Webology, Volume 12, Number 1, June, 2015

Home	Table of Contents	Titles & Subject Index	Authors Index

Aggregate ranking of the world's leading universities

Vladimir M. Moskovkin
Belgorod State University, Pobeda St., 85, 308015, Belgorod, Russian Federation.
E-mail: moskovkin@bsu.edu.ru

Nikolay A. Golikov
Independent Researcher, Kharkov, Ukraine. E-mail: kolia_forme@mail.ru

Andrey P. Peresypkin
Belgorod State University, Pobeda St., 85, 308015, Belgorod, Russian Federation.
E-mail: peressypkin@bsu.edu.ru

Olesya V. Serkina
Belgorod State University, Pobeda St., 85, 308015, Belgorod, Russian Federation.
E-mail: serkina@bsu.edu.ru

Received October 15, 2014; Accepted June 15, 2015

Abstract

The paper presents a methodology for calculating the aggregate global university ranking (Aggregated Global University Ranking, or AGUR), which consists of an automated presentation of the comparable lists of names for different universities from particular global university rankings (using Machine Learning and Mining Data algorithms) and a simple procedure of aggregating particular global university rankings (summing up the university ranking positions from different particular rankings and their subsequent ranking). The second procedure makes it possible to bring lists of universities from particular rankings, which are nonidentical by length, to one size. The paper includes a sample AGUR for six particular global university rankings as of 2013, as well as cross-correlation matrices and intersection matrices for AGUR for 2011-2013, all created by means of using the Python-based software.

Keywords

Aggregated Global University Ranking (AGUR); Times Higher Education (THE), QS World University Rankings; Academic Ranking of World Universities (ARWU); Higher Education Evaluation & Accreditation Council of Taiwan (HEEACT); CWTS Leiden Ranking; Scimago Institutions Rankings; University Ranking by Academic Performance (URAP); Webometrics; Machine learning; Data mining; Python; Cross-correlation matrix

Introduction

At the moment there are eight most significant global university rankings – THE, QS, ARWU, HEEACT, Leiden, SIR, URAP, and Webometrics. Each ranking relies on its calculation methodology and databases collected from universities (from 500 universities in HEEACT Ranking up to 25,000 universities in Webometrics Ranking). The first two rankings use both hard data and survey data, the next five rankings use exclusively hard data, and the last one uses the data obtained through testing sites with the help of Google and hard data from the SCImago laboratory on high-cited publications. THE, ARWU, HEEACT, Leiden, URAP rankings use in their methodologies hard data from The Web of Science, and the remaining three rankings rely on the Scopus database. We looked into these rankings in detail when we studied the process of joining them by the universities from the countries of the Mediterranean and Black Sea Regions (Moskovkin, Pupynina, Zaitseva and Lesovik, 2013). Considering different methodologies and scopes of global university rankings, we decided to develop in a software form a procedure of aggregating any number of particular global university ratings in the single aggregated ranking. We called this The Aggregated Global University Ranking (AGUR).

The task of constructing an aggregated global university ranking can be broken down into three sub-tasks:

a) selecting a mathematical model to construct an aggregated global university ranking;

b) receiving and pre-processing the information needed to build an aggregated global university ranking (a superposed ranking of universities based on particular global university rankings);

c) automated computation of an aggregated global university ranking and results analysis.

Further, these subtasks will be considered in this order.

Methods

a) Selecting a mathematical model to construct an aggregated global university ranking

There are a number of studies comparing ranked lists, including university rankings, among the most fundamental works is (Aguillo et al., 2010). The comparison of university rankings is studied in (Bar-Ilan et al., 2006). Of interest is the work by (Jurman et al., 2009), which considers the some general kinds of metrics (Canberra distance) applied in ranking. But the task of constructing an aggregated ranking and the task of comparing ranked lists are different tasks, though some approaches to solving the arising problems can be borrowed from the above-mentioned articles. To be more exact, the articles solved the problem of nonidentity of ranking structures, when some item could be present on one list, but absent on another. Applying this approach from the above articles to the problem of nonidentity of any two lists, we suggest calculating an aggregated ranking in the following way:

