EasyChair Preprint

№ 2145

Calculation of Scientometrical Indexes on the Base of Data from several Quotation Systems under the Conditions of Noncomplete Determination

Irina Bolodurina, Petr Boldyrev and Ivan Krylov

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 12, 2019

Calculation of Scientometrical Indexes on the Base of Data from several Quotation Systems under the Conditions of Noncomplete Determination

Bolodurina I.$^{1[0000-0003-0096-2587]}$, Boldyrev P.$^{1[0000-0001-7346-6993]}$, Krylov I.$^{1[0000-0002-8377-1489]}$

[1] Orenburg State University, Orenburg av. Pobedy 13, Russia
library_oit@mail.osu.com

Abstract. This work presented the problem of calculation of the main scientometrical indexes on the base of data from different quotation systems under the conditions of noncomplete determination caused by limited access to the quotation systems. It was offered the approach including following components: data collection and processing; the construction of aggregative publications list; the calculation of total number of quotations for each publication; the calculation of citation index and Hirsch index. The offered approach allowed make the calculation of citation and Hirsch index on the base of data from RSCI and SCOPUS quotation systems under the conditions of noncoplete determination. It was offered the calculation algorithm of total number of quotations the base of which makes mathematical apparatus of fuzzy decision trees. Fuzzy decision tree was learnt on learning data set. There were carried out the investigational studies of developed approach that allowed make the conclusion about the possibility of their usage for calculation of the citation and Hirsch indexes under the conditions of noncomplete determination.

Keywords: scientometrical indexes, citation index, Hirsch index, conditions of noncomplete determination, the construction of aggregative publications list, fuzzy decision trees, the calculation of total number of quotations.

1 Introduction

In recent years, an increasing publication activity of the scientific community has been observed in developing countries (China, Brazil, Turkey, Iran, and others) [1]. There is a low share of publications by Russian authors in the global share of publications (less than 2% of the total number of publications) and low positions of Russian universities in various international rankings. Therefore, in the Russian Federation at the state level, attempts are being made to increase the share of publications by Russian authors in the global fund of scientific publications.

At the moment to estimate the effectiveness of scientific work the scientometrical indexes are used together with experts' opinions. It's due to the presence of secure available for measurement and comparison information about the scientific researches results presented in different quotation systems [1, 2]. To the most popular foreign quotation systems can be referred Web of Science and Scopus[3] and to the Russian – RSCI [4].

However despite the big choice of quotation systems offering data for estimation of authors' publication activity there is number of problems preventing their wide usage in scientific and educational organizations [5, 6]. Quotation systems Web of Science and Scopus don't include the majority of publications in Russian. Quotation system RSCI doesn't have an access to the big number of foreign publications and also doesn't have the majority of works up to 2000. Special importance this problem gets under the conditions of limited access to scientometrical information because of high price.

2 Research Overview

Some scientists tried to develop software tools providing additional possibilities working with quotation systems [7, 8, 9]. However after the analyses of developed software tools following conclusions were made:

- there aren't software tools that can aggregate data and calculate scientometrical indexes taking into account Russian Science Citation Index;
- developed software tools are supposed to work with the availability of full saccess to the quotation systems and can't work under the conditions of noncomplete determination [6].

In that case under the conditions of noncomplete determination is meant the absence of possibility to identify uniquely and relate the quotations (the title of quoted work, the source and publication year) in foreign quotation systems with the list of quotations in RSCI for each publication. Whereas there is provided only the quantity of quotation for each chosen publication in foreign quotation system. As a consequence there is the problem of calculation of total quotations number for each publication [10, 11]. It's also impossible to calculate the main scientometrical indexes (quotation index, Hirsch index [12] and others).

Consequently the objective of this work is the realization of calculation of the main csientometrical indexes (quotation index and Hirsch index) on the base of data from RSCI and SCOPUS under the conditions of noncomplete determination.

3 Calculation of Scientometrical Indexes under the Conditions of Noncomplete Determination

To solve the presented problem there was developed the approach allowing make the analyses of the main scientometrical indexes such as quotation index, Hirsch index and others on the base of data from quotation systems under the conditions of noncomplete determination. The main components of developed approach and their relations are shown through IDEF0-diagram (see Fig. 1).

