DE GRUYTER
OPEN

Research Paper

Exploring the Potentialities of Automatic Extraction of University Webometric Information

Gianpiero Bianchi[1], Renato Bruni[2], Cinzia Daraio[2†], Antonio Laureti Palma[1], Giulio Perani[1], Francesco Scalfati[1]

[1]ISTAT, Italian National Institute of Statistics, Via Cesare Balbo 16, Rome 00184, Italy
[2]DIAG, Sapienza University of Rome, Via Ariosto 25, Rome 00185, Italy

Citation: Bianchi, G., Bruni, R., Daraio C., Laureti Palma, A., Perani, G., & Scalfati, F. "Exploring the potentialities of automatic extraction of university webometric information." *Journal of Data and Information Science*, vol. 5, no. 4, 2020, pp. 43–55. https://doi.org/10.2478/jdis-2020-0040

Received: Jul. 20, 2020
Revised: Oct. 28, 2020
Accepted: Nov. 9, 2020

Abstract

Purpose: The main objective of this work is to show the potentialities of recently developed approaches for automatic knowledge extraction directly from the universities' websites. The information automatically extracted can be potentially updated with a frequency higher than once per year, and be safe from manipulations or misinterpretations. Moreover, this approach allows us flexibility in collecting indicators about the efficiency of universities' websites and their effectiveness in disseminating key contents. These new indicators can complement traditional indicators of scientific research (e.g. number of articles and number of citations) and teaching (e.g. number of students and graduates) by introducing further dimensions to allow new insights for "profiling" the analyzed universities.

Design/methodology/approach: Webometrics relies on web mining methods and techniques to perform quantitative analyses of the web. This study implements an advanced application of the webometric approach, exploiting all the three categories of web mining: web content mining; web structure mining; web usage mining. The information to compute our indicators has been extracted from the universities' websites by using web scraping and text mining techniques. The scraped information has been stored in a NoSQL DB according to a semi-structured form to allow for retrieving information efficiently by text mining techniques. This provides increased flexibility in the design of new indicators, opening the door to new types of analyses. Some data have also been collected by means of batch interrogations of search engines (Bing, www.bing.com) or from a leading provider of Web analytics (SimilarWeb, http://www.similarweb.com). The information extracted from the Web has been combined with the University structural information taken from the European Tertiary Education Register (https://eter.joanneum.at/#/home), a database collecting information on Higher Education Institutions (HEIs) at European level. All the above was used to perform a clusterization of 79 Italian universities based on structural and digital indicators.

† Corresponding author: Cinzia Daraio (E-mail: daraio@diag.uniroma1.it).

Research Paper

Findings: The main findings of this study concern the evaluation of the potential in digitalization of universities, in particular by presenting techniques for the automatic extraction of information from the web to build indicators of quality and impact of universities' websites. These indicators can complement traditional indicators and can be used to identify groups of universities with common features using clustering techniques working with the above indicators.

Research limitations: The results reported in this study refers to Italian universities only, but the approach could be extended to other university systems abroad.

Practical implications: The approach proposed in this study and its illustration on Italian universities show the usefulness of recently introduced automatic data extraction and web scraping approaches and its practical relevance for characterizing and profiling the activities of universities on the basis of their websites. The approach could be applied to other university systems.

Originality/value: This work applies for the first time to university websites some recently introduced techniques for automatic knowledge extraction based on web scraping, optical character recognition and nontrivial text mining operations (Bruni & Bianchi, 2020).