Suppose we have a ranked list of items R_i and R_j being l_i and l_j long respectively. Let there be given an item u: $u \in R_i$ then its rank $r_i(u)$ on this list equals its number $n_i(u)$ on this list. Let $r_j(u)$ be a rank of item u on list R_j, in this case if $u \in R_j$ then $r_j(u) = n_j(u)$ where $n_j(u)$ is the number of item u on list R_j, otherwise $r_j(u) = l_j + 1$. That is, if u belongs to list R_j, then rank of u on list R_j equals the number of u on this list, otherwise rank of u on list R_j equals $l_j + 1$, being a unity longer than list R_j. Based on this principle, Judit Bar-Ilan et al. in (Bar-Ilan, Mat-Hassan, Levene, 2006) built inter-ranking distance between the partially overlapping lists.

Further, we suggest calculating the aggregate rank r_k of item u_k in the following way: $r_k = \sum_{j=1}^{N} r_j(u_k)$, where $r_j(u_k)$ is rank of u_k on list R_j, N is the number of ranked lists ($j = 1,2,3...,N$ are the numbers of the lists), which means that the aggregate rank simply equals the sum of all the ranks for all the lists. Other more sophisticated approaches (without attributing weight factors to each of the ranks), like those in (Jurman et al., 2009), would produce no significant result, as the final ranking in the aggregated ranking would not change. But the authors had no reason to attribute weight factors to each ranking without a preliminary study of this issue.

b) Receiving and pre-processing the information needed to build an aggregated global university ranking

As for the second sub-task, it is obvious that it can and should be automated. For this purpose, we developed software to collect data from sites (scrapping), to process the information received and to calculate an aggregated ranking. As a programming language we chose Python (Python. Retrieved from http://www.python.org/doc/), because it is a very powerful and probably the most flexible of the common programming languages with a great number of specialized libraries designed to solve various problems. In particular, this language (and its libraries urllib2, requests (http://docs.python-requests.org) and the framework Scrapy) is de facto considered the standard for Internet scrapping. An essential point here is that information was obtained from various sources (websites), which are in fact completely different html-files, possibly with javascript, which called for creating a unified interface for parsing the received html-files (with use of Python library lxml (http://lxml.de) and, in some cases (for the Leiden ranking) the use of special facilities to study an http-session between the browser ("pretending" to be a scrapper) and the server.

Another problem was how to correctly (unambiguously) match names of universities from different rankings. The names would have short forms, different word order, or even include characters other than Latin. But through parsing html-pages, we managed to obtain enough information to unambiguously match names of universities. The further processing of the data was carried out by means of pandas library (McKinney, 2012, http://pandas.pydata.org/pandas-docs/stable/), designed for analyzing and statistical data processing.

We should also note here that some data from a number of rankings are available in formats other than html, for example, pdf format in SCImago or xls format for Leiden rankings (except for the year of 2013). We did not process such cases as they would have required extra time.

Results and discussion

c) Automated computation of an aggregated global university ranking and results analysis

For preliminary experiments with the software we developed, we downloaded various numbers of different universities with their ranks from official sites of global university rankings over three years (Table 1).

Table 1. Rankings scope over years

Rankings	2011	2012	2013
THE	402	400	400
QS	724	873	834
Leiden	0	500	500
ARWU	500	500	500
Webometrics	0	500	500
URAP	2000	2000	2000

The calculated aggregated global university ranking for six particular rankings is provided only for 2013 (Table 2).