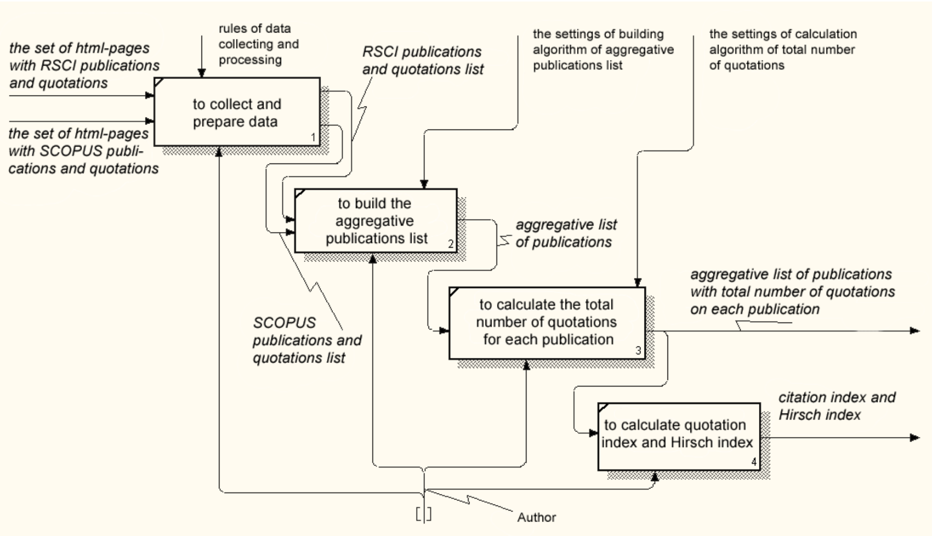

rules of data collecting and processing
the set of html-pages with RSCI publications and quotations
the set of html-pages with SCOPUS publi- cations and quotations
to collect and prepare data
1
RSCI publications and quotations list
the settings of building algorithm of aggregative publications list
the settings of calculation algorithm of total number of quotations
to build the aggregative publications list
2
aggregative list of publications
SCOPUS publications and quotations list
to calculate the total number of quotations for each publication
3
aggregative list of publications with total number of quotations on each publication
to calculate quotation index and Hirsch index
4
citation index and Hirsch index
Author

Fig.1. IDEF0-diagram

Let's consider the analyses of authors' publication activity on the base of RSCI and SCOPUS quotation systems.

At the stage of data collection and preparing is performed the review of html-pages with the list of publications and quotations from RSCI and SCOPUS quotation systems.

By building of aggregative list of publications it's being formed the list of author's publications on the base of data from RSCI and SCOPUS quotation systems in which there are no duplicated publications. The base of building algorithm of aggregative list of publications makes shingles algorithm [13].

The stage of data collection and preparing and also the stage of aggregative list of publications building are presented more detailed in work [14].

To calculate the total number of quotations for each publication the data used received at the stage of building of aggregative list of publications precisely the number of quotations in RSCI, number of quotations in SCOPUS and also the number of found duplicated publications in SCOPUS system.

The base of calculation algorithm of total number of quotations under the conditions of noncomplete determination makes mathematical apparatus of fuzzy decision trees [15]. This mathematical apparatus combine decision trees and fuzzy logic advantages: allows operate quality characteristics of the subject; used in situations when it's difficult to classify the subject exactly according to any attribute позволяет; provides training on comparable small data set.

By building fuzzy decision tree for each attribute were selected some linguistic variables and defined examples membership degree. Instead of number of examples for each knot fuzzy decision tree groups their membership degrees.

There were selected 2 target classes: «small portion of intersectional quotations» (negative result), «big portion of intersectional quotations» (positive result).

As attributes by which decision tree was built were selected the following: «the number of duplicated publications in SCOPUS system», «maximum possible number of duplicates between quotation systems». The attribute «the number of duplicated publications in SCOPUS system» was given by linguistic variable with the following term-set of meanings: «lack of duplicates», «ordinary number», «big number». The attribute «maximum possible number of duplicates between quotation systems» was given by linguistic variable with the following term-set of meanings: «small», «medium», «big» (see Fig. 2 and Fig. 3).

