The corporate identity of Italian Universities on the Web: a webometrics approach

G. Bianchi[2], R. Bruni[1], A. Laureti Palma[2], G. Perani[2], F. Scalfati[2]

[1] *bruni@diag.uniroma1.it*
Dept. of Computer Control and Management Engineering, Sapienza University of Rome, Rome (Italy)

[2.] *perani@istat.it*
ISTAT – Italian National Institute of Statistics, Via Cesare Balbo 16, 00184, Rome (Italy)

Abstract

In parallel with the increasing marketisation and globalisation of higher education, Universities' corporate websites have become institutional virtual storefronts largely contributing to reinforcing the organisations' brand, to disseminate information on their main achievements and to communicate with both enrolled students and potential "customers" worldwide. Thus, the effectiveness of Universities' websites to deliver value in terms of information on the organisations' activities and to interact with actual and potential students - as well as partner institutions in education and research projects - is to be regarded as a key objective of all Universities. The level of accomplishment of this task, measured so far mostly on a case-study basis, can be more extensively surveyed by adopting a webometric approach combining the use of web analytics as indicators of efficiency with selected indicators of contents collected through web scraping techniques. This approach has been tested on the websites of Italian Universities with the aim of classifying them in terms of quality and impact of their institutional websites, as well as to develop a permanent monitoring framework.

Introduction

The ability of academic institutions to effectively play the multiple roles of educational agencies, research hubs and drivers of innovation processes, in close connection with business enterprises and other organisations (Göransson and Brundenius, 2010), has become a key topic of research and policy action. The 'open innovation' paradigm – increasingly diffused in developed countries – assumes that knowledge can be freely transferred across economic sectors thus making attractive for businesses to give up large internal research facilities and replacing them with a network of potential partners – universities, research centres, start-ups, SMEs, customers, etc. – providing the technical and managerial knowledge needed to feed the innovation processes (Chesbrough, 2003). Universities are a privileged source of knowledge and innovations, as well as of educational services, and research has been long focusing on the measurement of the level of interaction between universities and external actors (West *et al.*, 2014). Leading universities are becoming more open (Lepori *et al*, 2015; Dennis *et al*, 2016; Pharr, 2016; Foroudi *et al*, 2019) as a condition not just for success but for survival in a context of increasing marketisation of higher education. Absolutely crucial is to develop their own identity (Steiner *et al*, 2013) and brand (Delmestri *et al*, 2015) in order to compete for attracting the most talented people (and, as a consequence, an increasing amount of funding). These efforts need to be pursued at a global scale, by adopting standard methods of communication and knowledge sharing and, most important, by extensively use digital technologies as key enablers. The digital transformation and the global competition are thus forcing the universities to foster their ability to communicate on the Web about their activities, capabilities and achievements. These two phenomena are intertwined, as a high-impact Web communication is a powerful driver to improve reputation, developing the brand, connecting with potential partners, and attracting funders and customers (including students).

Objectives of the study

Two methods are commonly used for evaluating the quality of universities, as well as their ability to meet the institutional objectives (higher education, research and the so-called third-mission, i.e. knowledge transfer): a) *institutional evaluation exercises* focusing either on academic programs or research outputs (long, complex, detailed and expensive exercises carried out at national level with a multi-year frequency); b) *rankings* (relying on informal data collections based on a range of available sources, quite often not very detailed, with annual frequency) (Shin *et al*, 2011). University rankings are increasingly using data freely available on the Web but often with a poor ability to properly check them for data quality.

Under the assumption that a corporate website is going to become the main communication channel between universities and external actors by delivering services (Bernier *et al*, 2002), attracting new students (Arslan *et al*, 2018) or interacting with potential or actual partners (Chu, 2005; Seeber et al, 2012 with reference to university-to-university links), a webometric approach can be adopted in order to draw a "university profile" as a result of the analysis of its corporate website. The webometric approach is extensively adopted to produce Web-based statistics (Thelwall, 2009, Thelwall *et al*, 2005; Björneborn and Ingwersen, 2004) and, more specifically, to collect information on higher education institutions from their websites. Several international universities' rankings use, at least partially, data extracted from the Web but no one of them relies exclusively on information available from universities' websites.