Keywords Development of data and information services; Webometrics indicators; Higher education institutions; Automatic extraction; Machine learning; Optimization

1 Introduction

The need for indicators to describe and evaluate universities is constantly increasing. Policy makers are facing opportunities generated by the availability of new types of information, including a large number of university rankings, which often attract policy and media attention, even if they may also receive harsh methodological criticism. Daraio and Bonaccorsi (2017) discuss new trends in the user requirements of the indicators and show, based on the case of European universities, how the intelligent integration of existing data may lead to an open-linked data platform which may in turn bring to the construction of new indicators. Recent significant trends and challenges in data integration for Research and Innovation activities have been discussed in Daraio and Glanzel (2016), which highlight the need for new ways of data integration and interoperability among many heterogeneous data sources. In the recent Springer Handbook of Science and Technology Indicators by Glänzel et al. (2019), there is a full section on new indicators for research assessment. It includes indicators coming from social media (Wouters, Zahedi, & Costas, 2019), a survey of the development of data collection procedures on the web focusing on current practices, data cleansing and matching, data quality and transparency (Bar-Ilan, 2019) and a review of webometric, altmetric, and other online indicators for the impact assessment of nonstandard academic outputs (Thelwall, 2019).

Webometrics relies on web mining methods and techniques to perform quantitative analyses of the web. This approach is already extensively adopted to produce statistics with data collected from the Web (Björneborn & Ingwersen, 2004; Thelwall, 2009; Thelwall, Vaughan, & Björneborn, 2005). Similar techniques have been recently specialized for the analysis and the evaluations of higher educational Institutions, by retrieving the institution information from their websites. In practice, a webometric approach can draw a "university profile" by analyzing many different aspects concerning the university website. Some limitations are given by the source from which the information is obtained.

One common and basic technique to obtain this information is through the automatic interrogations of search engines. For example, Aguillo, Ortega, and Fernández (2008) present a webometric ranking of world universities using a combined indicator called WR that takes into account number of web pages; number of rich files (pdf, ps, doc, and ppt format); number of articles; number of external inlinks. Those data are obtained by querying several search engines. Elgohary (2008) uses a webometric approach to investigate the Web Impact Factor (WIF) of 99 universities from 20 Arab countries by using AltaVista search engine. Islam and Alam (2011) examine the websites of 44 private universities in Bangladesh and identify statistics on web pages and links in order to calculate their overall WIF and absolute WIF, again by using AltaVista search engine. Pal, Sarkar, and Bhattacharya (2019) compute the WIFs for Indian Open University websites and identify three types of them.

Another technique to obtain the above information is by searching in social networks. In particular, Aguillo and Orduna-Malea (2013) observe that new scores can be obtained from open environments, especially through Web 2.0 tools, defined as the "21st century's new scholarly communication channel", and they propose to use academia, facebook, linkedln, twitter, wikipedia, and youtube. McCoy, Nelson, and Weigle (2018) present an alternative to the university ranking lists, which consists in mining a collection of university data, obtained from Twitter and publicly available online, to compute social media metrics that approximate the typical academic rankings of 264 universities in the United States. A third possible source of information is through the use of Web analytics services provided by some agency or application. For example, Bychkova1 and Okushova (2017) provide an analysis of the Internet presence of several main universities from Russia, the US and the UK using Google PR, Google analytics, Alexa Ranking, IQBuzz, and similar internet services. Many international rankings use data obtained from this type of source. However, all the above sources can provide only predetermined types of information, basically concerning with the counting of number of accesses, links, etc.

To attain superior flexibility, we propose in this work the use of a sequence of automatic knowledge extraction techniques, in order to obtain, directly from the universities' websites, more advanced analyses, which could even be defined on demand. This approach is here integrated with the batch interrogation of search engines (in particular Bing, www.bing.com) and the use of a Web analytics provider (in particular SimilarWeb, http://www.similarweb.com). Hence, this study constitutes an integrated application of the webometric approach, exploiting all three categories of web mining: web content mining; web structure mining; web usage mining. In more detail, the knowledge extraction operations rely on web scraping and text mining techniques, using tools inspired by those introduced in a different context by Bruni and Bianchi (2020).

By exploiting the increased flexibility offered by our approach, we also contribute to the literature on the proposal of new indicators for the profiling of universities or higher education institutions using information available on the web. Note, however, that the proposed approach opens the door to many other possible indicators, hence the innovative content of this work is not limited to the specific indicators described.

Finally, we use the obtained indicators to provide a clusterization of 79 Italian universities using k-means approach. Note that, for some large universities, a considerable part of the website is split into the departmental websites, which in practice constitute autonomous websites. Hence, the analysis has to be repeated for each of them, reaching 632 scraped websites and greatly increasing the computational demand.

