УДК 004.912

Wikipedia Index of scientists' popularity

Lande D.V., Dr. of Sciences, Andrushchenko V.B., Balagura I.V., PhD,
Institute for Information Recording of NAS of Ukraine, Kyiv,
dwlande@gmail.com

The new index of the scientists' popularity estimation is represented in the paper. The index is calculated on the basis of Wikipedia encyclopedia analysis (Wikipedia Index– *WI*). Unlike the conventional existed citation indices, the suggested mark allows to evaluate not only the popularity of the author, as it can be done by means of calculating the general citation number or by the Hirsch index, which is often used to measure the author's research rate. The index gives an opportunity to estimate the author's popularity, his/her influence within the sought-after area "knowledge area" in the Internet – in the Wikipedia.

Introduction

Today scientometric mostly uses several indices, according to which the scientists' rate and their impact on science and society are calculated. Thus, the simplest index is the number author's publications. It is clear that this index does not depict the qualitative parameters that are better reflected in another index - the number of citations. This index doesn't illustrate the overall performance of the author because the author of just the one, but very important work may exceed this indicator in comparison with scientists who regularly publish their results. In 2005 the physician Jorge E. Hirsch from the California University established the most popular index – Hirsch Index.

The principle of its calculation is quite simple, while it combines the advantages of the first and second approaches. The index calculation is based on the distribution of citations of the work of researcher. According to Hirsch scientist has index *h*, if h of his *Np* papers cited at least h times each, while both articles remaining (*Np – h*) quoted no more than h times each. This index gained the support and is used in such scientometric systems as Scopus, Web of Science, and Google Scholar Citations.

At the same time this indicator, which is focused on the scientific importance, significance of the author, not quite fully reflects the overall importance of the results that he/she received. For such an assessment it is appropriate to use non-fiction and open access systems. As one of the approaches to solve this problem, the authors proposed methodology for calculating the new index - the Wikipedia Index of authors' popularity [1].

This index can appear unimportant tool in combination and with other indices can provide a complete picture of influential scientific achievements of the author, not only in the research community, but the overall impact on the formation of perspective and fully understanding of research information by the users.

The network service Wikipedia - the largest and most democratic Internet encyclopedia is considered, access to which does not presume subscription and furthermore the system is available for download in full. Today Wikipedia (www.wikipedia.org) is the most visited site in the Internet. At this time only the English version of Wikipedia contains more than 5 million articles.

A sufficient amount of works and publications are dedicated to the research of subject areas as well as to the Wikipedia service that prove the relevance of the conducted studies [2]. The methods of building networks of co-authors, the definition of significant nodes of the network structure, research citations and appropriate buildings [3] are among them.

Based on the results of the processed data, we can assume the uniqueness of the proposed indices and value of the information that will be obtained by the computations to evaluate the level of certain data in the system of science popularization and accessibility of provided research information on specific issues. The use of indices is appropriate in different directions of evaluation and analysis of scientific activity, can also act as an additional tool for decision making, forming educational programs etc.

The rule of Wikipedia Index computation

The authors suggested the following rules for calculating Wikipedia Index of author's popularity. It is supposed that the references on the author are found in *N* Wikipedia articles.

Sorted by decreasing number of parameters that determine how many times author's name happens in bibliographic references of these articles we will denote as: R_1, R_2, ..., R_N.

Wikipedia Index of author's popularity (*WI*) corresponds to the maximum number of articles (*WH*) of Wikipedia, in which the number of references no more than the *WH* value, which is multiplied by a certain integral function, which is not decreasing (e.g., the square root is considered below) the N, that is:

$$WI = WH \times \sqrt{N} = \max\left(i : R_i > i\right) \times \sqrt{N}$$

Wikipedia Index of author popularity is ideologically close to the Hirsch index; however, it doesn't take into account the number of articles that refer to the author's article and citations to the work of the author and the number of articles from Wikipedia, which contain these data links. Another difference from the

Hirsch index is the multiplication by a function of *N*, reflecting the consideration it provides greater popularity and the more spread of index values for different authors.

It should be noted that the level of popularity of the author must be attached to his subject domain on one hand in order to avoid false counting for homonyms, and on the other - to ensure completeness on subject area.

Algorithm

In the process of the Wikipedia Index calculating there should be provided the procedure of Wikipedia resources scanning, corresponding to the subject area in which the author works. Accordingly, as "adverse product" of the Wikipedia-index computations, a model of the subject domain is being built, the model – is the network – nodes are concepts that represent articles from Wikipedia, and edges – are the hyperlinks between articles.

