doi: 10.2478/v10313-012-0002-y
2012 /15

The Analysis of Rating Data of the Latvian Higher Education Institutions Using Clustering

Peter Grabusts, *Rezekne Higher Educational Institution*

***Abstract* – Rating data of the Latvian public higher education institutions for the year 2012 has been used as the input data, and the aim of the experiment has been to show how by applying clustering methods the mentioned data can be analyzed in an alternative way. During the research an attempt has been made to group higher education institutions with the help of k-means clustering algorithm and to verify whether such division corresponds to the rate of a certain higher education institution in the rating data calculated mathematically. The validity of clustering has been evaluated with the help of Rand index.**

***Keywords* – clustering algorithms, cluster validity, k-means, Rand index, rating data**

I. Introduction

Rating data of the Latvian higher education institutions has been published since 2008 [9] – [13]. In general cases, the rating is made up of indicator values chosen in a definite way that can be multiplied by a significance measure – weight. The obtained numbers are summed and the resulting value defines the position in the rating table. The further analysis of such a rating table arouses interest. In the research, an attempt has been made to group public higher education institutions with the help of k-means clustering algorithm and to make sure whether such distribution corresponds to the mathematically calculated position of the institution in the rating table.

In order to evaluate the efficiency aspects of the performance of clustering algorithms, the following aim has been set – to perform the analysis of rating data of the Latvian public higher education institutions for the year 2012. Research tasks are subordinated to the aim set: to describe the changes in the number of clusters with respect to the data under analysis and to evaluate the reliability of clustering results. The research aims to show that by applying clustering methods it is possible to analyze such data in an alternative way.

II. Rating System

In order to evaluate the impact of parameters characterizing clustering results, the research has been conducted, where the rating table of Latvian public higher education institutions for the year 2012 has been used [13].

International ratings of higher education institutions are becoming more popular. Different methodologies exist with respect to determining the rating of higher education institutions.

Webometrics Ranking rates more than 20 000 higher education institutions in the world [14]. The rating is based only on the information about the institutions available on the Internet. Four main indicators are used: 10% of rank value is formed by the recognition of the institution in Google search engine, 50% – by the number of external links to the home page of higher education institution, 10% – by the academic and publishing activities in different file formats in Google search engine (.doc, .pdf, .ppt), 30% – by the number of electronic publications from Google Scholar (2007–2011) and data from SCImago Institutions Rankings (SIR) (2003–2010).

According to Webometrics Ranking, LU is ranked 796th, RTU – 1403rd, LLU – 2599th, RA – 3909th, DU – 4477th, RSU – 6971st.

The SCImago SIR rates 3042 higher education institutions in the world and is based on the data about the scientific activities of higher education institution [15]. Four indicators include the information about the number of publications (mostly SCOPUS), indicators of scientific cooperation, number of high level publications, etc. Among Latvian higher education institutions LU (rank No. 1565) and RTU (rank No. 2794) are mentioned here.

The QS World University Rankings consists of the list of 700 world's leading higher education institutions [16]. Six indicators are used: 40% is formed by academic reputation, 10% – by employer reputation, 20% – by citation of scientific papers, 20% – by the number of students, 5% – by the number of foreign students, 5% – by the number of international faculties. Latvian higher education institutions are not represented in the rating table of the QS World University Rankings.

The Times Higher Education World University Rankings (THE) forms the list of 400 world's leading higher education institutions [17]. 13 indicators divided into 5 groups are used: learning environment (30%), research activities (30%), citations (30%), innovations (2.5%), and foreign relations (7.5%). Latvian higher education institutions are not represented in the rating table of the Times Higher Education World University Rankings.

To form the rating of the Latvian higher education institutions, the evaluation criteria or indicators are the following [9]:

- I1 – the ratio of the number of students and academic staff (weight = 1);
- I2 – the number of the graduates (weight = 0.5);
- I3 – the number of academic staff employed on permanent contracts possessing a doctoral degree (among all higher education institutions) (weight = 1.5);
- I4 – the number of academic staff employed on permanent contracts possessing a doctoral. degree (in a definite higher education institution) (weight = 1);

- I5 – the number of academic staff employed on permanent contracts (weight = 0.5);
- I6 – the age structure of academic staff (the 30- to 50-year-old age group) (weight = 1);
- I7 – the number of foreign students (weight = 0.5);
- I8 – the number of publications per member of academic staff (weight = 2);
- I9 – the quality of education (excellent and good) (weight = 2);
- I10 – the popularity/ recognition of the higher education institution (weight = 1).

