

University Ranking Prediction System by Analyzing Influential Global Performance Indicators

Anika Tabassum*, Mahamudul Hasan[†], Shibbir Ahmed*, Rahnuma Tasmin^ℓ, Deen Md. Abdullah[‡],
Tasnim Musharrat[‡]

Department of Computer Science and Engineering

*Bangladesh University of Engineering and Technology, [†]University of Dhaka, [‡]Primeasia University,

^ℓMilitary Institute of Science and Technology, Dhaka, Bangladesh

Email: {annisizzler, munna09bd, shibbirahmedtanvin, rabi.mist, tasnim765}@gmail.com

Abstract—In this research, we present a technique of developing university ranking prediction system by analyzing global university performance indicators. Here, we consider standardized dataset of Times higher education world university rankings. Firstly, we perform country wise university ranking data analysis to observe the variation of performance indicators to find out the top influential features. To build the proposed prediction model, we split the ranking dataset into training and test data. Then, based on score of previous years we generate predicted score for each influential feature using our proposed outlier detection and rank score calculation algorithm. Later on, all the universities are ranked globally based on the predicted total score. Then, we evaluate the prediction system accuracy based on ROC curve, recall, number of matched rank against rank deviation. Finally it is justified that our proposed university ranking prediction system is well suited to assess the upcoming global university ranking.

Index Terms—Prediction System; Global University Ranking; Data Analysis; Machine Learning; Computational Intelligence.

I. INTRODUCTION

Ranking universities is a fundamental issue not only to the students and academics but also to university authorities, industry and even government. Global university ranking depends on several key factors such as teaching, research, citations, international outlook, industry income etc. Usually global university ranking is published on a yearly basis. Before publishing the present year's ranking, it is essential to assess the upcoming rank of certain university for several reasons. Prospective students require this for applying to specific universities based on upcoming ranks. University authorities should assess the varsity's upcoming rank to improve certain fields and upgrade their standards comparing with others. Industry and even government should try to forecast the ranking of universities for providing grants to the most eligible institutions for upcoming year. Concepts and techniques of data analysis and machine learning [1], [2] are very much useful to explain the past and predict the future by analyzing [3] and exploring the data.

There are several kinds of national and international university ranking systems in which different methodologies are adopted. The Times Higher Education World University Ranking, founded in the United Kingdom in 2010, is regarded

as one of the most influential and widely observed university ranking system [4]. There are other ranking systems such as Academic Ranking of World Universities, also known as the Shanghai Ranking, founded at China in 2003, the Center for World University Rankings (CWUR), comes from Saudi Arabia since 2012. A web based platform for predictive modeling and data analytics provides the dataset [5] containing all these three global university rankings of past few years in order to explore and investigate the best universities in the world. The Times Higher Education World University Ranking system provides with the global performance tables that justify research-intensive universities across all their prime objectives which are teaching, research, knowledge transfer and international outlook. The ranking system uses in total 13 calibrated performance indicators to provide the most comprehensive and balanced comparisons trusted by students, teachers, university authorities, industry and government as well.

In this paper, we develop a global university ranking prediction system by analyzing all the university performance indicators of Times Higher Education World University Rankings dataset. At first, we have made country wise university ranking data analysis in order to distinguish the actual effect of performance indicators and identify the top most influential factors. Then, based on previous years score we have generated predicted score for those influential attributes using our proposed outlier detection algorithm for those performance indicators and rank score calculation algorithm. After that, based on predicted total score we have ranked all the universities globally. We have split the ranking data into training and test data for the purpose of evaluating our proposed prediction system as well. We have evaluated the prediction system accuracy based on ROC curve, recall, number of matched rank against rank deviation. Thus we have found that our proposed university ranking prediction system is well suited for estimating upcoming global university ranking.