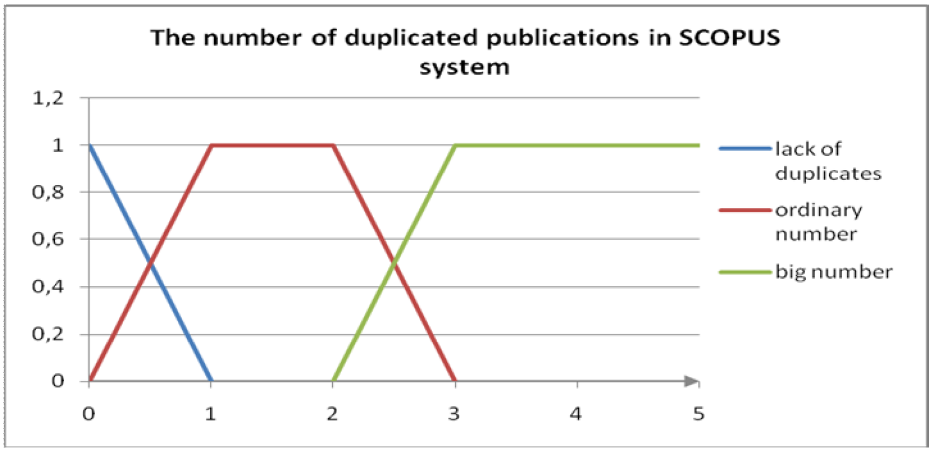


Fig.2. Membership functions for term-sets on attribute «the number of duplicated publications in SCOPUS system»

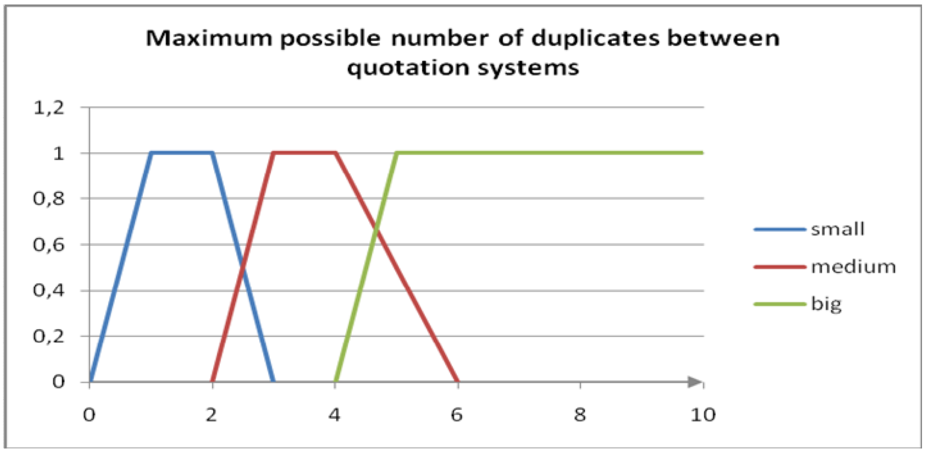


Fig. 3. Membership functions for term-sets on attribute «maximum possible number of duplicates between quotation systems»

Numeric value of attribute «maximum possible number of duplicates between quotation systems» is calculated in the following way:

$$x_i^1 = \min(K_i^{SCOPUS}; K_i^{RISC}), \tag{1}$$

where K_i^{SCOPUS}- the number of quotations in SCOPUS system for the current publication, K_i^{RISC}- the number of quotations in RSCI system for the current publication.

For construction of fuzzy decision tree is used the algorithm consisting of several stages.

At the first stage of algorithm work general entropy is calculated.

In the following stage are calculated coefficients P for each possible node. The calculation of coefficients P for each node N is accomplished in the following way:

$$P_i^N = \sum\nolimits_{S^N} \min(\mu_N(D_j), \mu_i(D_j)), \tag{2}$$

where $\mu_N(D_j)$ – membership level of the training example D_j to the node N, $\mu_i(D_j)$ – membership level of training example toward objective value *i*, S^N - the variety of all examples.