Table 2. Aggregated global university ranking, 2013

New rank	Universities	THE	QS	Leiden	ARWU	Web ometrics	URAP	Sum of ranks
1	Harvard University	2	2	5	1	1	1	12
2	Stanford University	4	7	3	2	3	4	23
3	Massachusetts Institute of Technology (MIT)	5	1	1	4	2	14	27
4	University of California, Berkeley (UCB)	8	25	7	3	4	5	52
5	University of Cambridge	7	3	24	5	20	10	69
6	University of Oxford	3	6	30	10	18	6	73
6	Yale University	11	8	10	11	15	18	73
7	Columbia University	13	14	19	8	11	17	82
8	University of Pennsylvania	16	13	18	15	10	13	85
9	University of California, Los Angeles (UCLA)	12	40	25	12	5	9	103
10	University of Chicago	9	9	16	9	23	42	108
10	Cornell University	19	15	32	13	8	21	108
11	California Institute of Technology (Caltech)	1	10	8	6	41	49	115
12	University of Michigan	18	22	40	23	7	8	118
13	Johns Hopkins University	15	16	36	17	33	2	119
14	Princeton University	6	11	4	7	19	86	133
15	University of Washington	25	59	22	16	6	7	135
16	Duke University	17	23	29	31	27	20	147
17	Northwestern University	22	29	17	30	28	28	154
18	UCL (University College London)	21	4	50	22	39	19	155
19	University of California, San Diego (UCSD)	41	63	15	14	30	12	175
20	University of Wisconsin-Madison	30	37	54	19	14	23	177
21	University of Toronto	20	18	88	28	24	3	181
22	New York University (NYU)	40	44	28	27	32	55	226
23	University of British Columbia	31	49	99	40	16	24	259
23	University of Texas at Austin	27	73	39	36	17	67	259
24	University of Illinois at Urbana-Champaign	29	56	67	25	26	74	277
25	University of Minnesota	46	103	61	29	9	35	283
26	University of Edinburgh	39	17	84	51	52	47	290
27	National University of Singapore (NUS)	26	24	73	114	54	41	332
28	University of California, Davis	53	85	82	47	44	32	343
29	McGill University	35	21	131	58	72	27	344
30	Boston University	50	79	37	75	62	58	361
31	University of Pittsburgh	78	106	65	63	42	22	376
32	University of California, Santa Barbara (UCSB)	33	130	2	35	68	110	378
33	Pennsylvania State University	49	107	112	54	12	50	384
34	Utrecht University	75	81	64	53	69	43	385
35	RMIT University	34	291	9	18	35	11	398
36	University of Southern California	71	125	52	48	37	70	403
37	Ohio State University	59	113	113	65	34	34	418
38	Carnegie Mellon University	24	57	21	52	38	234	426
39	Georgia Institute of Technology	28	99	34	105	61	125	452
40	University of Colorado Boulder	97	160	14	33	49	101	454
41	Brown University	52	47	74	67	85	141	466
41	Purdue University	62	101	124	57	22	100	466
42	Leiden University	67	74	58	74	125	72	470
43	University of California, Irvine	93	149	44	45	58	83	472
44	Australian National University	48	27	127	66	75	145	488
45	University of Copenhagen	150	45	132	42	96	25	490
45	Texas A&M University	159	153	12	46	29	91	490