II. MOTIVATION

University ranking is very significant to students, teachers, university board trustees and government. Different ranking systems publish varsity ranking annually or biannually. People

can observe the current rank of a university from those ranking providers. But ranking is a transitional and ever-changing process. If upcoming rank of a certain university can be predicted, it would certainly assist the prospective graduate students to choose desired institution in advance. Moreover the university authorities can observe their present situation and try to improve each of the influential performance indicators before publishing their upcoming rank. Prediction of university ranking will be very much helpful to the donors as well as government in order to make the decision of continuation or approval of grant for a specific university. So, the prior assessment of performance indicators such as teaching, research, citations, international outlook is not only necessary to predict a university rank but also to provide future insight of admission, provision of grants, university development indicators. That is why we are motivated to make analysis of the significant performance indicators of a well-reputed university ranking system and build a university ranking prediction system which can predict the rank of global universities precisely and effectively.

III. RELATED WORK

Ranking is a very important research topic in Information Retrieval. Feature selection is a fundamental issue in ranking model. An optimization method for feature selection in ranking has been briefly discussed in an influential research study [6]. Learning to rank is useful for document retrieval, collaborative filtering, and many other applications. The theory and algorithms of list-wise approach of learning to rank has been investigated in a paper [7]. It has been pointed out that to understand the effectiveness of a learning to rank algorithm, it is necessary to conduct theoretical analysis on its loss function. Another research study [8] is concerned with learning to rank which is to construct a model or a function for ranking objects by a newly proposed list-wise approach from Pairwise approach. There is also a significant work [9] which presents a new energy-based, list-wise approach to learning to rank. Moreover, there is another study [10] representing a new ranking scheme called collaborative ranking with three specific forms demonstrating its effectiveness on entity linking task. Meanwhile, a recent study [11] proposes CCRank, a parallel learning to rank framework based on cooperative co-evolution, aiming to significantly improve the learning efficiency while maintain accuracy.

There are very few research studies regarding the global ranking of higher education institutions. Among them, a recent study [12] closely examines the methodology and main features of each of all the existing global university ranking systems. There is also a detailed report [13] on Global university rankings and their impact stating the methodologies and potential impact of the most popular international and global rankings already in existence. University rankings theoretical basis, methodology and impacts on global higher education have been briefly illustrated in an impressive

research study [14]. Previously, another study [15] examined whether a university is to be ranked or not by analyzing the impact of global rankings in higher education. Again, there is an article [16] stating quantitative methodology of comparative analysis of global university rankings for the Mediterranean and Black Sea region. Moreover, there is a significant study [17] conducting a systematic comparison of national and global university ranking systems in terms of their indicators, coverage and ranking results. Another well-cited paper [18] illustrates a comparative analysis of the four rankings taking into account both the indicators frequency and its weights.

Recently there is increasing interest in university rankings as well as the comparison among the ranking systems. A recent study [19] has compared different world university rankings using a set of similarity measures. The comparisons in a research paper [20] revealed that ranking results can vary, sometimes dramatically, due to methodologies and emphases of various criteria. Considering all these comparative analysis, a unifying framework for ranking predictions has been proposed from training data called Boosted Ranking Model [21]. There is another study [22] showing the possibility to predict usability of a university website from university ranking systems. All these research studies encourage us to analyze the influential performance indicators of a well-reputed university ranking system and develop a university ranking prediction system which can predict the rank of global universities accurately.

IV. METHODOLOGIES

A. Data Analysis

For developing university ranking prediction system, here we exploit publicly available dataset of global university rankings [5]. At first, we analyze the dataset of global university rankings to find out the influential performance indicators shown in the table I. There are several attributes in the dataset which are university name, country of the university, score of teaching, research, citations, income, international, total score, number of students, ratio of student and staffs, number of international students, ratio of female and male, year of rank etc. described in the Table I.

