Improving the webometrics approach for profiling Italian Universities by using web analytics data.

**Keywords:** webometrics, web mining, text mining, machine learning, big data, methodological framework, profiling universities, open innovation.

# Introduction

The ability of academic institutions to effectively play a multiple role as educational agencies, research hubs and drivers of innovation processes, in close connection with the business enterprises [1], has become a key topic of research and is also attracting some interest by policymakers. 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 to build up a network of potential partners – universities, research centres, start-ups, SMEs, customers, etc. – which could provide most of the technical and managerial knowledge needed to feed the innovation processes [2]. Universities are a special source of external knowledge and innovations, and research has already measured the economic benefits of university technology provided to businesses [3].

A key driver of the diffusion of the open innovation paradigm is the digital transformation. It affects business enterprises, as well as all the actors interacting with them, by inducing a radical change in the organisation of public and private institutions with a high impact on the way they interact with external partners [4].

The implications for universities of these ongoing processes should not be neglected. On the one hand, universities have to become aware of the global competition emerging in the educational sector. Given the increasing international circulation of students and of the tertiary education’s teaching staff, the effort to attract – also in small universities – the most talented people (and, as a consequence, an increasing amount of funding) is often a condition not just for success but very much for survival. This means to become active at global level and to use effectively those digital technologies which can make they able to achieve this objective. 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 strictly intertwined as a high-impact communication on the Web is a powerful driver to improve reputation, to connect with potential partners, to attract funders and customers (including students).

# OBJECTIVES OF THE STUDY

The current methods for evaluating the efficiency of universities, as well as their effectiveness in meeting their institutional objectives (higher education, research and the so-called “third-mission”, i.e. knowledge transfer) are mostly two: a) institutional evaluation exercises (long, complex, very detailed and very 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). University rankings often use data available on the Web but with a poor ability to properly check for the quality of the data used.

In this perspective, it could be assumed that there is room for developing a partially new approach for measuring universities’ performances and capacities with reference to the Italian case [5-7]. It should be based on a few points:

* *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*.

This paper investigates the options for developing, with reference to Italian universities, an integrated data collection and data processing method with the following features:

* Data collection:
	+ Totally web-based
	+ Fully transparent/reproducible
	+ Based on three main data sources: administrative data sources (European Register of Tertiary Education institutions, Horizon database of EU-funded R&D projects, etc.), collection of data analytics for universities’ websites, web-scraping to implement a web-mining approach on universities’ websites.
	+ Updating dependent on data providers (national/international institutions and statistical agencies) for administrative data, every six months /one year for web scraping, monthly for web analytics.
* Data processing:
	+ Set of variables produced for all Italian universities on a regular basis
	+ Normalization of variables to prevent size biases when comparing large and small universities
	+ Discriminant analysis to identify the key factors which affect universities’ performance and competitiveness
	+ Text mining components of large-scale data retrieval from websites
	+ Advanced machine learning and cluster analysis to profile universities and to identify common patterns of activity.

# METHODOLOGY

The study can be described as an advanced application of the Webometrics approach. Webometrics is the “study of the quantitative aspects of the construction and use of information resources, structures and technologies on the web drawing on bibliometric and informatics approaches” [8]. Web pages allow us to study online web reality, such as hyperlinking among websites, as well as off-line hidden relationships, such as political attitudes reflected in blogs. In order, for instance, to sizing innovation activity, we make use of Webometrics technics to build metrics able to quantify the link interrelationships between university and enterprises. Webometrics is based on web mining methods and technics to perform quantitative analyses of the web.

To these aims, 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](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.

When using web 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 study has also included an advanced application of the webometric approach based on three categories of web mining: web content mining; web structure mining; web usage mining. 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).

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  |

The collected information have 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 check the quality level of the different sources of information used, considering that official data on the third mission of Italian universities do not exist, we have checked a positive linear dependence between key indicators. In figure 1a it is showed the linear dependence between Analytics (relevance) and scraped contents (collaboration, i.e. number of EU firms and research centres cited in the websites). In figure 1b, the eleventh indicator should allow for validating the joined use of this two sets of data, the picture 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).**

b)

a)

# Web mining application framework

Webometric indicators are derived by a big data methodological framework, based on web scraping, text mining and advanced machine learning components for data analytics of large-scale data retrieval from websites. In particular we focus on web sites of academic institutions and the directly related business enterprises.

In Figure 1, we show a schema of data warehouse environment on which is based the methodological framework [9]. This is an analysis framework for generating web-based academic innovation metrics. Alike to traditional processes, it starts from the statistical registers of academic institutions and enterprises.

Figure 1. big data methodological framework based on a data warehouse architecture.

In the source layer, the statistical registers’ URLs are identified on the web, checked and whenever necessary updated. In case of available URL from administrative sources the procedure checks URL’s validity. In particular, it does syntactic validation of the strings, the check of the recurring errors and extracts the authority of the web addresses. Then, in case of non-existing URL the procedure uses search engines in order to find the most similar URL and computes its probability of correctness by using a machine learning approach.

