An Index of Authors' Popularity for Internet Encyclopedia

Dmitry Lande, Valentyna Andrushchenko, Iryna Balagura

Institute for information Recording of NAS of Ukraine, Kyiv

dwlande@gmail.com

Abstract. The new index of the author's popularity estimation is represented in the paper. The index is calculated on the basis of Wikipedia encyclopedia analysis (Wiki-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. There are proposed algorithms and the technique of the Wiki-Index calculation through the network encyclopedia sounding, the exemplified calculations of the index for the prominent researchers, and also the methods of the information networks formation – models of the subject domains by the automatic monitoring and networks information reference resources analysis.

Keywords: Wikipedia, Author's popularity estimation, Wiki- Index, Information networks, Subject domains

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. In 2005 the physician Jorge E. Hirsch from the California University established the most popular index – Hirsch Index [1]. 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 researcher work. 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 Wiki-Index of authors' popularity [2].

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.

Today Wikipedia (https://www.wikipedia.org/, https://en.wikipedia.org/ – English version, https://zh.wikipedia.org/ – Chinese version etc.) is the most visited site in the Internet, and one of the most popular encyclopedic resources covering all the disciplines, it provides answers to the most search engines queries. At this time only the English version of Wikipedia contains more than 5 million articles (German - more than 2 million, Chinese, Russian – more than 1 million).

It is known that Wikipedia does not publish original research results, but at the same time all the information and references are verified according to the Wikipedia citing policy [3,4]. All Wikipedia articles are open to be edited, so it can improve information, but no new information which need to be approved will not be published freely.

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 [5]. The methods of building networks of co-authors, the definition of significant nodes of the network structure, research citations and appropriate buildings are among them [6]. Also authors have studied the array of publications relating to the approaches to the assessment of citations and other aspects of the update, existence, filling, editing of the encyclopedic resource Wikipedia [7-9].

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 Wiki-Index computation

The authors suggested the following rules for calculating Wiki-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:

Wiki-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:

Wiki-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 author popularity level 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.

Example:

Let assume that the Wikipedia article with the highest number of references to author George Smith (in a given subject area) contains 100 references. The second – 20 documents, a third - 10, fourth – 5, fifth – 5, 4 more – only one link. So we have a number of values:

=100, =20, =10, =5, =5, =1, =1, =1, =1
1 article contains the number of references least=100;
2 articles contain the number of references least=20;
3 articles contain the number of references least=10;
4 articles contain the number of references least=4.
5 articles contain the number of references least=5.
There are no 6 articles that contain the number of references least 6.
In this case:

As follows,

Algorithm

In the process of the Wiki-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 Wiki-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.

The process of the subject domain model of the author forming is possible in two ways:

- The use of Wikipedia dump database (not really relevant, but the link is available) by which the full range of all possible concepts and relationships. The advantage of this approach - completeness of information, disadvantage - possible loss of accuracy due consideration of homonyms, going beyond the subject area, considerable calculation time;

- The use of the principle of network services sounding (small sample volume of important contents of large information networks for technological reasons cannot be subjected to a complete scan). The advantage of this approach – getting accurate information strictly within a several subject domain, solvation of the homonyms problem and a short calculation time.The main drawback – the possible slight completeness, which may be assessed by additional experiments.

Authors chose the second approach for the Wiki-Index computation while building its corresponding domain model chose the second approach, which was implemented as a software as a service.

Formation of subject domain by sounding Wikipedia

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

On the main national Wikipedia page in the search line the initial word is given, e.g. (for English version - «Albert Einstein», for Chinese one –阿尔伯特·爱因斯坦, etc.).

- The search window opens. It contains information about concept, according to the task on the Step 1. The initial word/word combination is a graph vertex, which will be formed as the result of scanning.
- 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.
- The next transition is made by the first not involved hyperlink from the examining pages.
- In text on the page to which the transition has been made the search of shortened researcher’s name (e.g., Einstein, 爱因斯坦) or tag (e.g., physics, relativity, 物理学, 相对性) is to be carried out.
- In case, if there is a shortened researcher’s name or tag 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.
- If there is no word/word combination in the text – the given graph branch is considered to be built.
- 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.
- 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.

Calculation of the Wiki-Index of author’s popularity

To compute the Wiki-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 (or in sections «参考文献»,

«外部链接», «參考資料», «资料来源», «参考资料» for the Chinese Wikipedia) is provided.

Herewith, the number of these mentions, which correlates values, is counted. If , 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 sequenceis formed, which is used to calculate Wiki-Index, according to the rules above.