The article is organized as follows. Section 2 focuses on the purpose of this work. Section 3 describes the methodology and the data collection. Section 4 explains in detail the main results. Finally, Section 5 concludes the paper.

2 Purpose of the work

The main objective of this work is to show the potential of recently developed approaches for the automatic extraction of non-trivial information from university websites. Since the information is automatically extracted, it can potentially be updated more often than once a year. Moreover, when the information is obtained directly from the websites, superior flexibility in the definition of the indicators becomes possible. Furthermore, this information will be safe from mishandling, since there will be no intermediaries, and from misinterpretation, since no third person interpretation is needed. This approach allows to collect a series of indicators on the efficiency of university websites and their effectiveness in disseminating key

content. These new indicators can complement the traditional indicators of scientific research (e.g. number of articles and number of citations) and teaching (e.g. number of students and graduates) by introducing new dimensions for the "profiling" of the universities analyzed.

3 Methodology and data collection

Text Mining is a branch of Data Mining concerning the process of extracting high-quality information from texts (Aggarwal, 2018). When such a text is extracted from the web, text mining is called Web Mining. This is a prominent field of research due to the continuous expansion of the Internet and the consequent demand for always more effective information retrieval strategies. Recent web mining approaches take advantage from the integration of natural language processing with many advanced machine learning techniques, such as classification algorithms (Bruni & Bianchi, 2015) or logic-based information processing tools (Bruni et al., 2019).

Recently, Bruni and Bianchi (2020) have proposed an overall approach to websites categorization based on the use of automatic text scraping, image acquisition and processing, optical character recognition, natural language processing, and text mining to create data records representing in a standardized way the features of each website, and so they become able to use classification algorithms to perform a categorization. They apply this approach to the specific problem of the detection of websites providing e-commerce facilities. The present work is based on a similar sequence of web scraping and text mining operations to perform the quantitative analyses of the universities' websites, even if several modifications were introduced to deal with this completely different task.

In more detail, the first step consists of scraping the universities' websites. This went through three main steps: a) acquisition of the university's web addresses; b) validation of the websites; c) data extraction. Using official sources of information, the process of validation implemented a URL Retrieval technique which uses characteristics' identification to perform batch queries in the search engines. More specifically, it uses the denomination of the academic institutions as a search string. For each link found by the program, it evaluates the probability of correctness using a machine learning approach. The links whose probability exceeded a given threshold were accepted as valid ones. Even though the general rule was that of focusing on the corporate Web portal of universities, it has been needed—for a few large universities—to investigate also on university departments' websites, which in

practice constitute distinct websites. Extending the analysis to departments has been very demanding in terms both of data processing time and volume of downloaded data, since our analysis of 79 universities lead to 632 websites when counting the departmental ones.

The actual web scraping has started with the download of the websites' contents (texts, hyperlinks, HTML tags, meta-keywords, pdf files, etc.). Then, two types of web scraping have been applied: generic web scraping and specific web scraping. Generic web scraping assumes that the structure and the content of a website are not known in advance, so the site is scraped and processed in order to explore its structure first before starting with the information retrieval. In specific web scraping, instead, both structure and contents of the websites are known, so scraping programs just simulate the behavior of a user visiting the website and collecting all available information. As a standard practice, HTTP response codes and related errors are recognized and managed from website scraper program.

The scraped information has been stored in a NoSQL DB according to a semi-structured form to allow for retrieving information efficiently by text mining techniques (Bianchi, Bruni, & Scalfati, 2018). After this, Natural Language Processing steps have been used to determine the meaning of the free text. The tasks implemented for text analysis include information retrieval and information extraction. Information retrieval implements different techniques based on regular expressions (RegEx or RegExp) that are extremely useful in extracting information from texts by searching for one or more matches of a specific search pattern. Finally, a web-mining phase has been carried in order to identify several types of specific contents not easily recognizable with other techniques. In particular, we search for:

- research articles, by looking for their structure (name of journal, abstract, references, etc.) to compute indicator 1 of Table 1;
- content files (pdf, doc, ppt, etc.) to compute indicator 2 of Table 1;
- the names of European research institutions, of Italian research institutions, and of private companies having the top 900 highest budgets for Research and Development, to compute indicator 3 of Table 1;
- teachers' emails, to compute indicator 4 of Table 1.