B. Features Selection

The dataset of global university rankings consists of 13 features known as performance indicators of a university. We have analyzed country wise influence of all the performance indicators in last two years and found that the variation of scores in teaching, research, citations, international outlook mostly influence the ranking of universities as depicted in Fig. 1 and Fig. 2. So, we have developed effective and accurate university ranking prediction system by utilizing these features of the dataset. For example, if we look at Caltech University data, we see, the other indicators e.g., student staff ratio, number of students, international student acceptance rate and female to male ratio remain constant irrespective of ranking.

TABLE I
ANALYSIS OF DATASET FOR UNIVERSITY RANK PREDICTION

<i>university_name</i>	Different universities all around the world
<i>country</i>	Different Countries around the world
<i>teaching</i>	Score indicating the learning environment
<i>international</i>	Score combining staff, students and research outlook
<i>research</i>	Score based on volume, income and reputation in research
<i>citations</i>	Score representing research influence
<i>income</i>	Score indicating industry income by knowledge transfer
<i>total_score</i>	total score combining weighted scores of other performance indicators
<i>num_students</i>	number of full time equivalent students at the University
<i>student_staff_ratio</i>	ratio of full time equivalent students to the number of academic staff those involved in teaching or research
<i>international_students</i>	percentage of students originating from outside the country of the University
<i>female_male_ratio</i>	ratio of female to male students at the University
<i>year</i>	Years from 2011-2016

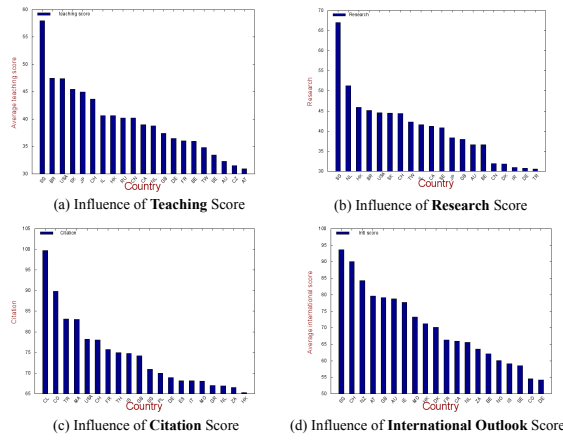


Fig. 1. Charts representing Influence of (a)Teaching, (b)Research, (c) Citation, and (d) International outlook Score of different country-wise universities in the year 2015

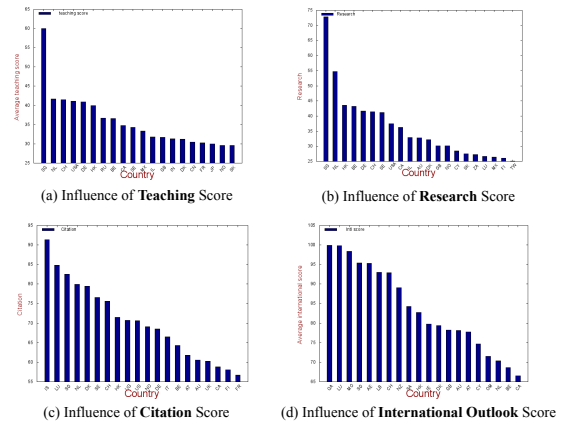


Fig. 2. Charts representing Influence of (a)Teaching, (b)Research, (c) Citation, and (d) International outlook Score of different country-wise universities in the year 2016

Therefore, we eliminate these features from the performance indicators at first. For the rest five features(3,4,5,6,7 of Table 1) we consider different weights to calculate total score and our predicted total score matches relatively more with the Times total score when we consider negligible weight for income.

C. Outlier Detection

We have found that a feature of a particular university does not follow a particular trend throughout the years. For example, analyzing the data of CalTech teaching score, we see that the points are not following any similar trend depicted in Fig. 3. Linear regression will not work for predicting the score of the next year. As, the regression line is going down, for the year 2016 it will also predict a downward trend score and so for the next year whereas actually the score for 2016 has gone up. Thus, a regression line will under fit the prediction for the next year.

