About Automatic Discovering Substantial Citations

Vitalii Velychko
Institute of Cybernetics, NASU,
Kiev, Ukraine
velychko@aduis.com.ua

Levon Aslanyan
Institute for Informatics and Automation Problems, NAS RA, Yerevan, Armenia
lasl@sci.am

Krassimir Markov
Institute of Mathematics and Informatics, BAS
Sofia, Bulgaria
markov@foibg.com

Krassimira Ivanova
University of National and World Economy
Sofia, Bulgaria
krasy78@mail.bg

Abstract

Due to the prevailing role of counting citations over the evaluation of added scientific value, the research activity becomes a kind of business, for instance, inroute of obtaining as more citations as possible. It is important, concurrently to counterbalance the growth of citations value by the introduction and use of effective qualitative criteria for citations. In this paper we continue our research on the approach based on the measure of "usefulness of scientific contribution" called "*USC-index*". An attention is paid to possibilities to discover automatically the substantial citations and in this way, to support the evaluation of the scientific contributions.

Keywords

Usefulness and Evaluation of the Scientific Contributions

1. Introduction

Perhaps the most popular understanding of the concept "scientific contribution" is: (1) the *added scientific value* of the researcher's published results; or (2) its impact on obtaining new scientific results registered by the corresponding *citations*.

As it was mentioned in [17], it is very difficult to measure the added scientific value automatically, without the help of experts.Because of this, in recent years, it became very popular to measure the second part – the citations.There are a number of ways to analyze the impact of publications of a particular research and researcher. A long-time favorite has been ISI's (Social) Science Citation Index, which has come to the web as Web of Science. The web has introduced a number of other tools for assessing the impact of a specific researcher or publication. Some of these are Google Scholar, Scopus, SciFinder Scholar, and MathSciNet among many others. In addition, we mention the system Publish or Perish, which uses data from Google Scholar, but it automatically does analysis on the citation patterns for specific authors. After searching for an author one can select the papers to analyze and to get metrics such as total citations, cites per year, h-index, g-index, etc. [23].

In parallel to the development described a specific negative tendency appears. *The prevailing role of counting citations over the added value evaluation distorts the scientific society* [5]. As a result, the scientific work becomes *a kind of business*, for instance, to obtain as more citations as possible. For examples see [10]. The issue is so broadly discussed and the systemic bottlenecks are so evident that we do not bring more examples.

Instead, our vision is that at least it is important to improve and counterbalance the role of counting citations by using more sophisticated and adequate qualitative criteria [5; 12].

The conceptual function of a citation may be very diverse. In a scientific article some references will represent works that are crucial or significant antecedents to the current work; others may just represent a more general background literature. In [25] five categories of citations were outlined: ***Negative***; ***Perfunctory***; ***Compared***; ***Used***; ***and Substantiated***. Thus, different functions that citations may have in a text are much more complex than merely providing documentation and support for particular statements [2].

Our approach for evaluating the "usefulness of scientific contribution" called "***USC-methodology***" was proposed in [17;18]. Here we will recall it and will discuss the general approach of automatic support to discovery of substantial citations as a part of process of evaluation of scientific contributions (applied) and as an algorithmic problem of a class of pattern recognition (theory).

2. Citation tracking and Evaluation of a Research

Citation tracking is an important task. It allows tracking of the author's own influence, and therefore the influence of his organization. It allows tracking the development of a particular technology, which may be the basis for progress undreamt of when a paper is written. Citation tracking provides information on other organizations and authors who are doing similar work, potentially for collaboration, and identifies publications that cover similar topics. Finally, tracking back in time can find the seminal works in a field [6].

Recent years have seen quantitative bibliometric indicators being increasingly used as a central element in the assessment of the performance of scientists, either individually or as groups, and as an important factor in evaluating and scoring research proposals. These indicators are varied (see [4]), and include e.g.: Citation counts of individual papers published by researchers; Journal metrics (the impact factors of the journals); Measures that quantify personal research contributions over an extended period. This is in conditions when not all researchers are in a similar environment. Research grants are short term and the basic journals involved in assessment systems require more time for publication. Publisher asks for financial support that is not covered by funds. But let us consider the technical part of the issue.