The resulting data of the rating of higher education institutions are shown in Table I. In the further research, the numeric values of these indicators have been used. Geographical, social and political aspects, as well as the obtained rank in the rating table have not been taken into consideration.

TABLE I

THE RATING DATA OF LATVIAN PUBLIC HIGHER EDUCATION INSTITUTIONS FOR -2012

Institution	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Rank
LU	57	83	100	66	89	46	76	75	91	99	1
RSU	66	88	30	55	100	42	100	95	100	100	2
RTU	61	67	96	68	90	36	59	54	97	99	3
REA	17	100	2	85	20	69	35	100	42	92	4
DU	48	76	22	63	99	52	3	68	43	96	5
LLU	44	67	38	70	60	38	2	20	73	99	6
BA	73	92	3	36	70	34	5	0	60	93	7
LJA	27	29	6	100	56	19	0	0	65	96	8
LMāA	8	70	3	14	100	56	0	0	75	97	9
RPIVA	72	87	9	46	86	42	2	0	31	89	10
RA	72	62	7	40	80	60	2	12	23	87	11
LMūA	6	79	3	14	97	45	2	0	76	94	12
VeA	19	67	4	37	86	41	1	20	44	90	13
LSPA	26	47	7	51	94	33	1	5	49	95	14
LiepU	40	77	9	54	43	38	0	8	34	93	15
LKuA	10	76	3	23	82	51	5	0	54	93	16
ViA	35	58	3	27	86	61	0	0	39	85	17
LNAA	1	17	2	25	100	75	7	0	41	88	18

III. CLUSTER ANALYSIS METHOD

Clustering algorithms are used to group some given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups [2] – [4]. Taking into account the important role of clustering in the data analysis, the ownership concept of the object is generalized in such a class function that determines affiliation of class objects to a concrete class.

All clustering algorithms have common parameters, the choice of which characterizes the effectiveness of clustering. The most important parameters characterizing clustering are the following: metrics (the distance between cluster elements and cluster centre), number of clusters k and cluster validity criteria.

In the data analysis, the k-means clustering algorithm is traditionally used [1]. It minimizes the quality index, which is set as a distance of all points belonging to cluster area to the centre of cluster (metrics). Metrics in this context is understood as the distance between the points included in the cluster [5], [6]. Usually, the vector of input data in clustering algorithms is compared to another or previously defined centre of cluster. Metrics of distance also shows affiliation to one or the other cluster, thus setting regularities in multidimensional data selections – by attributing the input data to one or the other class a.k.a. cluster.

One of the most widely used k-means clustering algorithm uses the Euclidean distance to measure the similarities between objects. K-means clustering algorithms need to assume that the number of groups (clusters) is known a priori. Table II outlines the k-means clustering algorithm [1].

TABLE II

AN OUTLINE OF K-MEANS ALGORITHM

K-means clustering procedure
1. Decide on a value for k.
2. Initialize the k cluster centres (randomly, if necessary).
3. Decide the class memberships of the N objects by assigning them to the nearest cluster centre.
4. Re-estimate the k cluster centres, by assuming the memberships found above are correct.
5. If none of the N objects changed membership in the last iteration, exit. Otherwise go to 3.

As a result of performance of the algorithm, final cluster centres are determined, considering the condition that the sum of distance squares among all points that belong to group j and the cluster centre should be minimal.

Important issue in the implementation of k-means algorithm is the determination of cluster number and initial centres. In

simpler tasks, it is assumed that the number of clusters is known a priori and that the first m values of dataset should be taken as the initial values of m cluster centres.

Advantages of the k-means algorithm could be considered popularity, good level of efficiency and simplicity of the procedure. However in case of heterogeneous disposition of objects, the algorithm could not provide good results. Then the parameters (number of clusters) should be changed and the operations of the algorithm should be repeated. In conclusion, the deficiency of this algorithm is that it is non-universal.

IV. Clustering Validity Measure

Cluster validity is a method to find a set of clusters that best fits natural partitions (number of clusters) without any class information.