Universities' corporate websites have thus become key sources of information being: a) the main gateways to access both general information and specific contents universities are disseminating; b) delivery points of educational services and c) contact points to develop research and co-operation networks. In this respect, to improve the quality of the universities' websites and their effectiveness as information hubs is a key corporate task (Kaur *et al*, 2016).

The exercise described below is aimed at collecting information on Italian universities through their corporate websites, similarly to existing rankings, but with two rather distinctive features:

- *to be exclusively based on information available on universities' websites, thus to be potentially updated with a frequency higher than once per year;*
- *to collect a set of indicators about the efficiency of universities' websites and their effectiveness in disseminating key contents in order to produce a "profiling" of Italian universities, rather than a ranking.*

Methodology

In order to achieve the mentioned objectives, a data collection and data processing exercise has been developed. Raw data have been collected from two main sources:

- A leading provider of Web analytics indicators (http://www.similarweb.com) has been used to draw a set of indicators about the quality/efficiency of Italian university websites.
- A web-mining task has extracted selected contents from the same websites.

Web analytics are regularly produced by highly specialised web companies that monitor the performance of commercial websites, in comparison to their direct competitors, in order to increase their attractiveness for customers and profitability. When using analytics of public or non-profit institutions' websites, the aim is usually that of assessing their effectiveness in communicating with the public or in delivering online services. In this study, still at a pilot stage, the analytics covering the last six months of activity of Italian universities' websites have been collected. The frequency of data collection and the related coverage of time (for instance, extending it to one year) will be increased in the next stage of the study. In parallel, by considering that the choice of the data provider could affect the quality of the data, several commercial data providers will be compared to select the most reliable source.

The study has also included an advanced application of the webometric approach. Shortly, webometrics uses three categories of web mining: web content mining; web structure mining; web usage mining[i]. In the first stage of this study the *web content mining* has been mostly adopted. It involves the analysis of unstructured text data in webpages in order to translate it into structured information (e.g. to find connections between academic web portals and external organisations, as in this study). The web scraping activity has been limited to the 'second level' of the websites' link structure in order to reduce the volume of scraped data.

The scraping activities on universities' websites has three main steps: a) acquisition of the universities' web addresses; b) validation of the websites; c) data extraction. By using official sources of information on the Italian tertiary institutions, the official names of universities have been used as *search strings* by search engines in order to keep web addresses with a high matching probability assessed via a machine learning approach. In addition to the identification of universities Web portals, it has also been needed – for a few large universities - to extend the analysis to University Departments' websites (increasing both data processing time and volume of downloaded data).

Data collection has started by downloading contents from the targeted websites (texts, hyperlinks, HTML tags, meta-keywords, pdf files, etc.) by using a web scraping procedure. This has allowed to explore the websites' structure and collect all available information by using text mining techniques. The scraped information has been stored in a semi-structured format to allow for efficient information retrieval (Bianchi *et al*, 2018a; Bianchi *et al*, 2018b). These require the integration of natural language processing (to extract meaning from free text) with advanced machine learning (Bruni and Bianchi 2019).

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

No.	*Indicators*	*Area*	*Description*	*Rationale*
1.	Relevance	Analytics	(1/national ranking by visitors) / log(number of students)	Websites' popularity at the national level is the key indicator of effectiveness for universities mainly enrolling Italian students
2.	Usability	Analytics	Percentage of contacts from mobile devices	Level of use by the mobile-oriented audience (largely including students).
3.	Identifiability	Analytics	100- bounce rate	A higher level of visitors leaving the website after the visualisation of the main page is an indicator of a low ability of the website (or the university) to be identifiable
4.	Intensity of use	Analytics	Number of pages visited * average time spent on the website	A key indicator of website effectiveness: the more time is spent on the website, the more relevant will be available contents for users
5.	International orientation	Analytics	Percentage of foreign contacts	Popularity abroad as a condition to attract customers (incl. students) and partners
6.	Visibility	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
7.	Use of social media	Analytics	Percentage of accesses from social media	Degree of orientation to the use of social media
8.	Access to information on teaching	Contents	Number of e-mail address / number of professors	Measures the ability of students to easily get in touch with professors
9.	Access to data and outcomes	Contents	Number of pdf documents / log(number of students)	Measures the ability of users to have access to relevant documents (including learning materials and research outcomes) produced by the university
10.	Orientation to external collaborations	Contents	Number of firms+research institutions (IT+EU) mentioned in the website	Measures the ability of the website to provide a comprehensive description of the extent of on-going research (or Third-mission) collaborations
11.	Link impact studies (URL-degree)	Contents	Number of hyperlinks pointing to each University website	Measures of the numbers of hyperlinks pointing to each website

