Impactful Scholar Map Based on Citation Network and Influence Numbers

Haoting Chen
hchen703@gatech.edu

Chen Lin
clin449@gatech.edu

Haojun Song
hsong343@gatech.edu

Zhaoyu Sun
zysun@gatech.edu

Cheng Zhang
czhang743@gatech.edu

1 INTRODUCTION

Over the last decades, multiple comprehensive University ranking analyses emerged to evaluate the higher education institutions through a variety of evaluation criteria [15]. For college and graduate school applications, people always refer to some well-known ranking organizations such as U.S. News Reports and QS World University Rankings for rankings [12]. However, for QS World, the methodology is based on academic review, faculty and student ratio, citations per faculty, employer reputation, international student ratio and international staff ratio. The number of citations per faculty counts toward 20 percent of the weighting, for which we claim the citation numbers here are biased [1, 7].

The purpose of having a global university ranking is to provide insight for scholars into the educational potential of a university [7]. It is a well-established fact that students tend to look for highly cited globally recognized researchers in their field. Therefore, a novel author citation-based visualization map is needed urgently for them.

2 PROBLEM DEFINITION

2.1 Formal Definition

The goal of the web interface system will provide a direct and interactive visualization which guides researchers to choose higher education organizations for studying abroad, scholar visiting, etc.

2.2 Jargon-free Definition

Here, we present to construct a web application that provides detailed author and citation information in each institution and have a ranking system with our customized impact factor models in a different field [11]. It will provide the top institutions around the globe or in a specific target area. And key authors in different subfields will be demonstrated with their notable journal paper that is ranked using our impact model. Furthermore, it helps researchers to discover the top impact authors' information in different region and academic field.

3 SURVEY

The citation number people often refer to as a standard for evaluating a paper is biased. For the citation number, according to the model of random citing scientist [14], the number might be high since a majority of scientific citations are copied from the list of references used in other papers causing the citation numbers of articles from the list increase exponentially, or the field of the paper is popular in recent years, or the paper introduces some major theorems from a subject. The number might be low if it is about a not that "popular" field of a subject or the average rate of citing decreases with the aging of a paper [6]. And the power-law distributions of citations of a paper published during the same year might also cause the various citation numbers [10]. The current citation network structures and their mathematical models only use one or two variables to find the new citation number for which we claim do not fit the current tendency of changes in citation numbers.

4 PROPOSED METHOD

4.1 Intuition

Since the current citation network structures and their mathematical models do not fit the current tendency of changes in citation numbers. We built directed graphs with citation networks and co-author networks for authors. The nodes represent individuals and links represent the citation interaction among those individuals referring to the structure mentioned in [9], which will be counted towards the influence number of authors.

Each node includes information about individuals' organizations, fields of study, and lists of citation numbers of publications. Our networks identify all authors related to a specific research topic and the subfield structured around it. Then use the citation networks to build a new model simulated by several mathematical models which consider all the factors that might cause an impact on citation numbers (refer Section 4.3.2 for details), assigns a new value for the "citation number" with citation interaction of a paper, called *influence number* of a paper, and gives authors' *influence scores* with influence numbers of individuals' publications and relation influence number with different ratios to show the key authors and their representative publications, and organizations and countries involved in the research [8].

For the influence number, since the traditional method to calculate authors' influence numbers only considers the following factors:

(1) using influence index to calculate the journal's influence, i.e., SNIP and SJR numbers
(2) using PageRank algorithm for the citation network such that the value for a paper is calculated by all the citations' influences recursively

For (1), SNIP (*Source Normalized Impact per Paper*) considers the ratio of a source's average citation and citation potential, and SJR (*SCImago Journal Rank*) only accounts the citations from high prestige journals being worth more than those from lower prestige. Though the methodologies are well-structured, (1) only uses a journals overall quality to judge an author's influence which is biased and not up-to-date. For (2), PageRank does not count the influences of new articles since usually new articles will be more innovative than the old ones though it may not have many citations yet but may be cited by authors with large values in the citation network.

4.2 Data

4.2.1 Dataset. We use SN SciGraph (Springer) and Semantic Scholar Academic Graph (S2AG) datasets.

Our raw dataset was retrieved from Springer, which aggregates data sources from Springer Nature and key partners in the academic field. The platform organizes information from the entire research field, including various datasets of articles, journals, people, institutions, etc [5].