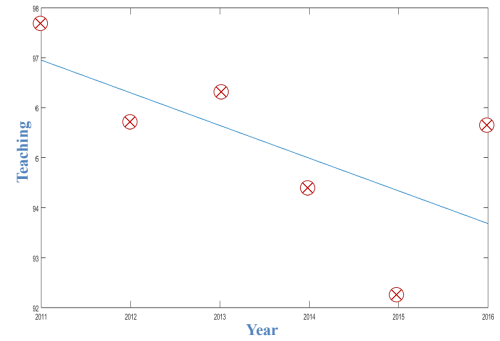


Fig. 3. Graph representing Teaching Score of CalTech university during year 2011 to 2016

That is why *performance_Indicators_OutlierDetection()* algorithm has been proposed to detect the outliers for a particular performance indicator in different years which is

illustrated in the Algorithm 1. In the Algorithm 1, line 1 indicates the total number of performance indicators, which is 4 in this case. Again, line 3 and 4 calculate the mean and standard deviation of each performance indicators. Finally, a f-score is calculated in line 6 and outlier-limit is depicted in line 10 based on t-distribution. So, if the f-score is greater than the outlier limit, then it is considered as the outlier and preserved in a set S in line 13.

Algorithm 1: performance_Indicators_OutlierDetection()

Input: *teaching, research, citation, international, scores* of all universities
Output: Set *S* of Outliers for each performance indicators

```

1   $N \leftarrow$  total number of Performance_Indicators
2  For each Performance_Indicator  $P_i \in P$ 
3      calculate average  $\bar{P}_i$  of  $P_i$ 
4      calculate standard deviation  $S_i$  of  $P_i$ 
5       $temp \leftarrow f_{abs}(P_i - \bar{P}_i)$ 
6       $f - score \leftarrow temp / S_i$ 
7       $P \leftarrow \frac{t_{\alpha}}{(2N)(N-2)}$ , where  $\frac{t_{\alpha}}{(2N)(N-2)}$ , taken from
   t - distribution
8       $temp_1 \leftarrow P^2$ 
9       $temp_2 \leftarrow \frac{temp_1}{(N-2+temp_1)}$ 
10     outlier - limitp  $\leftarrow \frac{N-1}{\sqrt{N}} \sqrt{temp_2}$ 
11      $S \leftarrow \phi$ 
12     If ( $f - score \geq outlier - limit_p$ )
13          $S \leftarrow S \cup P_i$ 
14     End If
15     return set S
16 End For
```

D. Rank Score Calculation

For building the prediction model, we split the ranking dataset into training and test data. After detecting outliers by using our proposed Algorithm 1, we have deployed a rank score calculation algorithm for each of the performance indicators i.e., teaching, research, citation, international. We have used prior scores of each of the performance indicators according to the weight of most recent year and compute new score for prediction. Then, total score has been calculated using the given weight to each of the features based on the influence of performance indicators on the university ranking [4]. The proposed *rank_score_calculate()* algorithm is depicted in Algorithm 2 where line 3 computes score for the selected performance indicators using empirical statistics weighted average method. Line 6 selects a weight w_i using moving average analysis for the year of the university rank we want to predict and computes the score for the current performance indicator.

E. University Rank Prediction System

At first, we have considered the university ranking dataset of Times Higher Education [5]. Then we have selected the

Algorithm 2: rank_score_calculate()