This methodology has been used to produce a set of eleven indicators (Table 1) combining Web analytics and website contents data. Seven variables (analytics) focus on the intensity of use of universities' websites, as well as highlighting some key features of users and their access modes (whether direct or indirect access, users from Italy or from abroad, Web traffic from social media, etc.). Four indicators have been drawn from scraped data: percentage of professors' e-mail available; number of pdf documents (weighted by university size) i.e. volume of information available to users; number of EU firms or research institutions mentioned in the website; number of hyperlinks pointing to each University website, as a link impact metrics. In particular, tenth indicator referring to develop international partnerships.

In order to identify the key features of the data collection stage of this study, it can be pointed out that it has been designed to be: a) Totally Web-based; b) Fully transparent/reproducible; c) Based on data from a leader company in Web monitoring and from systematic web-scraping to implement a web-mining approach on universities' websites. On the other hand, data processing was assumed to be: a) suitable for replication on a regular basis; b) effective in minimising the influence of university size on websites' effectiveness; c) based on state-of-the-art text mining and advanced machine learning techniques.

Moreover, at this stage, also considering that official data on the third mission of Italian universities do not exist, it is possible to confirm a positive linear dependence (Figure 1a) between key indicators of analytics (relevance) and scraped contents (collaboration, i.e. number of EU firms and research centres cited in the websites).

A critical issue of the study is the potential inconsistency between data analytics set and scraped data set. The eleventh indicator should allow for validating the joined use of this two sets of data. Figure 1b shows a remarkable concordance between the URL-degree (indicator 11) and Relevance (indicator 1) that supports the combined use of the two sets of data.

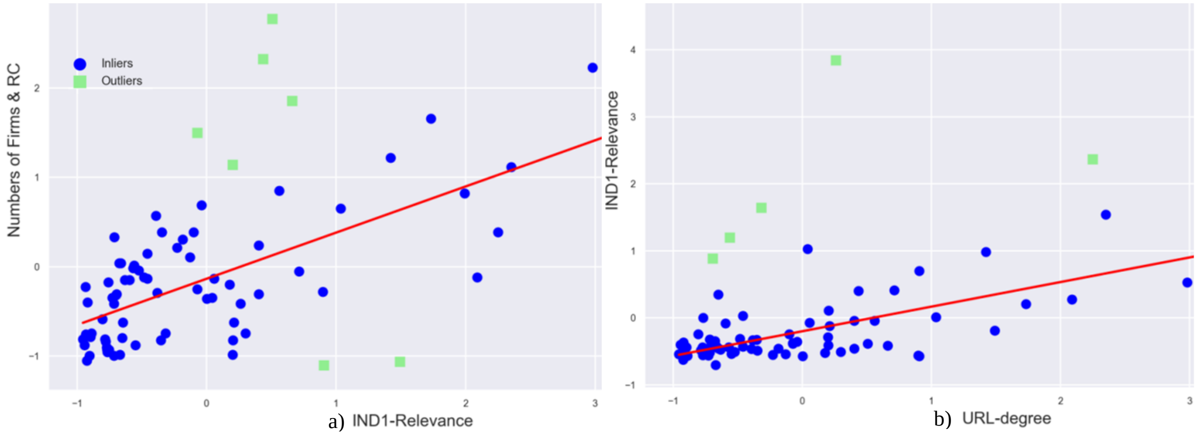

Figure 1. Linear regressions based on random sample consensus fits model. Scatter plot of: a) Relevance (Indicator-1) vs. Orientation to external collaborations (Indicators 10); b) URL-degree (Indicators 11) vs. Relevance (Indicators 1).