There are three fundamental criteria to investigate the cluster validity: external criteria, internal criteria, and relative criteria [3]. In this case, only external cluster validity index has been analyzed.

Given a data set X and a clustering structure C derived from the application of a certain clustering algorithm on X, external criteria compare the obtained clustering structure C to a pre-specified structure, which reflects a priori information on the clustering structure of X. For example, an external criterion can be used to examine the match between the cluster labels with the category labels based on a priori information.

Based on the external criteria, there is the following approach: comparing the resulting clustering structure C to the independent partition of the data P, which was built according to intuition about the clustering structure of the dataset [4].

If P is the pre-specified partition of dataset X with N data points and is independent of the clustering structure C resulting from a clustering algorithm, then the evaluation of C by external criteria is achieved by comparing C to P. Considering a pair of data points x_i and x_j of X, there are four different cases based on how x_i and x_j are placed in C and P.

- Case 1: x_i and x_j belong to the same clusters of C and the same category of P.
- Case 2: x_i and x_j belong to the same clusters of C but different categories of P.
- Case 3: x_i and x_j belong to different clusters of C but the same category of P.
- Case 4: x_i and x_j belong to different clusters of C and a different category of P.

Correspondingly, the numbers of pairs of points for the four cases are denoted as a, b, c and d. As the total number of pairs of points is N(N-1)/2, denoted as M, we have

$$M = a + b + c + d = \frac{n(n-1)}{2}, \qquad (1)$$

where n is the number of data points in the dataset. When C and P are defined, one can choose one of the many clustering quality criteria [4]. In the given research, the clustering quality criteria have been evaluated with the help of Rand index.

Rand index is calculated by using the following formula:

$$R = \frac{a + d}{M} \qquad (2)$$

Rand index suggests an objective criterion for comparing two arbitrary clusterings based on how pairs of data points are clustered. Given two clusterings, for any two data points there are two cases:

- The first case is that the two points are placed together in a cluster in each of two clusterings or they are assigned to different clusters in both clusterings.
- The second case is that the two points are placed together in a cluster in one clustering and they are assigned to different clusters in the other.

The value of Rand index ranges between 0 and 1. A higher index value indicates greater similarity between C and P.

V. Clustering Results

Experimental part of the research has been carried out in MatLab [7], and the obtained clusters have been compared to SPSS clustering results [8]. Sequentially choosing the number of clusters between 2 and 10 and by applying the k-means clustering algorithm, the corresponding clusters and their components have been obtained (see Table III).

TABLE III

The Obtained Clusters and Their Components

No.	Cluster content								
2	LU RTU RSU	Others							
3	LU RTU RSU	REA	Others						
4	LU RTU	RSU	REA	Others					
5	LU RTU	RSU	REA	LNAA	Others				
6	LU RTU	RSU	REA	LNAA	LJA	Others			
7	LU RTU	RSU	REA	LNAA	LJA	LMāA LMūA LKuA	Others		
8	LU RTU	RSU	REA	LNAA	LJA	LMāA LMūA LKuA	BA RPIVA RA ViA	DU LLU VeA LiepU LSPA	
9	LU RTU	RSU	REA	LNAA	LJA	LMāA LMūA LKuA	RPIVA RA ViA	DU LLU VeA LiepU LSPA	BA

The table shows that the higher education institutions present in the first three clusters are in the top of the rating table. Similarly, it can be concluded that with respect to clusters 6, 7, 8 and 9 as a result of applying the algorithm the content of the five calculated clusters is constant. Differences occur starting from the sixth cluster.

Dendrograms are often used for the purposes of visualizing clusters. If two clusters fall into one group at k-level and do not change at higher levels, such grouping is called a hierarchical clustering [1]. Each hierarchical grouping has a corresponding tree structure called a dendrogram that shows how clusters are grouped. Fig. 1 shows the dendrogram of

rating data of higher education institutions obtained as a result of hierarchical clustering.