To complement the above information, we also extract more elementary additional information by submitting batch queries to search engines (in particular Bing, www.bing.com) or by using a leading provider of Web analytics data (SimilarWeb, www.similarweb.com) and the university structural information taken from the European Tertiary Education Register.

4 Results

The described methodology has been used to produce a set of 10 indicators described in Table 1. For each of them, we report the source of information (scraping and mining; search engines; web analytics), the computation formula, and a brief explanation of the rationale.

Table 1. Selected indicators to be used for profiling Italian universities' websites.

No.	Indicator name	Source	Description	Rationale
1.	Access to research	Scraping and mining	Number of research articles / log(number of professor)	Measures the ability to provide access to publications produced by the university, slightly normalized by the size of the institution
2.	Access to content	Scraping and mining	Number of pdf, ppt, doc, rtf, ps / log(number of professor)	Measures the ability to provide consents, slightly normalized by the size of the institution
3.	Orientation to external collaborations	Scraping and mining	Number of research institutions (IT+EU) and research-oriented companies mentioned in the website/ log(number of professor)	Measures the ability of the website to provide a comprehensive description of the extent of on-going research (or Third-mission) collaborations, slightly normalized by the size of the institution
4.	Access to information on teaching	Scraping and mining	Percentage of teachers providing their emails	Measures the possibility to easily get in touch with professors
5.	Visibility	Search engine	Number of HTML pages pointing to the university website from external domains / log(number of professor)	Measures how the university website is visible from outside, slightly normalized by the size of the institution
6.	Usability	Analytics	Percentage of contacts from mobile devices	Level of use by the mobile-oriented audience (largely including students).
7.	Relevance	Analytics	(1/national ranking by visitors) / log(number of students)	Websites' popularity at the national level, slightly normalized by the size of the institution
8.	Intensity of use	Analytics	average time spent on the website / number of pages visited	A key indicator of website effectiveness: the more time is spent on the website, the more relevant will be the available contents
9.	International orientation	Analytics	Percentage of foreign contacts	Popularity abroad as a condition to attract customers (incl. students) and partners
10.	Direct access	Analytics	Percentage of direct accesses	Percentage of non-casual visitors as an indicator of popularity and ability to connect to a population of regular users

The data collection has been designed to be totally web-based and fully transparent/ reproducible, and the data processing was assumed to be: a) suitable for a replication on a regular basis; b) effective in minimizing the influence of university size on websites' effectiveness; c) based on advanced text mining and machine learning techniques.

As already mentioned, the aim of this study is to develop new indicators for profiling Italian universities according to the quality of their web activity. The indicators summarized in Table 1 can complement the many other traditional bibliometric indicators and altmetrics indicators developed for university activities.

The above indicators have also been used to perform a clusterization of Italian universities using the K-means algorithm, a well-established clustering technique. This algorithm aims at partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean (which actually constitutes the centroid of the cluster). The application of this principle leads to a partition of the data space into Voronoi cells. Therefor, data are iteratively clustered in n groups of similar variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

This algorithm has been applied by compressing the above indicators over two dimensions, called Impact and Website Quality, to allow an easier visualization. The first is obtained by a linear combination of indicators 1, 2, 3, 4, 5, 7 with unitary coefficients, while the second is a linear combination of the other indicators 6, 8, 9, 10 with unitary coefficients.

For this experiment, we use the data of 79 Italian universities, including all the Italian universities but excluding, to prevent comparability issues, two universities for foreign students and all the Italian online universities.

K-means algorithm requires the number of clusters k to be specified in advance. To determine the value of k representing the best compromise between distortion and number of clusters, it is often used the "elbow" method, which fits the model with a range of values for k, as reported in Figure 1.