Coefficient defining main characteristics of the node *N* is calculated in the following way:

$$P^N = \sum\nolimits_i P_i^N, \tag{3}$$

In the following stage is calculated the entropy that estimates the average number of information to determine the object class from the set P^N:

$$E(S^N) = -\sum\nolimits_i \frac{P_i^N}{P^N} \cdot \log_2 \frac{P_i^N}{P^N}. \tag{4}$$

Then the entropy for each attribute individually is calculated:

$$E(S^N, A) = \sum\nolimits_j \frac{P^{N|j}}{P^N} \cdot E(S^{N|j}), \tag{5}$$

where *N*|*j* – child of node *N*.

Then the information gain on each attribute is calculated:

$$G(S^N, A) = E(S^N) - E(S^N, A). \tag{6}$$

Finally as root attribute is chosen the attribute with the maximum information gain.

Then node *N* is devided into subnodes *N*|*j*. Membership level of each example D^k for node *N*|*j* is calculated from node *N*:

$$\mu_{N|j}(e_k) = \min(\mu_{N|j}(D^k); \mu_{N|j}(D^k, a_j)), \tag{7}$$

where $\mu_{N|j}(D^k, a_j)$ demonstrates the membership level D^k to the attribute a_j. In case if none of the examples belongs to the node *N|j,* this node is deleted.

The algorithm work continues until all the attributes are used or all the examples aren't classified.

The membership to the target class for the new recording is found in the following way:

$$\delta_j = \frac{\sum_l \sum_k P_k^l \cdot \mu_l(D_j) \cdot x_k}{\sum_l (\mu_l(D_j) \cdot \sum_k P_k^l}, \tag{8}$$

where P_k^l - coefficient displaying the correlation of examples *l* for target class values *k*, $\mu_l(D_j)$ - membership level of example *j* to the node *l*, x_k - membership of the target class value *k* to the positive result.

The calculation of total number of quotations is made in the following way:

$$S_i = K_i^{SCOPUS} + K_i^{RISC} - \min(K_i^{SCOPUS}; K_i^{RISC}) \cdot \delta_i \cdot \tag{9}$$

The generalized chart of calculating algorithm of total number of quotations by means of built decision tree was developed (see Fig. 4).

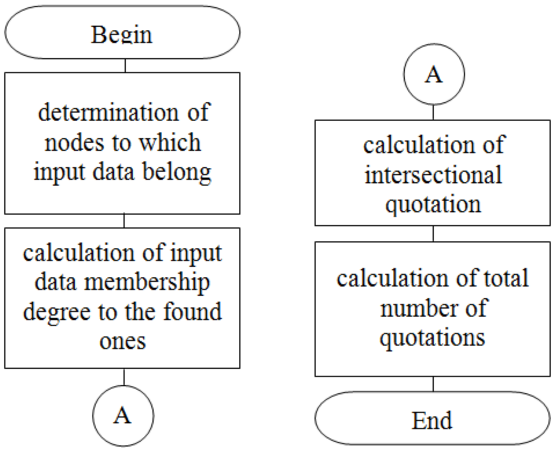

Begin
determination of nodes to which input data belong
calculation of input data membership degree to the found ones
A
A
calculation of intersectional quotation
calculation of total number of quotations
End

Fig. 4. Generalized chart of calculating algorithm of total number of quotations

For analyses of algorithm work results on calculation of total number of quotations for each publication at first stage was carried out the building of fuzzy decision tree on the base of training set.

Training set was formed on the base of publications list of 5 leading authors of Orenburg State University having fairly large number of duplicated publications in RSCI and SCOPUS quotation systems.

A fuzzy decision tree has been built (see Fig. 5).