To implement calculation of Wikipedia Index authors considered the following algorithm to form subject domains according to Wikipedia, avoiding the effect of topic drift):

1. On the main national Wikipedia page in the search line the initial word is given, e.g. (for English version - **Albert Einstein**).
2. The search window opens. It contains information about concept, according to the task on the Step1. The initial word/word combination is a graph vertex, which will be formed as the result of scanning.
3. All terms-concepts corresponding the hyperlinks on the chosen page, are added to the formed graph. All the words/words combinations are the nodes of the graph. The edges to them are formed from the initial node.
4. The next transition is made by the first not involved hyperlink from the examining pages.
5. In text on the page to which the transition has been made the search of shortened researcher's name (e.g., Einstein or tags (e.g., **physics, relativity**) is to be carried out.
6. In case, if there is a shortened researcher's name or tags is found, the transition to the Step 4 is made and accordingly from the node – word/word combination of the current search the new nodes are built.
7. If there is no word/word combination in the text – the given graph branch is considered to be built.
8. The next transition presumes pass to the page, which had been scanned – the word is not added as a graph node, and the feedback to the created node is formed.

9. All the operations under steps 4-9 repeat until the not involved hyperlinks, chosen from the page, are left. In another case the graph is considered to be built.

According to the suggested algorithm the data collection process in Wikipedia from the first node-notion is stopped when according to the algorithm transition to the new node is impossible (there are no more basic nodes for transition), so the "loop" is impossible.

To compute the Wikipedia Index it is necessary to make some changes to the suggested above algorithms, that is on the page, transition to which had been made by the hyperlink (5th Step of the algorithm), the search of author mentions in Publications, References, Further Reading sections is provided.

Herewith, the number of these mentions, which correlates values R_i, is counted. If $R_i = 0$, the article is not important, the concept is defined as the endnote and the transition to the Step 4 is provided. Of course, this rule narrows the scanning of Wikipedia pages list and results the completeness loss, though, as the real computations prove, has little effect on the overall results. Pages dedicated to the scientific concepts and those, which don't contain relevant publications, can be ignored – just skipped. Therewith, the time of Wikipedia target segment is significantly reduced.

As a result of the full network sounding, the sequence R_1, R_2, ..., R_N is formed, which is used to calculate Wikipedia Index, according to the rules above.

Experimental section

The represented algorithms were implemented as a software system, through which the subject domains models and Wikipedia-index are formed. Here are some examples of calculating Wikipedia Indices for three authors: Albert Einstein, Enrico Fermi, Benoit Mandelbrot.

In Fig. 1 shows the Gephi visualization of domain model fragments that were obtained by sounding Wikipedia according to the above algorithm. The parameters of obtained networks (subject domain models); nodes-concepts of Wikipedia are following.

For a network that meets the model of authors' subject domain:

Albert Einstein: nodes – 718, edges – 22111, $WI = 12 \times 11,5 = 138$ (128 articles with the references, WH=12), the largest nodes:

Consept	The node degree
Quantum_nonlocality	188

Alain_Aspect	181
Hermann_Weyl	177
Paul_Dirac	174
Electromagnetic_radiation	174
Isaac_Newton	169
Galileo_Galilei	169
Wolfgang_Pauli	169
General_relativity	167
Antimatter	167

Enrico Fermi: nodes – 605, edges – 22079, $WI = 7 \times 9,6 = 67$ (92 articles with the references, WH=7), the largest nodes:

Consept	The node degree
Enrico_Fermi	440
Nobelium	206
Transuranic_element	206
Particle_physics	204
Mendelevium	204
Einsteinium	204
Berkelium	203
Radioactive_decay	195
Radioactive	190
Particle_accelerators	188

Benoit Mandelbrot: nodes – 34, edges – 259, $WI = 6 \times \sqrt{11} \approx 20$ (11 articles with the references, WH = 6), the largest nodes:

Consept	The node degree
Benoit_Mandelbrot	22
Pattern	20
Chaos_theory	18
Patterns_in_nature	18
Hausdorff_dimension	17
Patterns	16
Fractal	15
Fractal_dimension	15
Fractal_geometry	15
Fractals	15

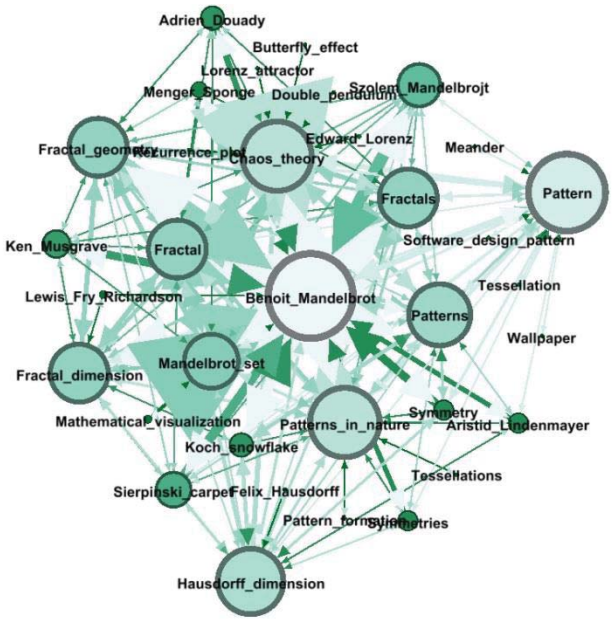

Fig.2. Fragments of subject domain "Benoit Mandelbrot"

There were provided comparisons of the results – Wikipedia Index, calculated on the research and the Hirsch-index, represented by the world's leading scientometric resources Scopus, Web Of Science and Google Scholar Citations (Table 1).