Let us remember that the real metric in business is a measure used to gauge some quantifiable component of an organization's performance, such as return on investment (ROI), or revenues. Metrics are part of the broad area of business intelligence used to help business leaders make more informed decisions. Organizations often use metrics to develop a systematic approach to transform the organization's mission statement and strategy into quantifiable goals, and to monitor the organization's performance in terms of meeting those goals [7]. At the knowledge market [16], the journal metrics are aimed for quantitative evaluation of the popularity and importance of the journals as well as their impact. These metrics have to be measured and used carefully. They are useful for publishers, librarians and administrators, but are not applicable for evaluating of personal scientific contributions.

Quantity measures that quantify personal research contributions over an extended period are based mainly on the idea of [11]. Several papers related to research indices were proposed to assess the quality of the academic research publications. Each one of those indices has its own strengths and weaknesses. The idea of having research indices started when J. Hirsh proposed the H-index [11].

Although the H-index has many limitations and seems biased or unfair in many cases, the other proposed indices such as: G-, H(2)-, HG-, Q^2-, AR-, M-quotient, M-, W-, H_w- ,E-, A-, R-, W-, J-index, etc. considered H-index as a suitable base to produce those other indices with some behavioral enhancements in order to overcome its limitations. In fact, all the other indices use calculations based on the number of citations (originally proposed in H-index) which the authors' papers received (very much similarly to the GOOGLE ranking system). The differences between those indices can be shown through how the index deals with the citation numbers, as in H-index, G-index, W-index, or in adding new attributes such as time, average, etc. as in Contemporary H-index, M-quotient, and AR- index [15]. A review focused in H-Index variants, computation and standardization for different scientific fields is given in [3].

3. Disadvantages of journal metrics and quantitative measures

At the first glance, the variety of scientific measures seems to be very great and great with all its differences.

Really, they all are based on counting the citations and similar formulas based or not on additional criteria like prestige of the journals, time periods, number of authors, etc. The indices for quantifying personal research contributions are based on the same idea of the Hirsh with modifications.

The subject of limitations in research indices is still evolving and with all proposed indices, there are still limitations and weaknesses. Moreover, the *large number of available indices may lead to the dispersion of the evaluation,* and therefore produce differences in values among research communities or even the countries [15].

References that are supposed to be positive they may also be ***negative***. An author may be cited for research of a controversial nature or for an error of the methodology. Here, citation does not always measure the quality of research but rather the impact of a particular piece of work or of an individual scientist [20].

At the end, if an academic shows good citation metrics, it is very likely that he or she has made a significant impact on the field. However, *the reverse is not necessarily true*. If an academic shows weak citation metrics, this may be caused by a lack of impact on the field. However, it may also be caused by: working in a small field; publishing in a language other than English (LOTE); or publishing mainly (in) books[9].

From the beginning, the quantitative measuring of scientific work has been criticized due to problems raised during evaluation of scientific results. Let point one of the earliest papers "Why the impact factor of journals should not be used for evaluating research" [24]. Its arguments are still valid [9].

The most notable and well-documented example of critical view on the H-index (and other "simple" measures of research performance) is the report by the joint Committee on Quantitative Assessment of Research [1]. In this report, the authors argue strongly against the use (or misuse) of citation metrics (e.g., the impact factor or the H-index) alone as a tool for assessing quality of research, and encourage the use of more complex methods for judging scientists, journals or disciplines, that combine both citation metrics as well as other criteria such as memberships on editorial boards, awards, invitations or peer reviews. With regard to the H-index (and associated modifications), specifically, [1] stress that its simplicity is a reason for failing to capture the complicated citation records of researchers, loosing thus crucial information essential for the assessment of a scientist's research. The lack of mathematical/statistical analysis on the properties and behavior of the H-index is also mentioned. This is in contrast to the rather remarkable focus of many articles to demonstrate correlations of H-index with other publication/citation metrics (i.e. published papers or citations received), a result that according to the authors is self-evident, since all these variables are essentially functions of the same basic phenomenon, i.e. publications [21].

Very important disadvantage of these quantitative measures is that they are applicable only to cited papers.