…

New rank	Universities	THE	QS	Leiden	ARWU	Webometrics	URAP	Sum of ranks
1826	University of Bolton	401	835	501	501	501	1971	4710
1827	Arkansas State University	401	835	501	501	501	1972	4711
1827	Johannes Kepler Universität Linz	372	835	501	501	501	2001	4711
1828	G B Pant University of Agriculture & Technology	401	835	501	501	501	1973	4712
1829	University of Calcutta	401	808	501	501	501	2001	4713
1829	Bozok University	401	835	501	501	501	1974	4713
1830	Dalian University	401	835	501	501	501	1975	4714
1831	University of Colombo	401	810	501	501	501	2001	4715
1831	Blekinge Institute of Technology	401	835	501	501	501	1976	4715
1832	Sastra University	401	835	501	501	501	1977	4716
1832	University of Buenos Aires	401	835	477	501	501	2001	4716
1833	Colby College	401	835	501	501	501	1978	4717
1834	University of Engineering & Technology (UET) Lahore	401	813	501	501	501	2001	4718
1834	Adelphi University	401	835	501	501	501	1979	4718
1834	University of Minho	379	835	501	501	501	2001	4718
1835	Seikei University	401	835	501	501	501	1980	4719
1836	Kanazawa Institute of Technology	401	835	501	501	501	1981	4720
1837	Indiana University – Purdue University Fort Wayne	401	835	501	501	501	1982	4721
1837	University of Quebec	401	835	501	482	501	2001	4721
1838	Maharshi Dayanand University	401	835	501	501	501	1983	4722
1839	University of Tunis	401	835	501	501	501	1984	4723
1840	United States Military Academy at West Point	401	835	501	501	501	1985	4724
1840	University of Rovira i Virgili	385	835	501	501	501	2001	4724
1841	National Technical University of Ukraine	401	835	501	501	501	1986	4725
1842	Universidade do Vale do Rio Dos Sinos	401	835	501	501	501	1987	4726
1842	National Autonomous University of Mexico	401	835	487	501	501	2001	4726
1842	Università degli Studi di Napoli Federico II	401	835	501	501	487	2001	4726
1843	Western Carolina University	401	835	501	501	501	1988	4727
1843	Università degli Studi di Genova	401	835	501	501	488	2001	4727
1844	University of Santo Tomas	401	823	501	501	501	2001	4728
1844	Lviv Polytechnic National University	401	835	501	501	501	1989	4728
1845	Universite Rennes 2 Haute Bretagne	401	835	501	501	501	1990	4729
1846	Free University of Bozen Bolzano	401	835	501	501	501	1991	4730
1847	Josip Juraj Strossmayer University of Osijek	401	835	501	501	501	1992	4731
1847	Federal University of São Paulo	401	835	492	501	501	2001	4731
1848	University of Kashmir	401	835	501	501	501	1993	4732
1849	North China University of Water Conservancy and Electric Power	401	835	501	501	501	1994	4733
1849	Federal University of Paraná	401	835	494	501	501	2001	4733
1850	Irkutskij Gosudarstvennyj Universitet	401	835	501	501	501	1995	4734
1851	Technological Education Institute of Athens	401	835	501	501	501	1996	4735
1851	University of Tromsø	396	835	501	501	501	2001	4735
1852	Lingnan University	401	835	501	501	501	1997	4736
1853	Mehmet Akif Ersoy University	401	835	501	501	501	1998	4737
1854	Vytautas Magnus University	401	833	501	501	501	2001	4738
1854	University of Miskolc	401	835	501	501	501	1999	4738
1854	Federal University of Viçosa	401	835	499	501	501	2001	4738
1854	University of Vigo	399	835	501	501	501	2001	4738
1855	West University of Timisoara	401	834	501	501	501	2001	4739
1855	Acharya Nagarjuna University	401	835	501	501	501	2000	4739
1855	York University	401	835	501	500	501	2001	4739

When collecting the data for 2013-2014, the Taiwanese website HEEACT Ranking was not in operation, and the SIR website held the annual reports in form of pdf-files that are difficult to discern. As the resulting rankings tables are voluminous, in the present paper we provide only the first and last lines for each table of the aggregated ranking, but no more than fifty lines.

We also provide the cross-correlation matrices for all three years (the elements of which were calculated for each pair of the ranking by using the Spearman rank correlation formula) (Tables 3-5).