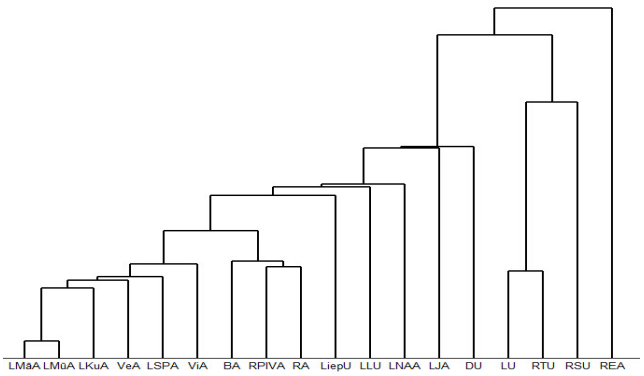

Fig. 1. Dendrogram of the rating data of higher education institutions

The analysis of the dendrogram indicates that the resulting clusters do not significantly differ from the clusters obtained by applying the k-means clustering algorithm.

In order to verify clustering validity, the quality index has been calculated – Rand index for ten clusters. Cluster structure C (consecutively with the number of clusters between 2 and 10 clusters) has been compared with specified divisions P containing various possible clusters.

Further, the total error has been calculated. The following errors of overall clustering have been calculated: 2 clusters – 5.56 %, 3 clusters – 77.8%, 4 clusters – 72.2%, 5 clusters – 88.9%, 6 clusters – 72.2%, 7 clusters – 94.4%, 8 clusters – 44.4%, 9 clusters – 72.2%, 10 clusters – 66.7%.

Among all structures, the lowest mistake occurs with 8 clusters, namely, the 8-cluster structure in this case is the most optimal. Fig. 2 shows the calculated Rand index for the 8-cluster structure.

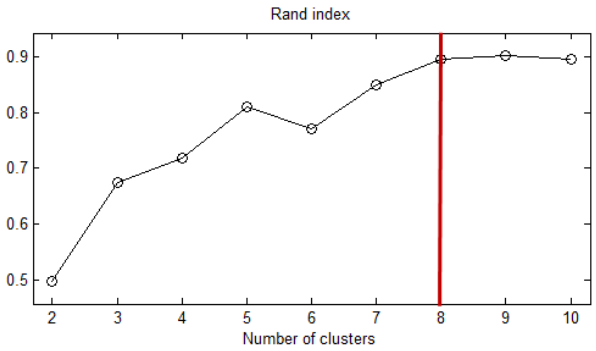

Fig. 2. Rand index in case of 8 clusters

Thus, it has been identified that the selection of the given data is best characterized by the 8-cluster structure. Taking into consideration the speech made in the public space regarding the necessity for restructuring the higher education institutions, from the mathematical point of view the calculated 8 optimal clusters could be further combined upon obtaining a "super cluster" with LU, RTU, RSU and REA. The resulting division into clusters is shown in Fig. 3.

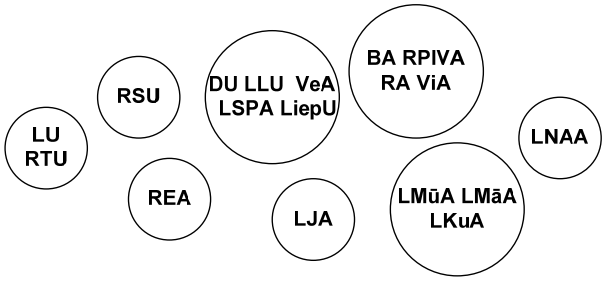

Fig. 3. Division of higher education institutions in case of 8 clusters

The results obtained in the research show that the higher education institutions are divided according to the measure of their "closeness" that is defined by index values.

Indexes characterizing the quality of clustering are useful for analyzing the performance of clustering algorithms. With their help, it is possible to choose an optimal cluster structure in cases when data distribution into clusters has not initially been set.

VI. Conclusions

In recent years, leaders of the rating have not changed – the first six positions are occupied by the following higher education institutions: LU, RSU, RTU, REA, DU, LLU (see Fig. 4).

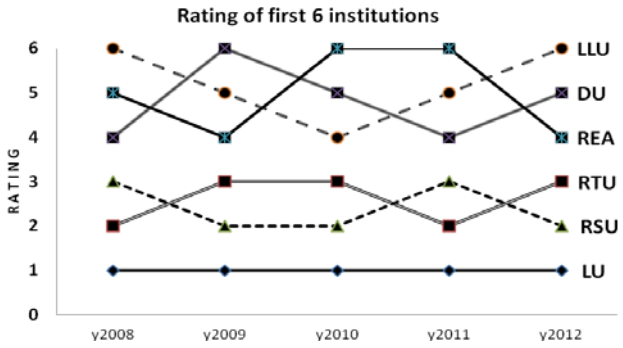

Fig. 4. Top 6 higher education institutions

Based on the figure, LU holds steady the first place, RSU and RTU share the 2nd and 3rd place, but the 4th – 6th positions are occupied by REA, DU and LLU with varying degrees of success.