This lead to the choice of 4 clusters. By analyzing the results, we note what follows.

Cluster 1 (blue) includes websites with a high number of visitors (which are indeed those of medium-large universities although the access rates were slightly normalized by university size) and providing extensive information about how to get in touch with the teaching staff (i.e. indirectly, to get information on teaching in general).

Cluster 2 (red) is influenced by the same indicators but rather with a negative sign: low access rates and poor information delivered to users. As a compensation, the quality level of these websites is, on average, higher than those of the other clusters.

Cluster 3 (green) includes several small and highly specialised universities, can be described as poorly performing in terms of Web quality while featuring not irrelevant access rates.

Finally, Cluster 4 (yellow) is representing, in the middle, the group of universities with an average quality level of these websites and significant number of visitors compared to the Cluster 2 and 3.

This exercise has been designed to deliver most of its potential by comparing the website performance over time, thus allowing for spotting any progress in the ability of universities to make their websites increasingly attractive or effective. As a consequence, the description which can be given of the current profiling may reflect existing rankings based on structural and economic indicators.

The above described techniques have been implemented by using the functions K-means and K-ElbowVisualizer, from the Python library scikit learn.

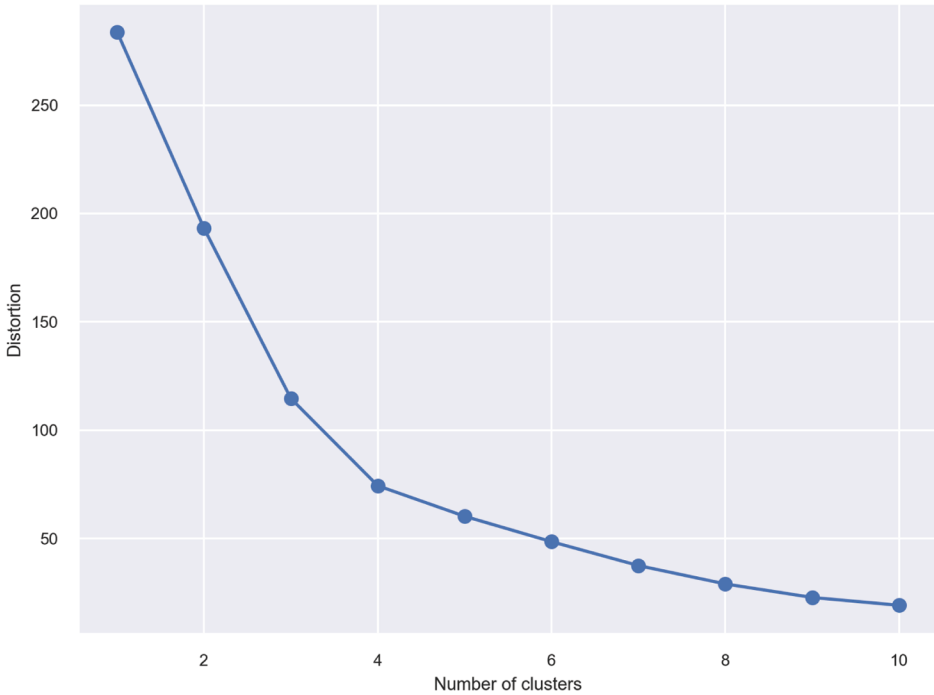

Figure 1. Elbow technique applied for the determination of the number of clusters.

Table 2. Clusters obtained for Italian universities.