Input: *teaching, research, citation, international, scores* of all universities for *i* different years excluding outliers
Output: predicted *rank - score* for each *performance_indicator*

```

1  For each university U
2      score  $\leftarrow 0$ 
3      For each Performance_indicator  $j$  of  $P_i$ 
4          statistically select the value of  $\alpha_j$ 
5          For each (current year  $i$ )
6               $score_j \leftarrow \alpha_j(weight_i \times x_i) + score$ 
7          end For
8          rank-scoreu  $\leftarrow score_j$ 
9      end For
10     return rank-scoreu
11 end For
```

most influential performance indicators by analyzing year wise variation of scores in various universities. To build as well as to evaluate the prediction model, we split the dataset as training data for year 2011 to 2015 and left the data of year 2016 for test purpose. Deploying the training dataset, we have detect outliers for each performance indicators using our proposed Algorithm 1. Then, we have calculated the predicted score of teaching, research, citations, international outlook using the proposed Algorithm 2. After that, we have generated total predicted rank score based on certain weight of each performance indicators. Finally we have ranked universities globally using the predicted total rank score. The entire process is illustrated in the flow diagram of Fig. 4.

V. EXPERIMENTATION

A. Experimental Setup

As stated earlier, we have analyzed the dataset of world university rankings of Times Higher Education using MATLAB [23]. To build the prediction model we have deployed the Algorithm 1 and Algorithm 2 and run the code in java language for efficient computation. We have used Gnuplot for representing graphical evaluation of our proposed prediction model. The technology and tools used for the university rank prediction system is given below:

- Language: Java
- Environment: Windows
- IDE: NetBeans IDE 8.0.2
- Dataset: World University Rankings of Times Higher Education
- Data Analysis Tool: MATLAB
- Graphing Utility: Gnuplot

B. Experimental Evaluation

We have divided the data set into test train split to evaluate our proposed university ranking prediction system. As, we have found no existing university ranking prediction system to

