved to be easily manipulated when using citation engineering practices (self-citations, citation clubs, co-authorship), AIS and SJR have been recognized to be a lot less influenced. The main reason of their performance is the exclusion of journal self cites in their input data.

The downside of excluding journal-level self-citations is the obvious downgrade for the small and niche journals. It is a compromise accepted by their authors, as they already excluded the journals issued for less than five years or publishing less than 12 articles per annum.

From this perspective, ARF presents a great advantage, using all citations available, but counting their relevancy for the present paper. Each citation is weighted by its added value to the new research, narrowing the contribution of empty citations to nearly zero. This means that "friendly" authors' citations are not contributing to the article score, unless it has a real contribution to their research.

Table 3 presents the pros and cons of the three most popular journal indexes, comparing them with the new Journal Relevance Score, over the six selected criterions

**Table 3. Bibliometric index comparison**

| | IF | AIS | SJR | JRF(ARF) |
|---|---|---|---|---|
| Capability to employ the whole citation network | - | + | + | + |
| Journal size independence | - | + | + | + |
| Sensitivity to recent journals | + | - | - | - |
| Sensitivity to research field differences | - | + | + | + |
| Capability to use/calculate citation weight | - | - | - | + |
| Proof to Citation Engineering Practices | - | - | - | + |

The scientific values of an article and implicitly of a journal are determined by two factors: popularity and prestige. While the former is a dimension of citations (IF), it can easily be admitted that it's error-prone, being sensitive to most of the distortion sources. The prestige on the other hand is a recursively weighted computation based on the prestige of the citing journals (AIS, SJR), the general misapplication being the translation of their scoring or ranking, over the articles that form them.

## 6. CONCLUSION

Nowadays, the major bibliometric indexes have proved to be insufficient and inadequate, leading to *index inflation*. A comparative analysis of the major indexes (Table 3), reveals bibliometrics inability to meet the scientists expectations in recognizing the truly valuable works.

In our quest to find an adequate solution, the study proposes an index based on citation value weighting (by a semantic similarity degree - *Equation 6*), which solves the issue without using journal bibliometric information or empirical constants (calculated for specific research areas). Applying the function of intertextual distance only on extracted conceptual references, obtained through semantic processing, has obvious benefits versus previous lexicometric approaches, invalidated in this case study. Thus, the proposed method is scientifically justified, being based on the principles of cognitive-discursive analysis (Ghiglione et al. (1998)), accordingly to any other discourse theory on semantic perspective.

Given the context of bibliometric indexes, the article relevance factor has several major advantages, which allow a wide range of applications and an easy adoption by the bibliometric information providers:

- A refined method to evaluate the real bibliographic impact of each scientific paper;
- The ability to process the information universe within each scientific paper and start off a new development of bibliometrics, with new indexes focused on journals, authors, data collections and publishers;

- The ability to recognize and properly evaluate *empty citations*, removing the side-effects of citation engineering practices.

The Article Relevance Factor is applicable to full-text databases, with a relevant coverage of one or more research fields (Engineering Village, PubMed, SAGE, Science Direct), representing an innovative implementation of current technologies for semantic/ontological processing, in line with the recent developments of information providers.

## REFERENCES


Amin, M., Mabe, M. (2000), Impact Factors: Use and Abuse, *Perspectives in Publishing*, vol. 1, 1-6

Baeza-Yates, R., Ribeiro-Neto, B. (1999), *Modern information retrieval*, ACM Press, New York

Braam, R.R., Moed, H.F., Van Raan, A.F.J. (1991), Mapping of Science by Combined Co-Citation and Word Analysis. II: Dynamical Aspects, *Journal of the American Society for Information Science*, vol. 42, 252-266

Brin, S., Page, L. (1998), *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford InfoLab

Deutsche Forschungsgemeinschaft (2010), *Funding Proposals and Final Reports to Include Fewer Publication Citations,* Berlin, http://www.dfg.de/en/service/press/press_releases/2010/pressemitteilung_nr_07/index.html

Elsevier (2002), *Elsevier and Koninklijke Bibliotheek finalise major archiving agreement*, Glasgow, http://www.elsevier.com/wps/find/authored_newsitem.librarians/companynews05_00020

Elsevier (2010), About SciVerse ScienceDirect, Amsterdam, http://china.elsevier.com/elsevierdnn/SM/tabid/70/Default.aspx

Feller, S.M. (2010), Beyond journal impact factors , *Cell Communication and Signaling*, vol. 8

Friedberg, E.C. (2010), A closer look at bibliometrics, *DNA Repair*, vol. 9, 1018–1020

Ghiglione, R., Landré, A., Bromberg, M., Molette, P. (1998), *L'analyse automatique des contenus*, Paris, Dunod

Glenisson, P., Glanzel, W., Janssens, F., De Moor, B. (2005), Combining full text and bibliometric information in mapping scientific disciplines, *Information Processing and Management*, vol. 41, 1548-1572

Houses of Parliament (2004), *Science and Technology - Tenth Report*, United Kingdom, http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm

Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H. (2005), Co-authorship networks in the digital library research community, *Information Processing and Management*, vol. 41, 1462–1480

Ortner, H.M. (2010), The impact factor and other performance measures – much used with little knowledge about, *Int. Journal of Refractory Metals and Hard Materials*, vol. 28, 559–566

Salton, G., Macgill. M.J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York

Smith, D.R., Rivett, D.A. (2009), Bibliometrics, impact factors and manual therapy: Balancing the science and the art, *Manual Therapy*, vol. 14, 456-459

Sternitzke, C., Bergmann, I. (2009), Similarity measures for document mapping: A comparative study on the level of an individual scientist, *Scientometrics*, vol. 78, 113 –130

West, J.D. (2010), *Eigenfactor: ranking and mapping scientific knowledge*, University of Washington

Woters, P. (1999), Beyond the Holy Grail: from citation theory to indicator theories, *Scientometrics*, vol. 44, 561-580

Yan, E., Ding, Y. (2011), Discovering author impact: A PageRank perspective, *Information Processing and Management*, vol. 47, 125–134