Cluster1 (blue)	unito,unimi, polimi, unipd, unibo, unifi,unipi, uniroma1, unina
Cluster2 (red)	unisg, hunimed, iusspavia, unibz, sissa, imtlucca, sssup
Cluster3 (green)	univda,liuc,uninsubria,iulm,unisr,uniurb,unicam,unitus,lumsa,uniroma4,unicampus,unint.eu,unieurm,unilink,unicas,unisannio,uniparth,unisob.na,univaq,unite,unich,unimol,unifg,lum,unisale,unibas,unicz,unirc,unime,unikore,uniss
Cluster4 (yellow)	polito,unige,unibocconi,unicatt,unimib,unibg,unibs,unipv,unitn,univr,unive,iuav,uniud,units,unipr,unimore,unife,univpm,unimc,sns,unisi,unipg,uniroma2,luiss,uniroma3,unisa,poliba,uniba,unical,unipa,unict,unica

Research Paper

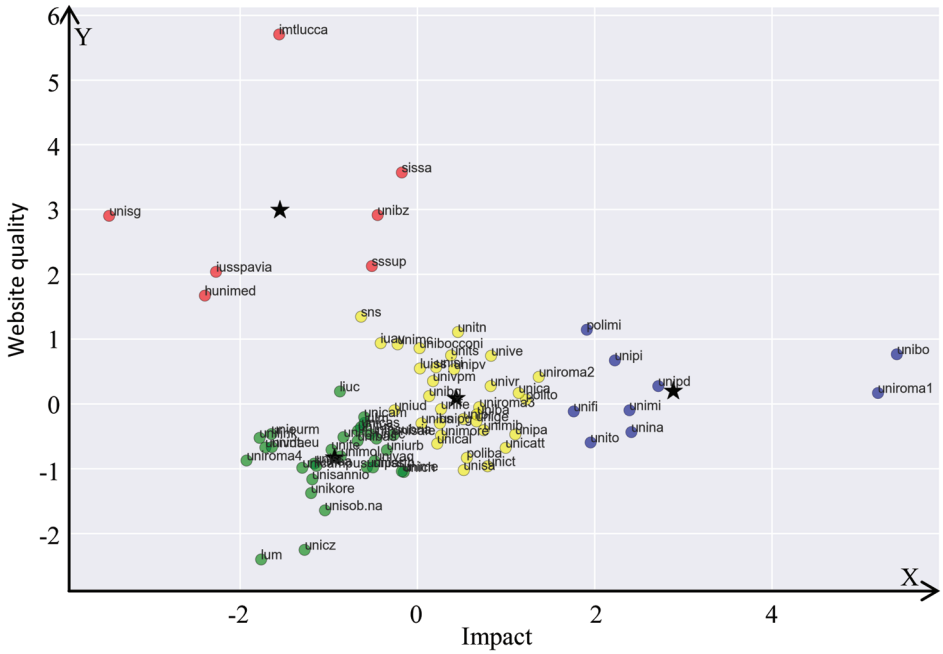

Figure 2. Websites of Italian universities grouped in clusters by quality and impact.

5 Conclusions

This work presents the original application to university websites of a method of quantitative profiling based on automatic knowledge extraction from the web by means of web scraping, text mining, machine learning, and optimization. The methodology proposed in this study and its illustration on the case of Italian universities show its practical relevance for characterizing and profiling the activities of universities using their websites. The proposed approach is completely formal, hence it could be applied to other university systems.

We show the advantages of the automatic extraction of website information to build indicators of quality and impact of the university activities on the web. These new indicators allow superior flexibility and can complement traditional indicators of scientific research (e.g. number of articles and number of citations) and teaching (e.g. number of students and graduates), introducing new dimensions for "profiling" the analysed universities. Moreover, these new indicators can be used to identify groups of universities with common features on the basis of their websites characteristics.

Several expansions of this work are in preparation and left for future studies, including the extension of the analysis to several time periods in order to assess the ability of Italian leading academic institutions to develop their Web strategies in a comparable way. On the methodological side, the areas where most of the development efforts will be focused include: improving the quality of Web analytics; testing a web-scraping activity reading even more of the Web portal structure; developing further machine learning routines to extract additional information from the scraped data.

Acknowledgments

This work is developed with the support of the H2020 RISIS 2 Project (No. 824091) and of the "Sapienza" Research Awards No. RM1161550376E40E of 2016 and RM11916B8853C925 of 2019. This article is a largely extended version of Bianchi et al. (2019) presented at the ISSI 2019 Conference held in Rome, 2–5 September 2019.