In 1991, David A. Pendlebury of the Philadelphia-based Institute for Scientific Information had published the startling conclusion that ***55%*** *of the papers published in journals covered by ISI's citation database* ***did not receive*** *a single citation in the 5 years after they were published* [8]. In his further publication, Pendlebury gave more concrete data[22]: "The figures -- 47.4% un-quoted for the sciences, 74.7% for the social sciences, and 98.0% for the arts and humanities -- are indeed correct".

In addition, we should measure the *mentoring abilities* of a scientist. Scientists do research and also mentor younger colleagues. Good mentoring should be a significant consideration of one's contribution to science. The H-index might measure research productivity, but currently there does not appear to be a "*mentoring index*" [13]. If the coauthors of a scientist are his or her own trainees or students and if they continue to make a scientific impact after leaving their supervisor, it does point to the quality of the mentoring by the scientist and to the impact made by the scientist, as a result of his/her mentoring abilities, in a given area during a given period. This is a very important but totally neglected aspect of the contribution made by a scientist or an academic. However, *we do not yet have a well–worked out formula to measure such mentoring abilities* [14].

4. Evaluation of Scientific Contributions

The real scientific product is not an object but the ***idea***, communication and the response to the ideas of others. While it is possible to track simultaneously scientists and the money invested, it is far more difficult to measure *science as a body of ideas*, or to grasp its interface with the economic and social system. For now, indicators remain essentially a unit of measure based on observations of science and technology as a system of activities rather than - a body of specific knowledge [19].

The Main Phases of the Science related activities are (1) Creation of a Scientific Result; (2) Registration of the Scientific Result; (3) Implementation and Use of the Scientific Result.

The bibliometric indexes analyze the second phase – registration of scientific result as (primary) publications and as (secondary) citations. The first and third phases are out of bibliometric scope. This way the evaluating of scientific work became partial and not significant.

A possible step to counterbalance is to analyze the publications and citations from the point of view of the third point above – ***implementation and using the scientific results*** [17].

A wide spread understanding is that only high qualified ***academic and industrial researchers*** can evaluate published ideas. They have knowledge and skills to continue research and developing of proposed ideas and via citations they recognize previous research done by other scientists or by themselves. In accordance to the *usefulness of cited ideas*, we may separate academic citations into the three main groups:

- ***Substantial citations***, which applied or substantially use the citing work indicating implementation and the direct use of the citied results, the "mentoring impact" included;
- ***Casual citations***, which includes only a general note or a review of the citing work;
- ***Refuting citations,*** which indicate that the citing work (possibly) has no scientific added value or similar.

The useful scientific result may cause a chain of publications which further use and develop the idea. This way, ***transitive citations*** is another value. Citation chain has to start from a substantial citation and to continue by the same type citations because casual citations could not generate such citation chain.

The influence of the scientific ideas is greatest when citation chain exists. Because of this, the *transitive substantial citations* have to be counted as native characteristic of the scientific publications. It is correct to assume that a transitive substantial citation is equal to the direct one. Similar issue is the cross references that is to be determined and eliminated. It can be outlined another situation when cited, the paper uses the citations of that paper, repeating them automatically giving them a value.

There is also a ***temporal dimension*** of the citation process. An article may first be cited for substantial reasons (e.g., its content has been used). Later when a paper is widely known and has obtained many citations the importance of the other mechanisms will increase (authors citing authoritative papers, the bandwagon effect, etc.). In other words, ***visibility dynamics*** become more important over the time because of the self-intensifying inside mechanisms. This explains why the relative differences in citation rates between poorly cited and highly cited papers increase over the time. Another temporal effect is the phenomenon termed "obliteration by incorporation", meaning that basic theoretical knowledge is not cited anymore. As a consequence, the most basic and important findings may not be among the most highly cited papers because they have been rapidly incorporated into the common body of accepted knowledge [2].

We have to draw attention to one very important fact.

A great number of publications have no chance to be viewed and further studied because they are published in media with ***limited and/or payable access***. In this case only well-known authors have chance to be recognized and possibly – cited.

Only what is needed is publications to be included in different digital libraries with *open access* and *as more such libraries exist* in the world so greatest chance these publications have to become useful. The variety of digital libraries and index data bases with open access to scientific publications and reviews is a crucial factor for further grow of the science.