Table 3. Cross-correlation matrix, 2011

	New rank	THE	QS	ARWU	URAP	Sum of ranks
New rank	1.00	0.48	0.28	0.46	0.94	1.00
THE	0.48	1.00	0.53	0.61	0.34	0.48
QS	0.28	0.53	1.00	0.44	0.06	0.28
ARWU	0.46	0.61	0.44	1.00	0.34	0.46
URAP	0.94	0.34	0.06	0.34	1.00	0.94
Sum of ranks	1.00	0.48	0.28	0.46	0.94	1.00

Table 4. Cross-correlation matrix, 2012

	New rank	THE	QS	Leiden	ARWU	Webometrics	URAP	Sum of ranks
New rank	1.00	0.50	0.36	0.50	0.49	0.54	0.88	1.00
THE	0.50	1.00	0.51	0.65	0.60	0.63	0.33	0.50
QS	0.36	0.51	1.00	0.47	0.45	0.50	0.06	0.36
Leiden	0.50	0.65	0.47	1.00	0.69	0.59	0.34	0.50
ARWU	0.49	0.60	0.45	0.69	1.00	0.53	0.36	0.49
Webometrics	0.54	0.63	0.50	0.59	0.53	1.00	0.39	0.54
URAP	0.88	0.33	0.06	0.34	0.36	0.39	1.00	0.88
Sum of ranks	1.00	0.50	0.36	0.50	0.49	0.54	0.88	1.00

Table 5. Cross-correlation matrix, 2013

	New rank	THE	QS	Leiden	ARWU	Webometrics	URAP	Sum of ranks
New rank	1.00	0.50	0.39	0.49	0.48	0.54	0.89	1.00
THE	0.50	1.00	0.53	0.64	0.60	0.63	0.31	0.50
QS	0.39	0.53	1.00	0.46	0.43	0.50	0.10	0.39
Leiden	0.49	0.64	0.46	1.00	0.69	0.60	0.33	0.49
ARWU	0.48	0.60	0.43	0.69	1.00	0.55	0.34	0.48
Webometrics	0.54	0.63	0.50	0.60	0.55	1.00	0.39	0.54
URAP	0.89	0.31	0.10	0.33	0.34	0.39	1.00	0.89
Sum of ranks	1.00	0.50	0.39	0.49	0.48	0.54	0.89	1.00

URAP is best correlated to AGUR (new rank, sum of ranks) and poorly with the rest of the particular rankings; THE is correlated to Leiden, ARWU and Webometrics; QS is correlated to THE; Leiden is correlated to THE, ARWU and Webometrics; ARWU is correlated to THE and Leiden; and Webometrics is best correlated to THE and Leiden.

We also built matrices of pair intersections of universities names in various rankings in absolute values (Tables 6-8).

Table 6. Intersection matrix, 2011

	THE	QS	Leiden	ARWU	Webometrics	URAP
THE		315	268	281	0	328
QS			0	307	0	404
Leiden				0	0	0
ARWU					0	405
Webometrics						0
URAP						

Table 7. Intersection matrix, 2012

	THE	QS	Leiden	ARWU	Webometrics	URAP
THE		312	274	276	268	309
QS			359	346	354	481
Leiden				361	287	379
ARWU					284	402
Webometrics						429
URAP						

Table 8. Intersection matrix, 2013

	THE	QS	Leiden	ARWU	Webometrics	URAP
THE		315	268	269	260	308
QS			348	331	347	499
Leiden				363	290	381
ARWU					247	399
Webometrics						432
URAP						

They are easy to compare for rankings of the same length, which is the case for Leiden, ARWU and Webometrics for which we downloaded TOP-500 universities (Table 9).

Table 9. Intersection matrix for rankings with the same number of university-participants, %

	Leiden		ARWU		Webometrics	
	2012	2013	2012	2013	2012	2013
Leiden	x	x	72.2	72.6	57.4	58.0
ARWU			x	x	56.8	49.4
Webometrics					x	x

As expected, the best intersection of universities names was obtained for Leiden and ARWU rankings, because their methodologies were based on hard data from Web of Science.