However, the citation and reference information of the article in Springer is insufficient. We can not get the relationship between paper and paper, and author and author from it. Therefore, we added the S2AG dataset. The most important thing is that it has the citation and reference relationship of the articles, and provides monthly snapshots of research papers published in all fields. The Springer dataset is a subset of S2AG. So we use Springer's DOI to find the citation and reference information of the article in S2AG. And use these two properties as a supplement to the Springer properties.

4.2.2 Data Cleaning. We need to perform data cleaning on Springer and S2AG, and integrate their processed results. Because the data size is huge (>200GB), we do the data analysis with PySpark on AWS.

Springer: Screened out the articles published after 1960 with a clear author, journal, and field of research from Springer. All articles were grouped into 157 second-level disciplines (subfields) according to the subject two-level classification method used in the ANZ Standard Research Classification organization.

S2AG: Screened from AI data, articles with clear citation and reference information published after 1960.

Integrated: Taking the filtered Springer as the main database, 157 subfields (Springer) and S2AG screened article entries are placed on the AWS platform for matching. And we can get the citation relationship between paper and paper in each subfield in Springer, as well as the citation relationship between journal and journal. Finally, we can know the citation network (directed graph) between paper and paper in each subfield, and the citation network (directed graph) between journal and journal. These relationships will serve as input to a mathematical model of subsequent author ratings. (*Figure*1).

There are some improvements about our data cleaning process, for example, as mentioned in [9], the author disambiguation problem can be solved by using uniform variant.

4.2.3 Database. After data cleaning, we use SQLite to create the database, and the figure shown below (*Figure*2) is the relational schema with the key information

Due to the limited number of papers in some subfields, we finally selected 19 subfields to display in the visualization section. We merge the papers, authors, author-publication, and author scores of these 19 fields together as db1.

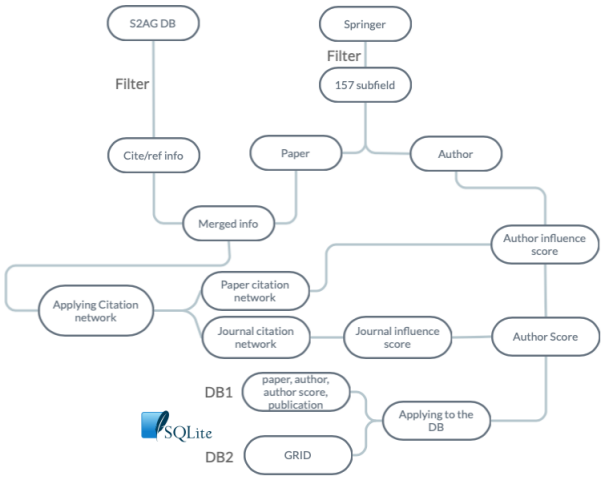

Figure 1: The Data Processing workflow

Name	Type	Schema
Tables (5)		
affiliation		CREATE TABLE affiliation (affiliation_id TEXT,affiliation_name TEXT,country TEXT)
ariticles		CREATE TABLE ariticles (DOI TEXT,title TEXT,citatiion_num INTEGER)
author		CREATE TABLE author (id TEXT,name TEXT,affiliation_id TEXT)
author _publication		CREATE TABLE author _publication (person_id TEXT,DOI TEXT)
author _score		CREATE TABLE author _score (id TEXT,score DECIMAL(10),field TEXT)

Figure 2: Database1 relational schema

We also use Global Research Identifier Database (GRID), to find the affiliation id, affiliation name, and country corresponding to the author in db1 as db2, to locate the author's current institution and the geographic location of the institution. In this way, it is convenient for the front end to retrieve information and visualize it.

4.3 Mathematical Model

4.3.1 Innovation. As mentioned in Chapter 3 and Section 4.1., for our model, we will calculate an author's score according to various factors by combining all the new scores we found based on the author's new influence index and the ratings of journals.

- Building citation networks to get authors' new influence numbers based on their field
- Building citation networks and math model to calculate new content rating scores for journals
- Creating new scores for authors

4.3.2 Citation Networks. For our citation network, by combining our old database with the new database, since it is hard to tell the first author of a paper, we choose to build directed acyclic citation network by the citation networks of papers (ex: *Figure*3). In this case, by

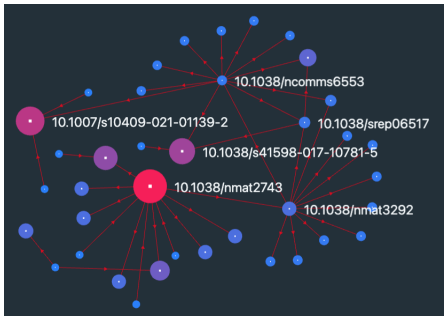

Figure 3: Citation Network of *10.1038/nmat2743*

choosing a most cited paper in the subfield, we can use it to construct the citation network for all the authors of the paper. For subfields of all subjects, we construct the working network which includes all the authors and the colleagues they published articles with. Since the graph is acyclic, it considers whether the citation is self-cited or not.