Results

The aim of this study was more that of profiling Italian universities according to their Web activity, rather than comparing them and their performances. The potential of the indicators in Table 1 has been tested by running a cluster analysis (FASTCLUS[ii] in SAS) which allowed for the identification of three main profiles. The analysis focused on 79 Italian universities (two universities for foreign students and all Italian online universities have been excluded).

In Figure 2, such profiles are described with reference to two canonical variables, respectively describing (X) the websites' impact on users (mostly based on indicators 1 'Relevance' and 8 'Information on teaching') and (Y) the level of websites' quality (indicators 2, 3, 4, 6 and 7). As a result, three clusters have been identified. Cluster 1 (red) includes websites with a high number of visitors (which are those of medium-large universities although the access rates were weighted 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 (blue) is influenced by the same indicators but rather with a negative sign: low access rates and poor information delivered to users. As compensation, the quality level of these websites is, on average, higher than that of the other clusters. Finally, Cluster 3 (yellow), including several small and highly specialised universities, can be described as poorly performing in terms of Web quality while featuring non-irrelevant access rates. 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 and effective. The description which can be given of the current profiling results may be neither relevant per se, nor totally new compared to existing rankings based on structural and economic indicators (Shin, 2011; Aguillo, 2010).

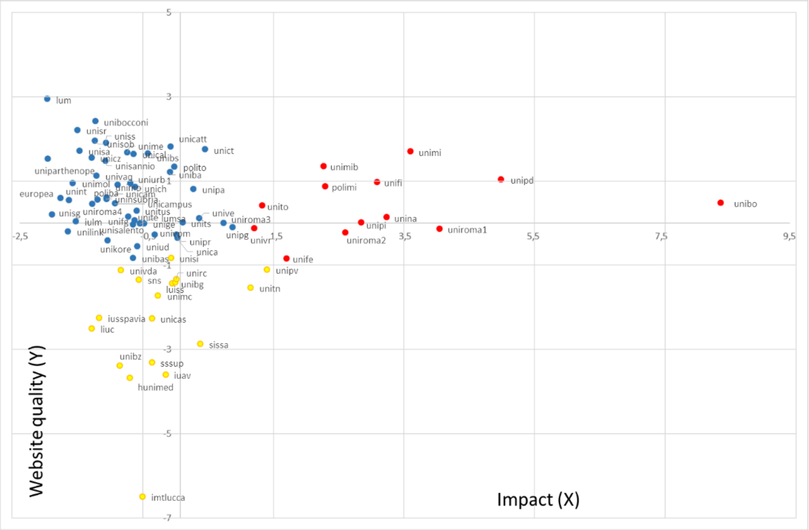

Figure 2. Websites of Italian universities grouped in clusters by quality and impact

Next steps

This project of profiling Italian universities by adopting a webometric approach aims at filling the need for a timely and neutral assessment of the ability to improve their competitiveness in a global context. The first stage is, necessarily, that of profiling them by using data available on the Web (i.e. the information available, in principle, to anyone would be interested to get in touch with them). A second stage will be that, of course, of extending the analysis to several time periods in order to assess the dynamics of an effort to develop Web strategies – including branding – over time in parallel with potential competitors/partners abroad. But again, the stage of profiling, focusing on how they communicate, is a preliminary stage for any meaningful measurement of effective performance in a digitalised environment. On the methodological side, the next stages of the study will improve the framework defined so far with respect to several actions of collecting and processing data. Areas where most of the development efforts will be focused will include: improving the quality of Web analytics; testing a web-scraping activity reading more than two layers of a Web portal structure (i.e. addressing the issue of website quality); developing more accurate machine learning routines to process scraper data.

References

Aguillo, F.I., Bar-Ilan, J., Levene, M & Ortega, J.L. (2010). Comparing university rankings. Scientometrics 85.1 p.243-256.