Certainly, for all higher education institutions the following issue is topical – what changes of the indicator values affect the overall rating. The analysis of the first three winners in the rating of higher education institutions for the year 2012 allows making the following assumptions:

- replacing weight values of all indicators to 1, the order is as follows: RSU, LU, RTU;
- changing indicator I8 weight value to 1 – the order of places does not change;
- changing indicator I7 weight value to 1 – the order of places is as follows: RSU, LU, RTU;
- changing indicator I2 weight value to 1 – the order of places does not change;
- without I9 and I10 the order of places does not change.

According the existing indicator weight values, in fact, RSU loses most of all.

To qualify for the leader's position in the rating of higher education institutions, it can be concluded from Table III that RTU should increase the number of the graduates (I2), as well as the number of foreign students (I7) and, especially, the number of publications (I8).

Analyzing the reviews expressed in the press with respect to the correctness of the rating of higher education institutions, it has been concluded that the main objections are as follows:

- it is not correct to compare the number of foreign students in private and public higher education institutions;
- it is not correct to state: "the larger university, the higher quality";
- more rating points in the rating table are gained by the institutions with a large number of students per instructor.

Taking into consideration that the current rating calculation methodology is criticized by the representatives of the institutions, it would be useful to develop a methodology satisfying the needs of the majority of institutions.

REFERENCES

[1] B.S. Everitt, *Cluster analysis*. Edward Arnold, London, 1993.
[2] L. Kaufman and P.J. Rousseeuw, Finding groups in data. An introduction to cluster analysis. John Wiley & Sons, 2005.
[3] R. Xu and D.C. Wunch, *Clustering*. John Wiley & Sons, 2009, pp. 263-278.
[4] G. Gan., C. Ma and J. Wu, *Data clustering: Theory, algorithms and applications*. ASA-SIAM series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.
[5] P. Grabusts, *Distance Metrics Selection Validity in Cluster Analysis.* Scientific Journal of Riga Technical University: Computer Science. RTU, Riga, Issue5, Volume 49, P. 72-77, 2011. Available: https://ortus.rtu.lv/science/lv/publications/12526
[6] M. Li, X. Chen, B. Ma and P. Vitanyi, *The similarity metric*. IEEE Transactions on Information Theory, vol.50, No. 12, pp.3250-3264, 2004.
[7] MathWorks homepage. [Online]. Available: www.mathworks.com [Accessed: Sept. 15, 2012].
[8] SPSS homepage. [Online]. Available: www.spss.com [Accessed: Sept. 15, 2012].
[9] Rating overview (in Latvian): 2008 year. [Online]. Available: http://la.lv/index.php?option=com_content&view=article&id=172572&catid=178:arhvs&Itemid=272 [Accessed: Sept. 15, 2012].
[10] Rating overview (in Latvian): 2009 year. [Online]. Available: http://alephfiles.rtu.lv/TUA01/000031826_e.pdf [Accessed: Sept. 15, 2012].
[11] Rating data (in Latvian): 2010 year. [Online]. Available: http://www.lu.lv/fileadmin/user_upload/lu_portal/zinas/kopsavilkums.xls [Accsessed: sept. 15, 2012].
[12] Rating data (in Latvian): 2011 year. [Online]. Available: http://la.lv/images/stories/2011/05/Reitingi.pdf [Accessed: Sept. 15, 2012].
[13] Rating data (in Latvian): 2012 year. [Online]. Available: http://la.lv/images/stories/2012/04/11/tabulas_augstskolas.pdf [Accessed: Sept. 15, 2012].
[14] Ranking Web of World Universities. [Online]. Available: http://www.webometrics.info/en/Europe/Latvia [Accessed: Sept. 15, 2012].
[15] SCImago Institutions Rankings. [Online]. Available: www.scimagoir.com [Accessed: Sept. 15, 2012].
[16] QS World University Rankings. [Online]. Available: http://www.topuniversities.com/university-rankings/world-university-rankings/2011 [Accessed: Sept. 15, 2012].
[17] The Times Higher World University Ranking. [Online]. Available: http://www.timeshighereducation.co.uk/world-university-rankings/2011-2012/top-400.html [Accessed: Sept. 15, 2012].