In our math model for the citation networks of authors, which will be discussed in next section, due to the limitation of citation lists of our old database, we were only able to build a undirected citation network. It is known that finding the longest path from a given node to every other node in general graph is NP-hard, and polynomial solutions for the problem only work for small classes of graphs. But after combining the new database, we restrict the input graph to be acyclic then we are able to solve the problem in polynomial time.

For the new journal ratings, we will also create citation networks (directed weighted graphs) for journals of every subfields in our database, generate adjacency matrix for each network, and build a new math model to calculate new journal ratings which will be discussed in Section 4.3.3.

4.3.3 Math Models. After cleaning the data, we will use the following equation to get author's influence index:

$$I = \sum \left(\frac{D_{c_i}}{D_{max}} \sum \lambda_i \right) \tag{1}$$

where I is the new influence number for the author, D_{c_i} is the max path between the author's node and a node in references, D_{max} is the max path of one node to other node in the subfield, and λ_i is 0.5 if there exits self-citation in the references for the i-th article found in our database, 1 if otherwise.

For the math model for journal rating, we will first divide all the journals by groups of different subfields, and for every subfield, we will define a constant and a random number which is calculated by journals SNIP and SJR numbers and use *analytical hierarchy process* (AHP) to decide the ratios of different factors. For every subfield, we will divide the n journals into $\lceil\sqrt{n}\rceil$ groups by their citation numbers with $\lceil\sqrt{n}\rceil$ in each group, assign initial values $\lceil\sqrt{n}\rceil$ for group 1, $\lceil\sqrt{n}\rceil - 1$ for group 2, ..., to X and create an equation for new rating of journals:

$$W_i = \alpha_0 X_i + \alpha_1 Y_i + \alpha_2 Z_i + S_i \tag{2}$$

where W_i is the new influence number of a journal i, X_i is the initial value number of the journal, S_i is the sum of SNIP and SJR scores of journal i, Y_i is the score if the journal is cited by journal from higher group, and Z_i otherwise such that

$$Y_i = \sum \theta P_{ij} X_j,\ Z_i = \sum \theta P_{ij} X_j, \tag{3}$$

where θ is 0 if barely no reference between journal i and j otherwise 1 (we decided to use median to determine the standard for "rarely"), and define $P_{ij} : P_{ij} = \frac{0.5Random}{e^{|X_i - X_j|/2}}$ where *Random* is a random number between 0 and 1 due to symmetry. For constant α_0, α_1 and α_2, i.e., ratios for X, Y and Z, we use AHP to create a standard to these numbers. Suppose we have a comparison matrix

$$A = \begin{bmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{bmatrix}$$

such that a_{xy} is the importance ratio of X and Y to W. Then by Thomas L. Saaty, who developed AHP, we can get the value of comparison matrix such that for our model, the impact ratio of X and Y to W is 5 and the impact ratio of Y and Z to W is 4. And by using Saaty scale, $\alpha : \alpha = (\alpha_0, \alpha_1, \alpha_2)^T$ and $\sum_{i=0}^{2} \alpha_i = 1$ (homogeneity) where α is the normalized eigenvector corresponds to the max eigenvalue of matrix A. To check the consistency of matrix A, following the properties [2], let $A = (a_{i,j})_{n\times n}$, if $a_{i,j}$ is

$$\begin{cases} a_{i,j} & > 0 \\ a_{i,j} & = \frac{1}{a_{i,j}} \\ a_{j,k}a_{i,j} & = a_{i,k} \end{cases}$$

such that $i, j, k = 1, 2, \ldots, n$, A is consistent. Since in general, A is inconsistent, the problem is to solve $A\mathbf{w} = \lambda_{\max}\mathbf{w}$ where $\lambda_{\max}$ is the largest unique eigenvalue that gives Perron eigenvector as an estimation of the priority vector. Then for the measure of inconsistency, by Saaty and [3], the consistency index CI and the consistency ratio CR are