Experimental section

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

In Fig. 1 shows the Gephi (http://gephi.org) visualization of domain model fragment 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,
- the largest nodes (Table 1):

Table 1. Description of the largest nodes for the Albert Einstein subject domain

Concept	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

(128 articles with the references, *WH* = 12)

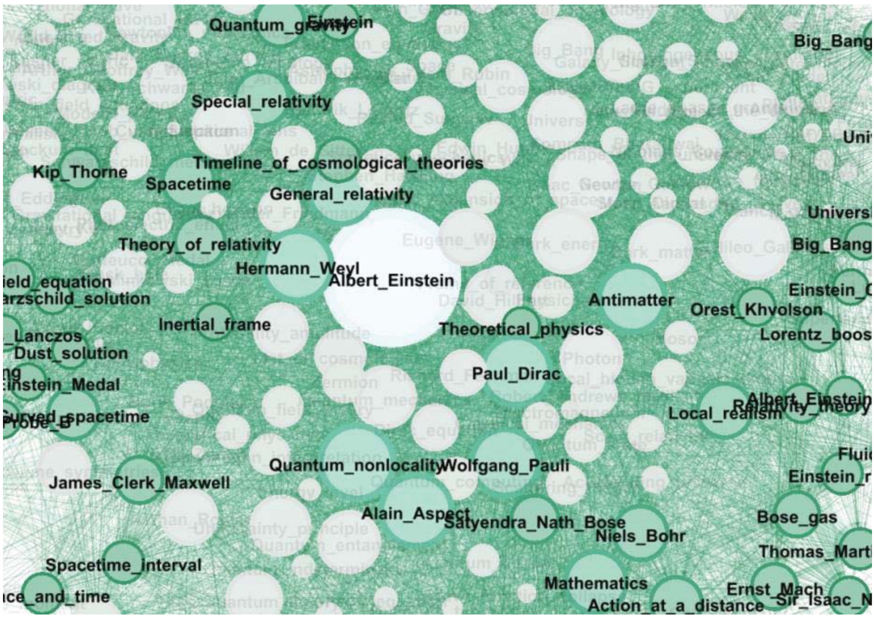

Fig. 1. Fragments of subject domains

Enrico Fermi:

- nodes – 605,
- edges – 22079,
- the largest nodes (Table 2):

Table 2. Description of the largest nodes for the Enrico Fermi subject domain

Concept	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

(92 articles with the references, WH=7)

Benoit Mandelbrot:

- nodes – 34,
- edges – 259,
- the largest nodes (Table 3):

Table 3. Description of the largest nodes for the Benoit Mandelbrot subject domain

Concept	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

(11 articles with the references, WH = 6)

There is an example on the Figure 2 of the subject domain fragment, which responds “A. Einstein” for the Chinese Wikipedia.

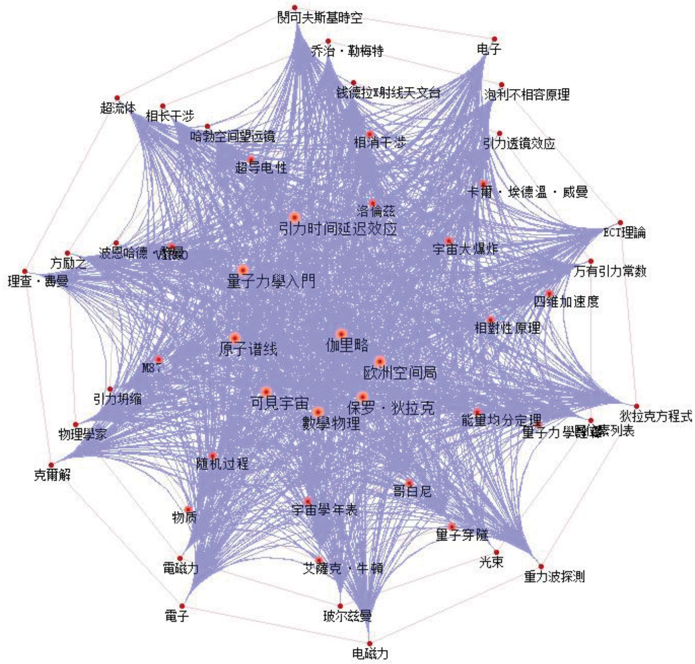

Fig 2. The fragment of subject domain for the Chinese Wikipedia

There were provided comparisons of the results – Wiki-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. Results are depicted in Table 4.

There are also were calculated the appropriate Wikipedia indices for the Chinese language Wikipedia that appeared an average 10-20% less.