5. USC-methodology

Following to considerations discussed above, we assume that for evaluating of usefulness of scientific contributions more-less important are:

- ***p*** – Number of the papers;
- ***q*** – Number of monographs;
- ***s*** – Number of the substantial citations;
- ***c*** – Number of the casual citations;
- ***r*** – Number of the refuting citations;
- ***Y*** = y_e - y_b +1 – Length of the interval of publications;
- ***z*** = y_c - y_b – Length of the interval of citations,

where

- y_b – starting year (beginning) of the period of publications;
- y_e – last year (end) of the period of publications;
- y_c – last year (end) of the period of citations.

In this list we have three different types of values, which we have to reduce to a common measurement unit. We propose to use "***paper***" as such unit because it may be assumed that *one paper represents a single idea*.

In accordance with this, we propose to use four coefficients of correlation:

- ***m*** – coefficient of the ***monograph correlation***:
  m : 1 monograph = m papers;
  example: if 1 monograph = 5 papers then m = 5;
- ***a*** – coefficient of the ***substantial citation correlation***:
  a : 1 substantial citation = 1/a paper;
  example: if 5 substantial citations = 1 paper then a=5;
- ***b*** – coefficient of the ***casual citation correlation***:
  b : 1 casual citation = 1/b paper;
  example: if 10 casual citations = 1 paper than b = 10;
- ***v*** – coefficient of the ***refuting citation correlation***:
  v : 1 refuting citation = 1/v paper;
  example: if 10 refuting citation = 1 paper then v = 10.

This way we have the methodological formula for *Usefulness of Scientific Contributions(**USC-index**)*:

$$usc = \frac{p + mq + z}{Y} + \frac{s}{aY} + \frac{c}{bY} - \frac{r}{vY}$$

This formula is *only a formal representation* of the understanding that the ***scientific contributions have to be evaluated completely*** taking into account as more parameters as possible. All types of publications have to be included in the evaluation process as well as mentoring activities, learning materials, and all types of citations including transitive citations, implementations, scientific projects, received funding, etc.

Special comment is needed for ***substantial self-citations***. They are indicators that the scientists provide long-time investigation and step by step publish new results. This is a normal cycle of science. Ignoring this means that we expect receiving the results in one "*genius*" invention. In addition, mentoring students and young researchers leads to publishing of co-authored papers which cause ***substantial citations from co-authors*** in their further independent work and publications. As the received knowledge is more qualitative so more important are the further citations from co-authors. Ignoring this means that we do not acknowledge the high level skills and leading ideas of the advisors.

6. Automated linguistic analysis

The problem of automated linguistic analysis of scientific publications, in accordance with USC-methodology and computing of its *USC-index* for different target scientific structures has to be solved.

It is complicated to compute USC-index for all scientists of a given organization and many times more complicated to do this for all researchers from the given scientific area. Because of this, the computer linguistic analysis of the scientific publications (to obtain values of the main parameters of USC-index) is a serious scientific problem which has to be solved.

Using an ontological approach to the solution of this problem is the most attractive. Suppose that for all the cited publications are constructed corresponding ontologies which identified the concepts and the relationships between them. Ontologies are built on the basis of text of articles by automated way. Comparing the ontologies of the article and cited publications can be exactly defining the category of citations. The problem of ontology automatic building for article is not yet solved, so consider other solutions.

Some preliminary considerations about possibility for solving it may be done.For the automated text analysis of an article can be used some open source text analysis tool, one the leading system of such type is GATE, a General Architecture for Text Engineering [26]. Using an automated text analysis in sections "Abstract", "Conclusions" we can select the combinations of words that indicate the main content of the article ("In this paper we report…", "In this paper we continue research…","Our attention is paid…", "We presented our approach…") and identify terms representing this content. Then allocated terms, and the terms directly related to it, are found in the article text. References connected with the terms of the main content of article are classified as substantial citation. Also in the text of the article can be found a number of word combinations directly indicating the type of citation ("…a sort of…realization of the approach suggested in this paper is the…", "…our approach is based on the …").