In case of non-available URL from administrative sources the procedure uses URL Retrieval techniques. It uses identification characteristics in order to perform batch queries on the search engines. In particular, it uses the denomination of the enterprise as a search string and for each obtained link it evaluates the probability of correctness using a machine learning approach. The link whose probability exceeds a given threshold is accepted as valid.

This is realized both for Academic and Enterprises contexts.

After, the URLs web addresses are passed to a web scraper. The web scraper is then used to download website content (texts, hyperlinks, HTML tags, meta-keywords, etc.) from the websites. The kinds of web scraping that we have been considered are: i) generic web scraping; ii) specific web scraping.

Generic web scraping assumes that the structure and the content of a website aren’t known in advance so the site is scraped and processed in order to collect information of interest. Therefore, specific information will be retrieved in the next phases.

Specific web scraping is performed in case both structure and content of websites to be scraped are well known. Scraping programs have to simulate the behaviour of an user visiting the website and collecting all needed information.

The scraped unstructured data are then organized and prepared for the text mining phase [10]. Text Mining is the branch of Data Mining concerning the process of deriving high-quality information from texts and requires the integration of natural language processing techniques with several advanced machine learning techniques. References can be found for instance in [11].

In the integration layer, text mining techniques are applied to extract information on academics’ innovation activities from the downloaded website content (publications, references, links, enterprises). Based on this information, novel metrics are constructed combining automatic process with *ad hoc* data mining activities. The framework maximize the ability of using both approach together with all necessary additional information useful to support the analysis (pre-classification, classification model selection based on enterprise characteristics, information from established innovation indicators etc.).

CLUSTER 1

CLUSTER 2

CLUSTER 3

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

# 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[[1]](#endnote-1) 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 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 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, 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).

Moreover, webometrics techniques are used to extend the functionalities of the Big data methodological framework addressing the web mining activities toward more structured processes suitable for official statistics.

# Conclusions

This project of profiling Italian universities by adopting a Webometric approach is aimed at filling the need of a timely and neutral assessment of the ability of these universities to stay competitive in the international contexts. For a similar exercise see [12,13]. The first stage is, necessarily, that of profiling them by using data available on the Web (which is the information available, in principle, to anyone would be interested to get in touch with them). A second stage would be that, of course, of extending the analysis to a range of foreign universities in order to compare the Italian leading academic institutions with their potential competitors/partners abroad. But again, the stage of profiling, by highlighting strengths and weaknesses of universities – also, and most important, in the way they communicate them – is the key stage for any meaningful measurement of effectiveness.

# References

1. Göransson, Bo, and Claes Brundenius, eds. *Universities in transition: The changing role and challenges for academic institutions*. Springer Science & Business Media, 2010.
2. Chesbrough, Henry, *Open Innovation. The New Imperative for Creating and Profiting from technology*, Harvard Business School Press, Boston, 2003.
3. West, Joel, et al. “Open innovation: The next decade”. *Research Policy*, Volume 43, Issue 5, June 2014, Pages 805-811.
4. Link, A.N., and Rees, J. “Firm size, university based research, and the returns to R&D”. *Small Business Economics* 2 (1), 1990, Pages 25-31.
5. Thelwall, Mike, Liwen Vaughan, and Lennart Björneborn. "Webometrics." *Annual review of information science and technology* 39.1 (2005): 81-135.
6. Thelwall, Michael. “Introduction to webometrics: Quantitative web research for the social sciences”. *Synthesis lectures on information concepts, retrieval, and services* 1.1 (2009): 1-116.
7. Aguillo, Isidro F. “Altmetrics of the Open Access Institutional Repositories: A Webometrics Approach”. *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS), 2018.
8. Björneborn, L., & Ingwersen, P “Toward a basic framework for webometrics”. *Journal of the American Society for Information Science and Technology*, 2004, 55(14), Pages 1216-1227.
9. Bianchi G., Laureti Palma A., Quaresma S., *“Prepare your data warehouse for a Big Future, by including Big Data”,* European Conference on Quality in Official Statistics 2018, Cracovia, Polonia. 26-29 June 2018.
10. Bianchi G., Bruni R., Scalfati F., “Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms”, *Mathematical Problems in Engineering*, vol. 2018, Article ID 7231920, 8 pages, 2018.
11. Bianchi G., Bruni R., “Effective Classification using a Small Training Set based on Discretization and Statistical Analysis”, *IEEE Transactions on Knowledge and Data Engineering* Vol. 27(9), 2349-2361, 2015.
12. Vásquez, Carmen, et al. “Cluster of the Latin American Universities Top100 According to Webometrics 2017”. *International Conference on Data Mining and Big Data*. Springer, Cham, 2018.
13. -Bianchi G., Bruni R., Laureti Palma A., Perani G., Scalfati F. “The corporate identity of Italian Universities on the Web: a webometrics approach”, Rome, 17th International Conference on Scientometrics & Informetrics, 2019.
1. [↑](#endnote-ref-1)