Table 4. Comparison of Wiki-Indices values with the Hirsch-index (Scopus, Web Of Science and Google Scholar Citations)

N	Scientist	Wiki-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

As a result of calculations and proposed approaches tests to the formation of popular author index due to the presence of references to his/her work and references in the largest encyclopedic resource – Wikipedia, following conclusions can be made:

1 The principle of Wiki-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.

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

3 Due to the use and promotion of proposed indices there can be a significant expansion of open access resources (available to be edited by Internet users).

4 Provided work may be continued by analyzing other resources and the formation of indicators to estimate and analyze the influence in a particular environment.All the obtained results can be compared with those, which can be reached by analyzing other resources and the open access research on-line systems.

It is also necessary to note a fundamental difference between the proposed approach of automatic subject domains models formation 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. After that, the program uses the knowledge that is implanted by different languages Wikipedia articles' authors, tags defined by internal hyperlinks. This way expert area is widely extended.

References

1. Hirsch, Jorge E., An index to quantify an individual's scientific research output // E-preprint ArXiv. arxiv.org/abs/physics/0508025
2. Lande D.V., Andrushchenko V.B., Balagura I.V. Wiki-index of authors' popularity // E-preprint ArXiv. arxiv.org/abs/1702.04614
3. M. Pei, K. Nakayama, T. Hara and S. Nishio, Constructing a Global Ontology by Concept Mapping Using Wikipedia Thesaurus,22nd International Conference on Advanced Information Networking and Applications - Workshops (workshops 2008), Okinawa, 2008, pp. 1205-1210.
4. https://en.wikipedia.org/wiki/Wikipedia:Verifiability
5. 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.
6. Fei Wu and Daniel S. Weld. Automatically refining the wikipediainfobox ontology. In Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 2008, pp. 635-644.
7. Norlidah Alias, Dorothy DeWitt, SaedahSiraj, Sharifah Nor Atifah Syed Kamaruddin, MohdKhairulAzmanMdDaud, A Content Analysis of Wikis in Selected Journals from 2007 to 2012, Procedia - Social and Behavioral Sciences, Volume 103, 2013, pp. 28-36
8. Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, James R. Curran, Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence, Volume 194, 2013, pp. 151-175
9. F. Abbas, M. K. Malik, M. U. Rashid and R. Zafar, WikiQA — A question answering system on Wikipedia using freebase, DBpedia and Infobox,2016Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, 2016, pp. 185-193.

INTERNATIONAL CONFERENCE

COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

COLINS 2017

PROCEEDINGS

KHARKIV, UKRAINE

21 APRIL 2017

COLINS 2017

The 1st International Conference

COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

Proceedings of the Conference

21 April 2017

Kharkiv, Ukraine

The 1st International Conference
COMPUTATIONAL LINGUISTICS AND
INTELLIGENT SYSTEMS

PROCEEDINGS

Kharkiv, Ukraine
21 April 2017

Preface

The volume contains the papers presented at COLINS 2017: the 1st International conference “Computational linguistics and intelligent systems”.

The main purpose of the CoLInS conference is a discussion of the recent researches results in all areas of Natural Language Processing and Intelligent Systems Development.

The conference is soliciting literature review, survey and research papers comments including, whilst not limited to, the following areas of interest:

- mathematical models of language;
- artificial intelligence;
- statistical language analysis;
- data mining and data analysis;
- social network analysis;
- speech recognition;
- machine translation, translation memory systems and computer-aided translation tools;
- information retrieval;
- information extraction;
- text summarization;
- computer lexicography;
- question answering systems;
- opinion mining;
- intelligent text processing systems;
- computer-aided language learning;
- corpus linguistics;

The language of COLINS Conference is English.

The conference took the form of oral presentation by invited keynote speakers plus presentations of peer-reviewed individual papers. There was also an exhibition area for poster and demo sessions. A Student section of the conference for students and PhD students run in parallel to the main conference.

This year Organization Committee received 13 submissions, out of which 12 were accepted for presentation as a regular papers. The papers are submitted to the following tracks: corpus linguistics (1 paper), computational lexicography (2 papers), automatic ontology building (2 papers), morphological analisys (2 papers), content analysis (2 papers), intelligent computer systems building (2 papers) and problem of classification (1 paper). The papers directly deal with such languages: Ukrainian, Russian, Spanish, French, English, Polish and Danish.

COLINS 2017 Organization Committee:
Olga Kanishcheva (National Technical University “KhPI”, Ukraine)
Olga Cherednichenko (National Technical University “KhPI”, Ukraine)
Natalia Borysova (National Technical University “KhPI”, Ukraine)
Victoria Vysotska (Lviv Polytechnic National University, Ukraine)

Table of Contents