$$\mathrm{CI} = \frac{\lambda_{\max} - n}{n-1},\ \mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \tag{4}$$

where RI is the average consistency index. When CR $<$ 0.1, the estimation is accepted otherwise we need a new comparison matrix until CR $<$ 0.1. In our model, by Saaty's experiment, for $n = 3$, we have RI $= 0.58$ and CR $= 4.7 \times 10^{-3} < 0.1$, so

$$\alpha = (\alpha_0, \alpha_1, \alpha_2)^T = (0.485, 0.415, 1) \tag{5}$$

4.3.4 Future Work. As mentioned in Seciton 4.3.1., the PageRank algorithm has large time complexity; however, in March 2022, [4] introduces a new ranking algorithm based on PageRank algorithm which solving the problem of bias. In the future, if we want to build larger citation networks, we may revise some parts of our method. And as mentioned in Section 4.3.3., due to the database we chose, there are some limitations of our model. Though our database updated the citation lists every month, we can add a model to mining periodic activities to see the distribution of citations referring the modified model of random-citing scientists [14] in the future. Also, unlike CSRanking, the website includes the current database of computer science faculty and the numbers of their publications (but its adjust number only count the total number of publications divide by the number of co-authors), whereas we only refer to the authors' publications in our database. In the future, we can pull out the total publications number of every author in our database from google scholars and the publications of current faculties (of the U.S. organizations) to get more precise results. Then for key authors, we can use the centrality metrics (total degree centrality), and for key papers, we can use the in-degree ranking (and even able to see the connection between country-to-country and institution-to-institution) to get more precise results than CSRanking. For journal math model, we can add new factors such as the different ratio α for different subfields, etc.

4.4 User Interface

4.4.1 Frontend Basic. In order to create a website that not only provides the data analysis and visualization,

but also explains our idea, algorithm, contact info and about us (like a real website). We chose React as it is a main stream front end framework from Meta (Formerly named Facebook)'s open-source JavaScript library, and Material UI from Google that is complementary to the react framework. The React is known for its lightweight progressive web apps that can mix uses HTML with JavaScript[13]. Combining the React and Material UI allows us easily create the necessary component like buttons or banners with a great simple design and use it as component to be used repeatedly. Additionally with the D3.js to create our dynamic interactive data visualization on the web.

The first thing users will select is the academic field list when they visit our website and enter it into our citation map visualization page. The page will send AJAX requests to our database API to request the information of the corresponding field. Then the map based on D3.js will refresh (*Figure*4).

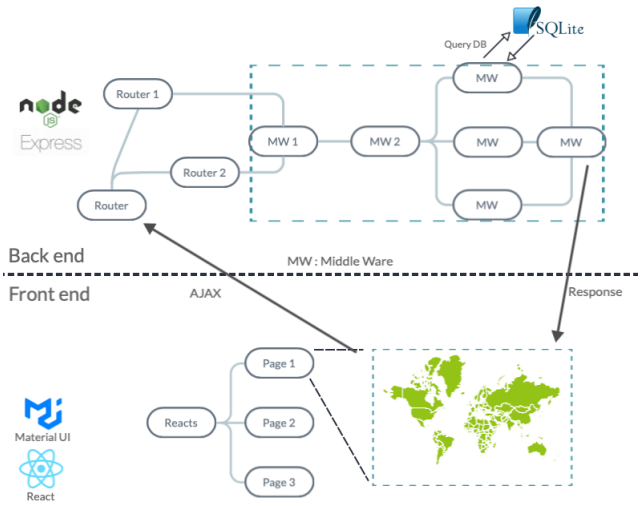

Figure 4: The front-back end Mechanism

4.4.2 Frontend Main. Our website provides three different ways to access our data: interactive visualized world map, table and API. The map and table will send AJAX requests to our back-end, and fit the data into the designated place after receiving the response. While the API will visit the URL directly. In addition to the pages on data, we have a home page where you can navigate to learn more about our project and get the contact information of us. These pages of main functions surely can be navigated easily at the top bar.

4.4.3 Backend. Our back end runs on Node.js, and we use the "Express" as our back-end framework. First, mount initial middle wares at the enter file and import our router. Then, mount different middle wares for different routes. Different middle wares on different routes take in different request parameters, which are extracted from the request. After routes matching, the query sentence will be modified accordingly, and then the query database. Finally, the data will be processed into a specific JSON format and responded to the client.

