Ranking Universities via Clustering

George Stoupas[1], Antonis Sidiropoulos[2], Dimitrios Katsaros[3], Yannis Manolopoulos[1,4]

[1]Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[2]International Hellenic University, 57001 Thessaloniki, Greece
[3]University of Thessaly, 38221 Volos, Greece
[4]Open University of Cyprus, 2220 Nicosia, Cyprus

{grgstoupas,manolopo}@csd.auth.gr, *asidirop@ihu.gr*, *dkatsar@uth.gr*, *yannis.manolopoulos@ouc.ac.cy*

Introduction

University rankings have been approached by academicians, politicians, journalists, and policy makers, even though it is surrounded with scepticism (Manolopoulos & Katsaros, 2017). University rankings are especially conspicuous for the top universities (Angelis, Bassiliades & Manolopoulos, 2019).

All university rankings basically agree on the order of the top-20 or top-30 universities, with major disagreements appearing thereafter. However, one could still wonder whether there is *really* a *serious* performance or prestige discrepancy between universities e.g., either in the range [400-500] or even in the ranges [2-5] or [6-15].

Problem description

Following the reasoning that there is small justification in providing absolute and rigid ranked lists of universities, but it is more meaningful to provide ranked sets, we will use clustering methodologies to achieve our goal. So, the aim of the present study is to use (and evaluate the appropriateness of) popular clustering algorithms to develop university rankings with the defining characteristic that competing universities are “organized” into sets where ranking/ordering is imposed only among the sets, whereas the elements within any set are considered unordered. This concept comprises a departure from existing methodologies where – at least for the first, say, 150 positions – there is a *strict* ranking/ ordering among universities, i.e., total order, whereas we seek for an approach developing partially ordered sets of universities (posets).

Dataset

We choose to work with the National Taiwan University Ranking (NTU) (http://nturanking.csti.tw), founded by the Higher Education Evaluation and Accreditation Council of Taiwan (HEEACT), because it is based exclusively on verifiable research performance indicators. NTU ranking is based on eight features categorized into three categories: *research productivity*, *research impact* and *research excellence* with weights 25%, 35% and 40%, respectively. We work with the top-500 universities of the NTU list.

Experimentation

We used the Weka software (Witten, Frank & Hall, 2011), which offers machine learning libraries, including many algorithms for clustering. For our experiments, we choose the most popular representative of each of three well-known families of clustering algorithms, namely:

- Expectation Maximization (EM) (model-based algorithm)
- DBSCAN (density-based algorithm)
- *k*-means (center-based algorithm)

Our experimental methodology was the following: First, we used the EM-algorithm with the default Weka values, and got 12 clusters. Second, we examined DBSCAN with its default Weka values, except that we set *minpoints*=1, so that we do not lose any outlier, plus we varied parameter *e* with step 0.01 to get the number of clusters. By using the Elbow method (Thorndike, 1953), we came up with 43 clusters. Using the previous findings, we performed a set of experiments on the 3 algorithms for 12 and 43 clusters. From Table 1 we see that DBSCAN does not provide any useful insight for the particular dataset, whereas EM and *k*-means give similar results in terms of the maximum cluster size and the number of singleton clusters.

Table 1. Statistics of the examined algorithms.

	#clusters	max cluster size	#singleton clusters
DBSCAN	12	488	10
	43	430	40
EM	12	90	1
	43	33	2
k-means	12	99	1
	43	23	1

To rank clusters, we assign to each cluster a value equal to the sum of the median values of each of the 8 features mentioned in the “Dataset section”. Thus, we order the clusters from 1 (the highest quality) to 12 (or 43).

Table 2 shows the size of each cluster for each of the three algorithms, with EM and *k*-means clusterings bearing similarity in terms of clusters’ cardinality.

Table 2. Size of the 12 clusters per algorithm.