Conclusion

Thus, the paper presents a methodology for calculating the (Aggregated Global University Ranking, or AGUR), which consists of an automated presentation of the comparable lists of names for different universities from particular global university rankings (using Machine Learning and Mining Data algorithms) and a simple procedure of aggregating particular global university rankings (summing up the university ranking positions from different particular rankings and their subsequent ranking). The first procedure allowed us to solve the problem of creating a unified interface for parsing the received html-files based on diverse sources (sites), the information from which is provided in completely different html-files. Besides, the problem of unambiguous matching the names of universities in various global university rankings was solved. As a result of parsing html-pages, we can extract enough information to unambiguously identify the names of the universities. The further processing of the data was carried out by means of pandas library. The second procedure (based on Aguillo et al., 2010; Bar-Ilan et al., 2006; Jurman et al., 2009) makes it possible to bring lists of universities from particular rankings, which are nonidentical by length, to one size. As an example of the functioning of the Python-based software, the paper includes AGUR calculations for six particular global university rankings as of 2013, as well as cross-correlation matrices and intersection matrices for AGUR for 2011-2013.

URAP is best correlated to AGUR (new rank, sum of ranks) and poorly with the rest of the particular rankings; THE is correlated to Leiden, ARWU and Webometrics; QS is correlated to THE; Leiden is correlated to THE, ARWU and Webometrics; ARWU is correlated to THE and Leiden; and Webometrics is best correlated to THE and Leiden.

We have built the matrices of pairwise intersections of university names in various rankings in absolute values. They are easy to compare for same-size rankings, which is the case for Leiden, ARWU and Webometrics, for which we downloaded TOP-500 universities. The best intersection of university names was recorded, as we had expected, for Leiden and ARWU rankings because their methodology was based on hard data from Web of Science.

For further AGUR calculations for 2014 and following years, we find it possible to add SIR, since Scimago Lab has begun to provide the data in normal editable formats, rather than pdf-files, HEEACT Ranking, which has obtained support from the National Taiwan University (until 2012 it had been supported by the Higher Education Evaluation and Accreditation Council of Taiwan (HEEACT)). Besides, we consider it feasible to use a number of new rankings which have recently appeared (Round University Ranking (RUR), U-Multirank, and Global World Communicator (GWC) - Worldwide Professional University Rankings).

Acknowledgements

Research was done according to the Government task of the Ministry of Education and Science of the Russian Federation for 2015, project code -516.

References

Aguillo, I.F., Bar-Ilan, J., Levene, M., & Priego, L.O.J. (2010). Comparing university rankings. *Scientometrics*, 85(1), 243-256.

Bar-Ilan, J., Mat-Hassan, M., & Levene, M. (2006). Methods for comparing rankings of search engine results. *Computer Networks*, 50(10), 1448-1463.

Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C. (2009). Canberra distance on ranked lists. In S. Agrawal, C. Burges, & K. Crammer (Eds.), *Proceedings of Advances in Ranking NIPS 09 Workshop* (pp. 22-27).

McKinney, W. (2012). *Python for data analysis*. Sebastopol, CA: O'Reilly Media, Inc.

Moskovkin, V.M., Pupynina, E.V., Zaitseva, N.P., & Lesovik, R.V. (2013). Methodology for comparative analysis of university rankings with the Mediterranean and Black Sea region countries taken as an example. *Middle-East Journal of Scientific Research*, 18(11), 1656-1665.

Pandas (2015). *Pandas: Powerful Python data analysis toolkit*. Retrieved March 10, 2015, from http://pandas.pydata.org/pandas-docs/stable/

Python (2015). *Python*. Retrieved March 10, 2015, from http://www.python.org/doc/

Requests (2015). *Requests: HTTP for Humans*. A Kenneth Reitz Project. Retrieved March 10, 2015, from http://docs.python-requests.org

LXML (2015). *lxml - XML and HTML with Python*. Retrieved March 10, 2015, from http://lxml.de

Bibliographic information of this paper for citing:

Moskovkin, Vladimir M., Golikov, Nikolay A., Peresypkin, Andrey P., & Serkina, Olesya V. (2015). "Aggregate ranking of the world's leading universities." *Webology*, 12(1), Article 133. Available at: http://www.webology.org/2015/v12n1/a133.pdf