Arslan, Y., Evren, S. & P. Soner. (2018). The Relationship between International Students Attitude toward the University Website and University Attractiveness. *Journal of the Faculty of Education* 19.1 p. 200-211.

Bernier, J.L., Barchéin, M., Cañas, A., Gómez-valenzuela, C. & Merelo, J. (2002). The services a university website should offer. *Information Society and Education: Monitoring a Revolution. Serie Sociedad de la Educación* 9 p.1746-1750.

Bianchi, G., Laureti Palma, A. & Quaresma, S. (2018a). Prepare your data warehouse for a Big Future, by including Big Data. *European Conference on Quality in Official Statistics 2018, Krakow*.

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

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

Bruni, R., Bianchi, G. (2019). Robustness Analysis of Classifiers for Website Categorization: the Case of E-commerce Detection. *Expert Systems With Applications*, to appear.

Chesbrough, H. (2003). Open Innovation. The New Imperative for Creating and Profiting from technology. *Harvard Business School Press, Boston*.

Chu, H. (2005). Taxonomy of inlinked Web entities: What does it imply for webometric research? *Library & Information Science Research* 27.1 p. 8-27.

Delmestri, G., Oberg, A. & Drori, G. S. (2015). The unbearable lightness of university branding: Cross-national patterns. *International Studies of Management & Organization* 45.2 p. 121-136.

Dennis, C., Papagiannidis, S., Alamanos, E. & Bourlakis, M. (2016). The role of brand attachment strength in higher education. *Journal of Business Research* 69.8 p. 3049-3057.

Pantea, F., Qionglei, Y., Suraksha, G. & Mohammad M. F. (2019). Enhancing university brand image and reputation through customer value co-creation behaviour. *Technological Forecasting and Social Change* 138 p. 218-227.

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

Huang, M. (2012). Opening the black box of QS World University Rankings. *Research Evaluation* 21.1 p. 71-78.

Kaur, S., Kaur, K. & P. Kaur. (2016). An empirical performance evaluation of universities website. *International Journal of Computer Applications* 146.15 p. 10-16.

Lepori, B., Seeber, M. & Bonaccorsi A. (2015). Competition for talent. Country and organizational-level effects in the internationalization of European higher education institutions. *Research policy* 44.3 p. 789-802.

Montazer, G. (2018). University Website Quality Improvement Using Intuitionistic Fuzzy Preference Ranking Model. *Quarterly Journal of Iranian Distance Education* 1.2 p. 9-30.

Pharr, J M. (2016).University Branding 2.0 Harnessing the Power of Social Media for Open-Source Branding and Brand Co-Creation of Colleges and Universities. *Kennesaw State University*, paper.

Rauschnabel, P.A., Krey, N., Babin B.J. & Ivens B.S. (2016). Brand management in higher education: the university brand personality scale. *Journal of Business Research* 69.8 p. 3077-3086.

Seeber, M., Lepori,B., Lomi,A., Agiullo,I. & Barberio,V. (2012). Factors affecting web links between European higher education institutions. *Journal of informetrics* 6.3 p. 435-447.

Shin, J. C., Toutkoushian, R. K. & Teichler, U. (Eds). (2011). University rankings: Theoretical basis, methodology and impacts on global higher education. *Vol. 3. Springer Science & Business Media*.

Steiner, L., Sundström, A. C. & Sammalisto, K. (2013). An analytical model for university identity and reputation strategy work. *Higher Education* 65.4 p. 401-415.

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, L. & Yang, R. (2013). Web traffic and organization performance measures: Relationships and data sources examined. *Journal of informetrics* 7.3 p. 699-711.

West, J. Salter, J.A., Vanhaverbeke, W. & Chesbrough H.W. (2014). Open innovation: The next decade. *Research Policy*, Volume 43, Issue 5, p. 805-811.

[i] Web mining techniques have been used to extract the Web analytics used in variables 1 to 7 (Table 1).

[ii] Additional details on the analysis carried out and regression results are available on request.