For another instance, it is typical that the introduction of a scientific article is structured as a progression from the general to the particular. References have been found to be most frequent in the introductory section of paper. Thus, in the introduction, an article typically refers to more general or basic works within a field. The net effect of many articles referring to the same general works, therefore, is that such contributions get a very large number of citations. References to highly cited publications seemed to occur more in the introduction than anywhere else in the articles. Similarly, since most scientific articles contain a methodology section in which the methods applied in the study are documented, authors typically cite the basic papers describing these methods. This may explain why some papers containing commonly used methods sometimes receive a very large number of citations [2].

At this point we addressed analyzing, modeling and automation issues of the citation framework. Our current input is in the USC methodology developed, and in specific text analysis automation that provides the necessary information about references. Another activity is devoted to the validation framework. At the end we use the pattern recognition, classification algorithms. We suppose a large learning set of citations is prepared manually and in an automated way and that domain experts classified the citations and evaluated the papers, authors and journals. Using the pattern recognition algorithm, say the Voting algorithms by Yu. Zhuravlev [27], we solve the optimization problem of evaluating the model parameters so that the classification fits well to the expert knowledge mentioned above. The total result involving analysis, multi-parametric model, automation for citations, and the validation by the pattern recognition will be given in extended paper after this abstract of the work.

Conclusion

It is not permissible to replace the quality of a scientific publication, with qualities of the media in which it has been published. In science, the incorrect management decisions lead to a decline in its development. *If a complete scientific "industry" is not developed, the "complete" administrative attitude to science grows, which inevitably will kill it.*

Exuberant dependence on single numbers to quantify scientists' contribution and make administrative decisions can affect their career progression or may force people to somehow enhance their H-index instead of focusing on their more legitimate activity, i.e., doing good science. Considering the complex issues [28] associated with the calculation of scientific performance metrics, it is clear that a comprehensive approach should be used to evaluate the research work of a scientist. We should not rely excessively on a single metric [14].

Although the use of such quantitative measures may be considered at first glance to introduce objectivity into assessment, the exclusive use of such indicators to measure science "quality" can cause severe bias in the assessment process when applied simplistically and without appropriate benchmarking to the research environment being considered [4].

A common conclusion of these studies is the recognition of the important role of peer review in the quality assessment of research, and the recommendation to apply bibliometric performance indicators with great caution, and only by peers from the particular discipline being reviewed [4].

We have to underline, that USC-methodology is aimed only to turn process of evaluation of scientific contributions back to human responsibility of authors, reviewers, and publishers. Modern science is distributed all over the world and concentration of any its part in one or two monopolies is absolutely inadmissible. To ensure growing of science we are obligated to provide for growing of variety of possibilities for doing science – financial resources, publishing opportunities, scientific indexing systems, and distributing organizations [17].

Special comment was done for *substantial self-citations*. They are indicators that the scientists provide long time investigation and step by step publish new results. In addition, mentoring students and young researchers leads to publishing of co-authored papers, which causes *substantial citations from co-authors* in their further independent work and publications.

Automated linguistic analysis of usefulness of the scientific contributions is a very important step in evaluation process and developing corresponded tools is actual and substantial. In this paper we have outlined an approach to a possible solution.

USC-index is *only a formal representation* of the understanding that the *scientific contributions have to be evaluated completely* taking into account as more parameters as possible. All types of publications as well as mentoring activities, learning materials, and all types of citations including substantial self-citations, substantial citations from co-authors, transitive citations, implementations, scientific projects, received funding, etc. have to be included in the evaluation of usefulness of scientific contributions. The validation stage that incorporates the experts' knowledge is very hard but important and even with all these – citation analysis at its current level of formalization and automation can only be a recommendation and not a decision making value.

REFERENCES

[1] Adler, R., Ewing, J. and Taylor, P., "Citation Statistics", *Joint IMU/ICIAM/IMSCommittee on Quantitative Assessment of Research*, 2008.

[2] Aksnes D.W. "Citations and their use as indicators in science policy",*Dissertation for the doctoral degree of the University of Twente*, 2005.

[3] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera. "h-Index: A review focused in its variants, computation and standardization for different scientific fields",*Journal of Informetrics 3*, pp. 273–289, 2009.