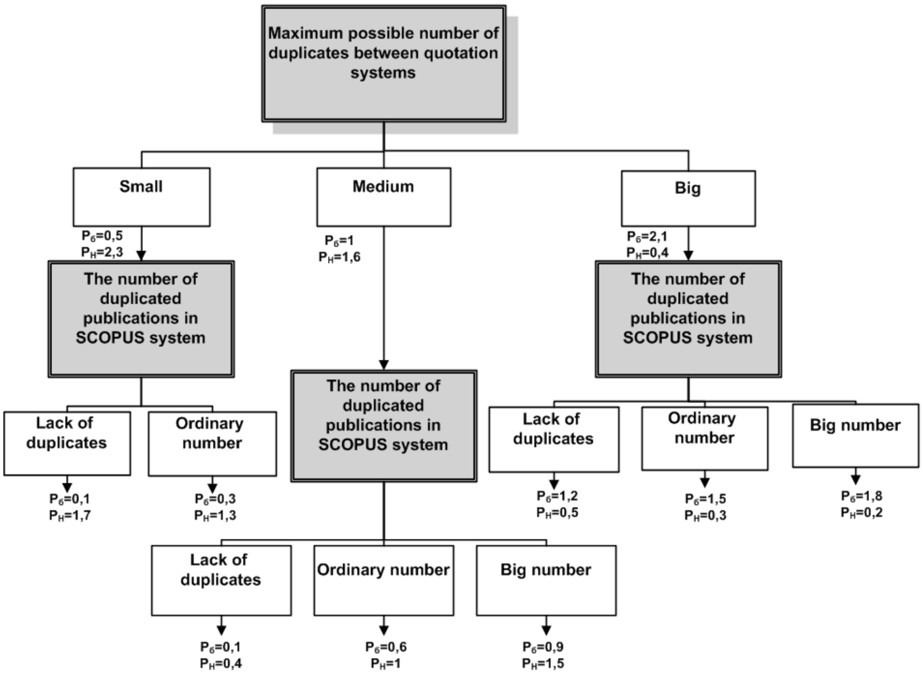


Fig. 5. Constructed fuzzy decision tree

At the next stage trained tree was tested. The test set was formed on the base of list of publications of 3 leading authors of Orenburg State University that aren't included in training set and having fairly large number of duplicated publications in RSCI and SCOPUS systems.

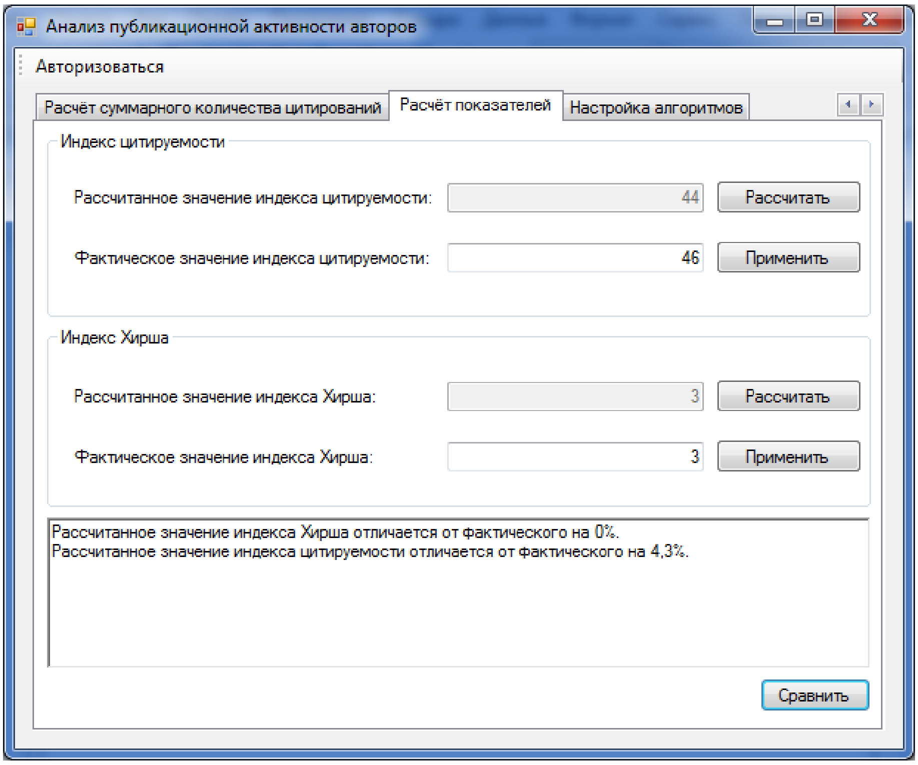
As result of algorithm work in training set were received the following results: for more than 82% publications were noticed the coincidence of total number of quotations calculated by means of developed and real algorithm; for 18% of publications calculated total number of quotations by means of developed algorithm differed slightly from real value.