Peter Grabusts was born in Rezekne, Latvia. He received his Dr.sc.ing. degree in Information Technology from Riga Technical University in 2006. Since 1996 he has been working at Rezekne Higher Education Institution. Since 2008 he is an Associate Professor at the Department of Computer Science.
His research interests include data mining technologies, neural networks and clustering methods. His current research focuses on techniques for clustering and fuzzy clustering.
E-mail: peter@ru.lv

Pēteris Grabusts. Latvijas augstskolu reitinga analīze ar klasterizācijas palīdzību
Latvijas augstskolu reitinga dati tika publicēti jau piekto reizi. Reitinga izveides pamatā izvēlēta metodoloģija, kurā izmanto 10 vērtēšanas kritērijus jeb indikatorus. Pētījumā tika veikts mēģinājums sagrupēt augstskolas ar klasterizācijas algoritma k-means palīdzību un pārliecināties, vai šāds sadalījums atbilst matemātiski izskaitļotajai augstskolas vietai reitingu tabulā. Pētījuma mērķis bija raksturot klasteru skaita izmaiņas un novērtēt klasterizācijas rezultātu ticamību. Par sākotnējiem datiem tika izmantota Latvijas valsts dibināto augstskolu reitinga tabula par 2012. gadu, un eksperimenta uzdevums bija parādīt, kā ar klasterizācijas metodēm alternatīvā veidā var analizēt šādus datus. Klasterizācijas pareizība tika novērtēta ar Rand indeksa palīdzību. Pētījumā tika izmantotas reitinga indikatoru skaitliskās vērtības, netika ņemti vērā ģeogrāfiskie, sociālie un politiskie aspekti, kā arī iegūtā vieta reitingu tabulā. Secīgi izvēloties klasteru skaitu robežās no 2 līdz 10 un pielietojot klasterizācijas algoritmu k-means, tika iegūti attiecīgie klasteri un tajos ietilpstošās augstskolas. Lai pārbaudītu veiktās klasterizācijas ticamību, tika izskaitļots kvalitātes rādītājs - Rand indekss desmit klasteriem. Reitingu datu klasterizācijā un pēc tās veiktās klasterizācijas indeksa izskaitļošanas par optimālāko tika izvēlēta klasteru struktūra ar astoņiem klasteriem. Pētījuma rezultāti liecina, ka augstskolas klasteros iedalītas pēc to „tuvības" mēra, ko nosaka indikatoru vērtības. Tāpat tika secināts, ka vietu reitinga tabulā būtiski ietekmē indikatora I8 (publikāciju skaits) vērtība. Tāda veida datu analīzi ar klasterizācijas palīdzību var uzskatīt par papildu līdzekli tradicionālajām datu apstrādes procedūrām, un tās rezultāti ir rūpīgi jāanalizē.

Петерис Грабуст. Анализ рейтинга высших школ Латвии с помощью кластеризации
Рейтинговые данные высших школ Латвии опубликованы пятый раз подряд. За основу рейтинга взята методология, использующая 10 критериев или индикаторов оценки. В исследовании произведена попытка сгруппировать высшие школы с помощью алгоритма кластеризации k-means и убедиться в соответствии такого распределения математически вычисленному месту высших школ в рейтинговой таблице. Целью исследования являлись характеристики изменения количества кластеров и оценка качества кластеризации. В качестве исходных данных использовались только численные значения индикаторов рейтинговой таблицы за 2012 год, не учитывались географические, социальные и другие аспекты, а также место высшей школы в таблице. Последовательно выбирая количество кластеров в пределах от 2 до 10 и применяя алгоритм кластеризации k-means, были получены соответствующие кластеры с входящими в них высшими школами. Для проверки достоверности результатов кластеризации был вычислен показатель качества – индекс Рэнда. После кластеризации рейтинговых данных и вычисления индекса кластеризации оптимальной была признана структура из 8 кластеров. Результаты исследования показали, что высшие школы в кластерах распределены соответственно мере „близости" значений индикаторов. Такой анализ рейтингов с помощью кластеризации может использоваться как дополнительное средство к традиционным методам обработки данных.