Table 1. **Comparison of Wikipedia Indices values with the Hirsch-index (Scopus, Web Of Science and Google Scholar Citations)**

N	Scientist	Wikipedia Index	h-index Scopus	h-index Web Of Science	h-index Google Scholar Citations
1.	Albert Einstein	141	36	6	110
2.	Enrico Fermi	67	26	1	49*
3.	Benoit Mandelbrot	20	31	36	90

*Profile missing, the value was calculated for: "E. Fermi" according to the Google Scholar (Google Scholar Calculator) service.

By comparison, we can see and estimate the role of information on research and publications on open-access resources in comparison with data that consider purely scientific information with a certain restrictions set.

Conclusions

The principle of Wikipedia Index forming differs primarily from those, which currently is used in scientometrics with consideration of citation from not only scientific papers but popular service Wikipedia (separately for each language version). This way the index of author's popularity within this service can be obtained. This is an important issue, considering the fact that Wikipedia is currently the largest and most popular encyclopedic resource.

There is suggested the technique of the Wikipedia Index quick calculation, which allows to realize computation as a separate service, and also automatically form the subject domain.

Provided work may be continued by analyzing other resources and the formation of indicators to estimate and analyze the influence in a particular environment. It is also necessary to note a fundamental difference between the proposed approach of automatic subject domains models formation [4] and those that already exist, based on direct participation of experts in selecting specific nodes and links. In cases, as it depicted in the work, the researcher uses only a small share of knowledge represented by the name of the scientist, his writing abbreviated names of several key terms, concepts to construct an appropriate network.

Bibliography

1. Lande D.V., Andrushchenko V.B., Balagura I.V. Wiki-index of authors' popularity // E-preprint ArXiv arxiv.org/abs/1702.04614
2. Zareen Saba Syed, Tim Finin, Anupam Joshi. Wikipedia as an Ontology for Describing Documents, Proc. 2nd Int. Conf. on Weblogs and Social Media, AAAI Press, March 2008., pp. 136-144.
3. Lande, D. V., Snarskii, A. A., Bezsudnov, I. V. (2009), Internetika: Navigation in complex networks: models and algorithms [Internetika : navigatciia v slozhnykh setiakh : modeli i algoritmy], Librokom (Editorial URSS), Moscow, 264 p. (In Russian).
4. Lande D.V., Andrushchenko V.B., Balagura I.V. Formation of the Subject Area on the Base of Wikipedia Service // Open Semantic Technologies for Intelligent Systems: Proceed. of Intern. sci-tech conf., ISSN 2415-7740; Issue 1 (Minsk, 16-18 feb 2017). – Minsk : BSUIR, 2017. – pp. 211-214.

Міністерство освіти і науки України
Національний технічний університет України "Київський
політехнічний інститут імені Ігоря Сікорського"
Факультет прикладної математики
ННК "Інститут прикладного системного аналізу"
Російська асоціація штучного інтелекту
Білоруський державний університет
інформатики і радіоелектроніки
Видавництво «Просвіта»

XVII Міжнародна наукова конференція імені Т. А. Таран

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ІНФОРМАЦІЇ

IAI-2017

Київ, 17 - 19 травня 2017 р.

Збірка праць

Рекомендовано Вченою радою
факультету прикладної математики

Київ
«Просвіта»
2017

УДК 004.8/.9+001.102](06)
ББК 32.973я43+73я43
І-73

Головний редактор к.т.н., доц. Сирота С.В.
Відповідальний редактор Темнікова О.Л.
Відповідальний за випуск Копичко С.М.
Рекомендовано Вченою радою факультету прикладної математики
КПІ ім. Ігоря Сікорського протокол №10 від 29 травня 2017 р.

І-73 XVII Міжнародна наукова конференція імені Т.А. Таран «Інтелектуальний аналіз інформації» ІАІ- 2017, Київ, 17–19 травня 2017 р. : зб. пр. – К. : Просвіта, 2017. – 252 с. : іл.
ISBN 978–617–7010–13–4

У збірці опубліковані доповіді, представлені на конференції з наступних напрямів: мережеві і багатоагентні моделі, знання і міркування, онтологічний інжиніринг, аналіз даних, м'які обчислення, обробка природної мови, соціальні проблеми і освіта.

УДК 004.8/.9+001.102](06)
ББК 32.973я43+73я43

ISBN 978–617–7010–13–4