	DBSCAN	EM	*k*-means
1	1	1	1
2	1	17	16
3	1	42	28
4	1	34	13
5	1	48	14
6	1	33	55
7	1	44	29
8	2	49	77
9	1	42	41
10	1	90	99
11	1	70	72
12	488	30	55

In Table 2, Harvard is always the (singleton) cluster #1. Cluster #2 of *k*-means and EM consists of: Berkeley, Cambridge, Columbia, Imperial College, Johns Hopkins, Michigan/Ann Arbor, MIT, Oxford, Pennsylvania, Stanford, Toronto, UCollege London, UCLA, UCSD, UCSF and Washington/Seattle. In addition, EM's cluster #2 includes Tsinghua. On the other hand, DBSCAN is much closer to the existing methodologies of university rankings.

Although NTU provides an ordered list, there are tie cases that can be considered as ranked clusters; thus, NTU rank can be viewed as a set of 406 clusters. Table 3 shows the Rand Index (RI) (Rand, 1971) for each pair of clustering algorithms as well as with respect to the NTU ranking for 12 (in parenthesis, for 43) clusters.

Table 3. Rand Index for 12 (43) clusters.

	EM	*k*-means	NTU
DBSCAN	0.147 (0.284)	0.167 (0.280)	0.048 (0.261)
EM	--	0.878 (0.973)	0.890 (0.971)
k-means	--	--	0.878 (0.973)

Oppositely, we view each cluster of the EM or *k*-means algorithm as a list of equal performing universities, ordered according to the position in the NTU list. Thus, we apply the Spearman rank correlation coefficient (ρ) for all pairs of algorithms in Table 3, getting Table 4.

Table 4. Spearman ρ for 12 (43) clusters.

	EM	*k*-means	NTU
DBSCAN	0.230 (0.509)	0.239 (0.500)	0.242 (0.503)
EM	--	0.905 (0.968)	0.960 (0.978)
k-means	--	--	0.896 (0.973)

It seems that *k*-means is the most appropriate clustering algorithm among the three to carry out our goal. Using this algorithm, we investigated the following question which might be crucial for university administrators: Is there any particular feature or set of features (among the eight ones examined) that affect the clustering the most, and which are the methodologies to discover it/them?

Weka has already built-in support for such kind of questions. In Table 5, we see that for the *k*-means, the *average number of citations in the last 11 years* (AveCit, cell in grey) is the most important and affects the ranking dramatically; the rest of the features present very similar rankings with the case when all the 8 features are used. For instance, the features HiCit (*number of citations in the last 11 years*) and CurCit (*number of citations in the last 2 years*) do not affect the ranking as their Spearman (Rand Index) value is close 1.

Table 5. Spearman *ρ* (Rand Index) for feature elimination in the *k*-means.

	8 features without HiCit	8 features without CurCit	8 features without AveCit
Full set of 8 features	0.9671 (0.9834)	0.9568 (0.9396)	0.5190 (0.8519)

Conclusions

We argue for the representation of university rankings in the form of ordered clusters. In the present work, we clustered universities from the NTU ranking with DBSCAN, EM and *k*-means, the last one being the most appropriate. Furthermore, looking at the clusters produced, with the exception of the top (singleton) cluster, the second position is occupied by 16 (17) universities, and based on the concepts developed here, these universities are performing equally well, and thus they can be ranked into the same position.

References

Angelis, L., Bassiliades, N. & Manolopoulos Y. (2019). On the necessity of multiple university rankings. *Collnet Journal of Scientometrics & Information Management*, 13(1), 11-36.

Manolopoulos, Y. & Katsaros, D. (2017). Metrics and rankings: Myths and fallacies. In *Revised selected papers, 18th DAMDID Conference*, pp.1-16, Springer.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *JASIST*, 66(336), pp. 846-850.

Thorndike, R. (1953). Who belongs in the family?, *Psychometrika* 18(4), pp.267-276.

Witten, I., Frank, E., Hall, M. & Pal, C. (2011). *Data mining: Practical machine learning tools and techniques*, 3rd edition, Morgan Kaufmann.