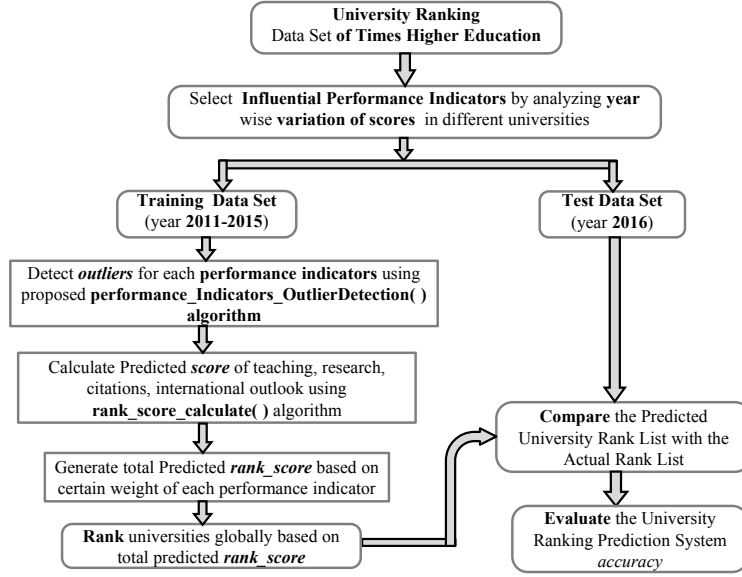


Fig. 4. Flow Diagram of University Rank Prediction System

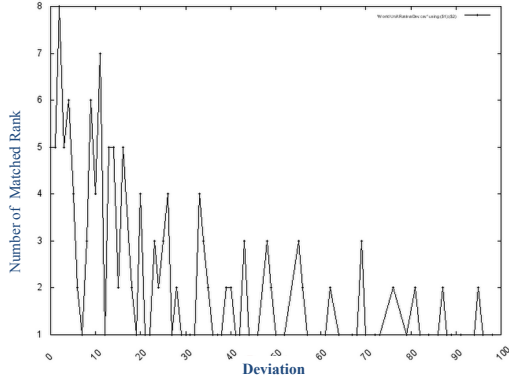


Fig. 5. Number of Matched Rank against Deviation graph

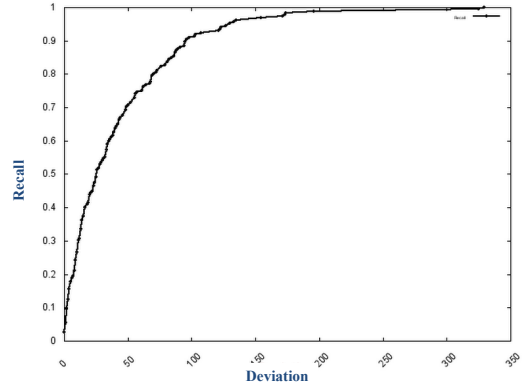


Fig. 6. Recall against Deviation graph

be compared with our proposed system, we have used below three different evaluation measures to justify the accuracy and explain the performance of the prediction model.

1) *Number of Matched Rank Vs Deviation*: The total number of matched predicted ranks (N_m) of universities are plotted against the predicted rank of a university which is within deviation d with the rank given in Times data ($Rank_{p_d}$) using the equation 1.

$$N_m = \sum Rank_{p_d} \quad (1)$$

It is observed from Fig.5 that, the number of predicted ranks of universities which is within deviation i with the actual rank of Times data decreases with the increase of deviation. We also observe that highest number of universities found within 2 deviation. So, it justifies the accuracy of the prediction system.

2) *Recall Vs Deviation*: The Recall has been calculated from dividing the number of rank within deviation(d) by total

universities(T) using the equation 2 where predicted rank of a university is denoted by $Rank_p$ and rank given in Times data is depicted as $Rank_m$.

$$Recall = \sum_{i=0}^d \frac{(Rank_p - Rank_m)}{T} \quad (2)$$

It is illustrated in the Fig.6 that the recall increases exponentially with the increase of deviation. So, it provides a justification for the accuracy of the prediction system.

3) *Accuracy Vs Deviation*: Here, the term Accuracy(Acc) is defined as the difference between predicted rank($Rank_p$) and the rank given in Times data($Rank_m$) of that university within deviation (d) divided by total number of the universities(T) which is given in the equation 3.

$$Acc = \frac{\sum_{i \in d} (Rank_p - Rank_m)}{T} \quad (3)$$

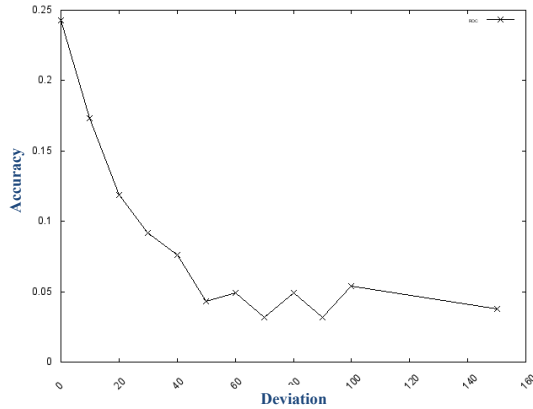


Fig. 7. Accuracy against Deviation (ROC curve) graph

For illustrating the ROC curve, we have plotted *Accuracy* (obtained from equation 3) against the *Deviation* and found that accuracy is higher in case of lower deviations as depicted in Fig. 7. From this figure, we can speculate that accuracy is higher in case of lower deviation, i.e., our algorithm can predict correctly ranks of the universities similar to Times data rank and at 0 deviation of the accuracy of our predicted university rank is the highest, and very small number of universities have been found whose predicted rank deviates highly from the actual rank. It clearly proves that the proposed prediction system is much more precise and effective.

VI. CONCLUSIONS

In this research, we present a technique of developing a global university ranking prediction system by analyzing all the university performance indicators. Here, we have considered the dataset of Times Higher Education World University Rankings which comprises of global performance tables that judge universities based on several calibrated performance indicators. We have split the ranking dataset into training and test data for the purpose of building and evaluating our proposed prediction system. At first, we have made country wise university ranking data analysis to observe the variation of performance indicators and figure out the most influential factors. Then, based on previous years score we have generated predicted score for those influential attributes using our proposed outlier detection algorithm and rank score calculation algorithm to predict score of specific features for upcoming year. After that, based on predicted total score we have ranked all the universities globally. Finally, we have evaluated the prediction system accuracy based on recall, number of matched rank and ROC curve with respect to the rank deviation. Thus, we have found that our proposed university rank prediction system is acceptable to estimate upcoming global university ranking.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge collaborators of Kaggle for their generous assistance to get the

necessary data of World University Ranking for this research study.