Author contributions

All authors contributed equally to the research and to the preparation of the article.

References

Aggarwal, C.C. (2018). Machine learning for text. Springer.

Aguillo, I.F., Ortega, J.L., & Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. Higher education in Europe, 33(2–3), 233–244.

Aguillo, I.F., & Orduna-Malea, E. (2013) The Ranking Web and the "World-Class" Universities: New Webometric Indicators Based on G-Factor, Interlinking, and Web 2.0 Tools. In book: Building World-Class Universities pp. 197–217. doi: 10.1007/978-94-6209-034-7_13

Bar-Ilan, J. (2019). Data Collection from the Web for Informetric Purposes. In Springer Handbook of Science and Technology Indicators (pp. 781–800). Springer, Cham.

Bianchi, G., R. Bruni, & F. Scalfati, (2018). Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms. Mathematical Problems in Engineering, 7231920.1-7231920.8.

Bianchi, G., Bruni, R., Laureti Palma, A., Perani, G., & Scalfati, F. (2019). The corporate identity of Italian Universities on the Web: a webometrics approach. In the Proceedings of the 2019 ISSI Conference ISSI (pp. 2273–2278).

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 55(14), 1216–1227.

Bruni, R., & Bianchi, G. (2015). Effective classification using a small training set based on discretization and statistical analysis. IEEE Transactions on Knowledge and Data Engineering, 27(9), 2349–2361.

Bruni, R., Bianchi, G., Dolente, C., & Leporelli, C. (2019). Logical Analysis of Data as a Tool for the Analysis of Probabilistic Discrete Choice Behavior. Computers & Operations Research, 106, 191–201.

Bruni, R., & Bianchi, G. (2020). Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. Expert Systems with Applications, 142, 113001.

Bychkova1, M.N., & Okushova, G.A. (2017). Methods of analysis of a modern university's presence in the Internet communicative space. AI & Society, 32, 89–100.

Daraio, C., & Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. Journal of the Association for Information Science and Technology, 68(2), 508–529.

Daraio, C., & Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. Scientometrics, 108(1), 391–400.

Daraio, C., Bruni, R., Catalano, G., Daraio, A., Matteucci, G., Scannapieco, M., ... , & Lepori, B. (2020). A tailor-made data quality approach for higher educational data. Journal of Data and Information Science, 5(3), 129–160.

Elgohary, A.E. (2008). Arab universities on the web: A webometric study. Electronic Library, 26(3), 374–386.

Glänzel, W., Moed, H.F., Schmoch, U., & Thelwall, M. (2019). Springer Handbook of Science and Technology Indicators. Springer Nature.

Göransson, B., & Brundenius, C. (2010). Universities in transition: The changing role and challenges for academic institutions. Springer Science & Business Media.

Islam, M.A., & Alam, M.S. (2011). Webometric study of private universities in Bangladesh. Malaysian Journal of Library and Information Science, 16(2), 115–126.

McCoy, C.G., Nelson, M.L., & Weigle, M.C. (2018) Mining the Web to approximate university rankings. Information Discovery and Delivery, 46(3), 173–183.

Pal, A., Sarkar, A., & Bhattacharya, U. (2019). Webometric analysis of open universities in India. Library Philosophy and Practice. 3038.

Seeber, M., Lepori, B., Lomi, A., Aguillo, I., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. Journal of informetrics, 3, 435–447.

Thelwall, M. (2019). Online Indicators for Non-Standard Academic Outputs. In Springer Handbook of Science and Technology Indicators (pp. 835–856). Springer, Cham.

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. Synthesis lectures on information concepts, retrieval, and services 1.1 p. 1–116.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). "Webometrics." Annual Review of Information Science and Technology 39.1 p. 81–135.

Vaughan, Liwen, & R. Yang. (2013). Web traffic and organization performance measures: Relationships and data sources examined. Journal of Informetrics 7.3 p. 699–711.

Wouters, P., Zahedi, Z., & Costas, R. (2019). Social media metrics for new research evaluation. In Springer handbook of science and technology indicators (pp. 687–713). Springer, Cham.