4 Results of experiments

The results of citation index and Hirsch index calculation for one of the authors included in test set are presented in table 1 and on figure (see Fig. 6).

Table 1. The results of citation index and Hirsch index calculation

The name of index	Index value in RSCI	Index value in SCOPUS	Calculated index value	Index real value
Citation index	35	13	44	46
Hirsch index	2	2	3	3



Рассчитанное значение индекса цитируемости: 44

Фактическое значение индекса цитируемости: 46

Рассчитанное значение индекса Хирша: 3

Фактическое значение индекса Хирша: 3

Рассчитанное значение индекса Хирша отличается от фактического на 0%.
Рассчитанное значение индекса цитируемости отличается от фактического на 4,3%.

Fig. 6. The program at the stage of the main scientometrical indexes calculation

Conclusion

The results of citation index and Hirsch index on the base of developed approach allowed make following conclusions: Hirsch index calculated value matches with real value; calculated value of citation index differs slightly from the real one.

As it can be seen from the above it's possible to talk about acceptable results of developed approach and the possibility of its further usage for citation index and Hirsch index calculation.

References

1. Kotsemir M.: Publication Activity of Russian Researches in Leading International Scientific Journals. Acta naturae. Vol. 4 N. 2(13), 15-35 (2012).
2. Buela-Casal, G.: Assessing the quality of articles and scientific journals: Proposal for weighted impact factor. Psychology in Spain. Vol. 8, N 1, 60–76 (2004).
3. Scopus, https://www.scopus.com, last accessed 2018/10/10.
4. RSCI: Russian Science Citation Index, https://elibrary.ru/project_risc.asp, last accessed 2018/10/10.
5. Van Noorden, R.: A profusion of measures. Nature 465, 864–866 (2010).
6. Podlubny, I.: Comparison of scientific impact expressed by the number of citations in different fields of science. Scientometrics. Vol. 64, N.1, 95–99 (2005).
7. Baneyx, A.: “Publish or Perish” as citation metrics used to analyze scientific output in the humanities: international case studies in economics, geography, social sciences, philosophy, and history. Archivum Immunologiae Et Therapiae Experimentalis. Vol. 56 N. 6, 363-371 (2008).
8. Yang, K., Meho, L.: Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. Proceedings of the American Society for Information Science and Technology. V. 43, I. 1, 1-15 (2006).
9. Garfield, E., Paris. S, Stock, W.: HistCite: a software tool for informetric analysis of citation linkage. Infometrics 57, 391-400 (2006).
10. Schreiber, M.: Modification of the h-index: The h(m)-index accounts for multi-authored manuscripts. Journal of Informetrics. Vol. 2, N 3, 211–216 (2008).
11. Garcia-Perez, M.: The Hirsch h index in a nonmainstream area: methodology of the behavioral sciences in Spain. The Spanish Journal of Psychology. Vol. 12, N. 2, 833–849 (2009).
12. Hirsch, J.: An index to quantify an individual’s scientific research output. Proceedings of the National Academy of Science 102, 16569-16572 (2005).
13. Boldyrev, P., Krylov I.: Providing quality search in the electronic catalog of scientific library via Yandex.Server. EMIT 2018 Internationalization of Education in Applied Mathematics and Informatics for HighTech Applications : proceedings of the workshop, pp. 42-46, Leipzig, Germany (2018).
14. Boldyrev, P., Krylov I.: Several characteristics of existing automated systems according to survey of russian scientists publishing activity. Theoretical & Applied Science 5(25), 6–9 (2015).
15. Janikow, C.: Fuzzy Decision Trees: Issues and Methods. IEEE transactions on systems, man, and cybernetics – part b: Cybernetics. Vol. 28, N. 1, 1-14 (1998).