[4] "On the use of bibliometric indices during assessment" EPS Statement, 2012,

http://c.ymcdn.com/sites/www.eps.org/resource/collection/B77D91E8-2370-43C3-9814-250C65E13549/EPS_statement_June2012.pdf.

[5] San Francisco Declaration on Research Assessment, American Society for Cell Biology (ASCB), 2012,http://am.ascb.org/dora/.

[6] Susan Fingerman "Web of Science and Scopus: Current Features and Capabilities",*Electronic Resources Reviews*, 2006, http://www.istl.org/06-fall/electronic2.html.

[7] State Government of Victoria State Services Authority, A guide to people metrics// The State Services Authority, Melbourne, 2010.

[8] David P. Hamilton," Research Papers: Who's Uncited Now?",*Science*, 251:25, 1991.

[9] Anne-WilHarzing "Google Scholar - a new data source for citation analysis", Research in International Management Products & Services for Academics, 2008.

[10] Harzing A.W. "How to become an author of ESI Highly Cited Papers?",University of Melbourne, 2012.

[11] Hirsch, J. E., "An index to quantify an individual's scientific research output", Proceedings of the National Academy of Sciences of the United States of America, 102(46), pp. 16569–16572, 2005

[12] Join Initiative to put an end to the misuse of the journal impact factor (JIP), Initiative for Science in Europe (ICE).http://www.no-cuts-on-research.eu/news/dora.html.

[13] Kuan-TehJeang, "H-index, mentoring-index, highly-cited and highly accessed: how to evaluate scientists?",*Retro virology*, vol. 5, Number 106, Springer, 2008.

[14] M. Jagadesh Kumar "Evaluating Scientists: Citations, Impact Factor, h-Index, Online Page Hits and What Else?",*IETE TECHNICAL REVIEW*, Vol. 26, ISSUE 3, pp. 165 - 168, 2009.

[15] MajdiMaabreh and Izzat M. Alsmadi "A Survey of Impact and Citation Indices: Limitations and Issues", *International Journal of Advanced Science and Technology*, Vol. 40, pp. 35-54, 2012.

[16] K. Markov, K. Ivanova, I. Mitov, N. Ivanova, A. Danilov, K. Boikatchev "Basic Structure of the Knowledge Market", *IJ ITA*, V.9, No.4, pp.123-134, 2002.

[17] Kr. Markov, Kr. Ivanova, V. Velychko "Usefulness of Scientific Contributions", *International Journal "Information Theories and Applications"*, Vol. 20, Number 1, pp. 4-38, 2013

[18] Kr. Markov, Kr. Ivanova, V. Velychko. "Evaluation of Scientific Contributions",*International Scientific Conference "Modern Informatics: Problems, Achievements, and Prospects of Development"*, devoted to the 90th anniversary of academician V. M. Glushkov. Kiev, Ukraine, 2013 (in print).

[19] National Science Foundation, "Science & Engineering Indicators" 1989, National Science Foundation, Washington, DC

[20] Okubo, Y., "Bibliometric Indicators and Analysis of Research Systems: Methods and Examples", *OECD Science*, Technology and Industry Working Papers, OECD Publishing, 1997.

[21] John Panaretos and Chrisovaladis Malesios, "Assessing scientist research performance and impact with single indices", Scientometrics, Vol. 81/3, pp 635-670, 2009

[22] David A. Pendlebury, "Science, Citation, and Funding", Science, 251:1410-1411, 1991

[23] Peper M. Publish or Perish. May 22, 2009.

[24] Per O Seglen "Why the impact factor of journals should not be used for evaluating research",*BMJ*, 314:498–502, 1997.

[25] Small H.,"Citation Context Analysis", In B. Dervin & M. Voigt (Eds.), Progress in communication sciencesVol. 3, pp. 287-310, Norwood: Ablex. 1982.

[26] Cunningham H, Maynard D, Bontcheva K, Tablan V, Aswani N, et al. (2011) Text Processing with GATE (Version 6). The University of Sheffield,

http://tinyurl.com/gatebook.

[27] Zhuravlev Yu., Selected papers, Magister, Moscow, 1998, 418p.

[28] Complex Systems Society, http://http://cssociety.org