4.4.4 Citation network with bar chart. An interactive webpage that can reveal the world map is used, the visualization tool d3 extracts the geojson data along with the data (Country-ID, Country-score) from our SQLite data-frame to provide a choropleth map that can have different color intensity where darker blue means higher score obtained by the country, and Gray area showed no data retrieved.

By clicking on the colored country on the map, a bar chart that has the top 10 institutes for that country will be demonstrated, providing more insights for the user. Each bar of the organization is clickable that will be redirected to detailed author information and cited data web page for more information.

5 EXPERIMENTS/ EVALUATION

5.1 Model Evaluation

For our math models, Springer divides all articles into groups by ANZ Standard Research Classification organization, take *0101 Pure Mathematics* (mainly from Mathematical Physics) top 8 authors as an example,

Authors Ranked by Our Models

	AIS	JIS	CN-SS	CN-GS
N.Seiberg	395.47	614.09	49,209	71,970
E.Witten	259.12	444.88	143,905	230,548
A.Sen	167.34	302.21	24,156	35,457
M.R.Douglas	144.51	245.26	21,194	29,695
A.Connes	122.62	232.47	25,631	45,479
A.Strominger	117.37	222.43	37,099	58,380
S.Minwalla	122.07	209.13	10,792	
J.M.Maldacena	119.17	199.46	68,390	85,885

where AIS is the author influence score, JIS is the journal influence score, and CN-SS and CN-GS are the citation numbers from Semantic Scholar and Google Scholar. The table above shows that if sorted by the citation numbers from the resources, or only used the authors' influence scores, the ranking will be differ

than ours. Our models account various factors that cause the influence of a author such as his/her influence around collaborators and the research field, and the new rankings of journals, etc. And from the results above, we can see our new ranking system does provide different aspect of author's impact with powerful mathematical models.

5.2 UI Evaluation

For our UI interfaces, we conducted a questionnaire survey online and the users can give feedback with an option of inputting their background so that the survey is completely anonymous. But majority at nearly 60 percent claimed them as international students with total of 11 participants. The background of the participant showed the relatedness of our web server, and gave us confidence that their suggestions and feedback are valuable and constructive. The survey they completed consists of the question below and UI survey in *Figure*5:

(1) Do you classify yourself as an international student/scholar or who is willing to study aboard?
(2) Have you had a hard time choosing the institution when you thinking about studying abroad based on authors impact in a rank from 1 to 10?
(3) Do you think you need a citation-based author info map when you are studying abroad in a rank from 1 to 10?
(4) How intuitive is the interface to use in a rank from 1 to 10?
(5) Suggestions for any other features to include?

The average difficulty when choosing the institution with thinking about the author's impact is 8.3 which leans toward what we hypothesized and most of the participants think that our interface is intuitive. Moreover, there are two participants suggest us to adding a feature that can show the different authors score in a different time period like being able to choose from the start date of 2000 to the end date of 2022 and feature like zoom in view of the states of United States from the global view where they think there are more authors info about the US should be displayed. *Figure*5).

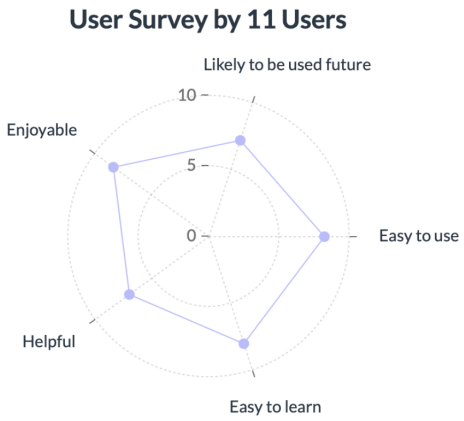

Figure 5: The User Survey on UI

6 CONCLUSIONS AND DISCUSSION

Our project collected more than 200 gigabytes data through Springer and S2AG, and processed for our citation network models. After applying the new author and journal indices models, information were extracted. We designed an interactive web browser to help researchers around the world better understand research progress. This website can display top institutions and top authors from different countries according to the research areas of interest to users.

There are also some future works for the improvement of our project as mentioned in Section 4.2.2. and Section 4.3.4. The survey about intuitive web design and the UI aesthetic provide some features that will be our future goals.