REFERENCES

- [1] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [2] C. C. Aggarwal, *Data mining: the textbook*. Springer, 2015.
- [3] D. J. Power, R. Sharda, and F. Burstein, *Decision support systems*. Wiley Online Library, 2015.
- [4] "Times world university rankings," <https://www.timeshighereducation.com/world-university-rankings>, accessed: 2016-10-30.
- [5] "Dataset of world university rankings," <https://www.kaggle.com/mylesoneill/world-university-rankings>, accessed: 2016-10-10.
- [6] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 407–414.
- [7] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: theory and algorithm," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1192–1199.
- [8] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 129–136.
- [9] M. N. Volkovs and R. S. Zemel, "Boltzrank: learning to maximize expected ranking gain," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1089–1096.
- [10] Z. Chen and H. Ji, "Collaborative ranking: A case study on entity linking," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 771–781.
- [11] S. Wang, Y. Wu, B. J. Gao, K. Wang, H. W. Lauw, and J. Ma, "A cooperative coevolution framework for parallel learning to rank," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 12, pp. 3152–3165, 2015.
- [12] M. Rayanakorn, "Global ranking of higher education institutions," 2011.
- [13] A. Rauhvargers, *Global university rankings and their impact: Report II*. European University Association Brussels, 2013.
- [14] J. C. Shin, R. K. Toutkoushian, and U. Teichler, *University rankings: Theoretical basis, methodology and impacts on global higher education*. Springer Science & Business Media, 2011, vol. 3.
- [15] S. Marginson and M. Van der Wende, "To rank or to be ranked: The impact of global rankings in higher education," *Journal of studies in international education*, vol. 11, no. 3-4, pp. 306–329, 2007.
- [16] V. M. Moskovkin, E. V. Pupynina, N. P. Zaitseva, and R. V. Lesovik, "Methodology for comparative analysis of university rankings, with the mediterranean and black sea region countries taken as an example," *Methodology*, 2013.
- [17] M. P. Çakır, C. Acartürk, O. Alaşehir, and C. Çilingir, "A comparative analysis of global and national university ranking systems," *Scientometrics*, vol. 103, no. 3, pp. 813–848, 2015.
- [18] G. Buella-Casal, O. Gutiérrez-Martínez, M. P. Bermúdez-Sánchez, and O. Vadillo-Muñoz, "Comparative study of international academic rankings of universities," *Scientometrics*, vol. 71, no. 3, pp. 349–365, 2007.
- [19] I. F. Aguillo, J. Bar-Ilan, M. Levene, and J. L. Ortega, "Comparing university rankings," *Scientometrics*, vol. 85, no. 1, pp. 243–256, 2010.
- [20] M.-H. Huang, "A comparison of three major academic rankings for world universities: From a research evaluation perspective," *Journal of Library and Information Studies*, vol. 9, no. 1, pp. 1–25, 2011.
- [21] K. D. Rosa, V. Metsis, and V. Athitsos, "Boosted ranking models: a unifying framework for ranking predictions," *Knowledge and information systems*, vol. 30, no. 3, pp. 543–568, 2012.
- [22] L. Hasan and E. Abuelrub, "Is it possible to predict usability of a university website from university ranking systems?" in *Proceedings of the World Congress on Engineering*, vol. 2, 2013, pp. 3–5.
- [23] S. J. Chapman, *MATLAB programming for engineers*. Nelson Education, 2015.