6.0.1 Distribution of Team Member effort. All team members have contributed a similar amount of effort.

NAME	ACTIVITIES
C. Zhang	Data cleaning, processing
C. Lin	Data processing, model construction
H. Chen	Data cleaning, database creation
Z. Sun	Frontend, backend development
H. Song	Data cleaning, visualization, poster Design

REFERENCES

[1] Farzana Anowar, Mustakim A. Helal, Saida Afroj, Sumaiya Sultana, Farhana Sarker, and Khondaker A. Mamun. 2015. A Critical Review on World University Ranking in Terms of Top Four Ranking Systems. In *New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering*, Khaled Elleithy and Tarek Sobh (Eds.). Springer International Publishing, Cham, 559–566.

[2] Julio Benítez, Xitlali Delgado-Galván, Joaquín Izquierdo, and Rafael Pérez-García. 2014. Achieving matrix consistency in AHP through linearization. *Applied Mathematical Modelling* 35, 9 (2014), 4449–4457. https://doi.org/10.1016/j.apm.2011.03.013

[3] J. Benítez, J. Izquierdo, R. Pérez-García, and E. Ramos-Martínez. 2014. A simple formula to find the closest consistent matrix to a reciprocal matrix. *Applied Mathematical Modelling* 38, 15 (2014), 3968–3974. https://doi.org/10.1016/j.apm.2014.01.007

[4] Sartawi B. Dayeh, M. and S. Salah. 2022. A Bias-Free Time-Aware PageRank Algorithm for Paper Ranking in Dynamic Citation Networks. *Intelligent Information Management* 14, 2 (2022), 53–70. https://doi.org/10.4236/iim.2022.142004

[5] Till Haselmann, Gunnar Thies, and Gottfried Vossen. 2010. Looking into a REST-Based Universal API for Database-as-a-Service Systems. In *2010 IEEE 12th Conference on Commerce and Enterprise Computing*. 17–24. https://doi.org/10.1109/CEC.2010.11

[6] Feng Hu, Lin Ma, Xiu-Xiu Zhan, Yinzuo Zhou, Chuang Liu, Haixing Zhao, and Zi-Ke Zhang. 2021. The aging effect in evolving scientific citation networks. *Scientometrics* 126, 5 (May 2021), 4297–4309. https://doi.org/10.1007/s11192-021-03929-8

[7] Mu-Hsuan Huang. 2012. Opening the black box of QS World University Rankings. *Research Evaluation* 21, 1 (02 2012), 71–78. https://doi.org/10.1093/reseval/rvr003 arXiv:https://academic.oup.com/rev/article-pdf/21/1/71/4632997/rvr003.pdf

[8] Fahmi H Kakamad, Q.S Rawezh, H.M Shvan, A.H Dahat, A.H Hunar, Snur Othman, Hawbash Mohammed, S.A Masrur, Jaafar Omer, and Hiwa Baba. 2018. Paper ranking; a strategy to embed research culture in developing countries. *International journal of surgery open* 15 (2018), 56–59.

[9] Miray Kas. 2011. *Structures and Statistics of Citation Networks.* Master's thesis. CARNEGIE-MELLON UNIVERSITY DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, PITTSBURGH, PA.

[10] Yurij L. Katchanov. 2015. Towards a simple mathematical theory of citation distributions. *SpringerPlus* 4, 1 (2015), 677–692. https://doi.org/10.1186/s40064-015-1467-8

[11] L Leydesdorff, S Carley, and I Rafols. 2013. Global maps of science based on the new Web-of-Science categories. *Scientometrics* 94, 2 (2013), 589–593.

[12] Adina-Petruta Pavel. 2015. Global University Rankings - A Comparative Analysis. *Procedia Economics and Finance* 26 (2015), 54–63. https://doi.org/10.1016/S2212-5671(15)00838-2 4th World Conference on Business, Economics and Management (WCBEM-2015).

[13] Farrukh Shahzad. 2017. Modern and Responsive Mobile-enabled Web Applications. *Procedia Computer Science* 110 (2017), 410–415. https://doi.org/10.1016/j.procs.2017.06.105 14th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2017) / 12th International Conference on Future Networks and Communications (FNC 2017) / Affiliated Workshops.

[14] M. V. Simkin and V. P. Roychowdhury. 2007. A mathematical theory of citing. arXiv:physics.soc-ph/physics/0504094

[15] Paul Taylor and Richard Braddock. 2007. International University Ranking Systems and the Idea of University Excellence. *Journal of Higher Education Policy and Management* 29, 3 (2007), 245–260. https://doi.org/10.1080/13600800701457855 arXiv:https://doi.org/10.1080/13600800701457855