4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15090

Changes in corporate websites and business activity: automatic classification of corporate webpages

Joan Manuel Valenzuela Rubilar, Josep Domenech, Ana Pont
Universitat Politècnica de València, Spain.

Abstract

Every time a firm or institution performs an activity on the Web, this is registered, leaving a "digital footprint". Part this digital footprint is reflected on their websites as these officially represent them on the Web. We plan to automatically monitor the changes that periodically occur in a website to relate them with the business activity. The aim of this paper is to propose a theoretical classification of corporate webpages to associate changes that occur on them with the regular activity of the firms, and to evaluate the possibility of an automatic categorization using classification models. To generate the classification of corporate webpages, a significant number of today corporate webpages were analyzed and observed, distinguishing four theoretical types of corporate webpages. To evaluate the automatic categorization of corporate webpages, a dataset of 1005 today corporate pages was generated by manually labeling them and evaluating their automatic categorization using classification models.

Keywords: *Corporate websites; Webpages classification; Websites changes.*

1. Introduction

Nowadays, the daily activities of entities and individuals on WWW (the Web) generate tons of fresh and digitized data. These data are commonly referred to as "Big Data". Properly analyzing these data could help to reveal trends and monitor economic, industrial and social behavior or magnitude (Blazquez and Domenech, 2018). Every time a firm or institution performs an activity on the Web, this is registered, leaving a "digital footprint". Part this digital footprint is reflected on their websites as these officially represent them on the Web.

Corporate websites help to obtain direct information about the strategic variables that can define firms and allow to describe the corporate profile of the firms (Llopis et al., 2019) and institutions as they are regularly updated by the firms and institutions themselves. Thus, firm data are available on corporate websites almost in real time (Crosato et al., 2021) and can therefore be considered up-to-date and reliable sources of information on the activities of firms and institutions that also reflect, to a large extent, their health and behavior.

Today's websites are composed of several webpages. An important characteristic of web content in general is the constancy of its change (Han et al., 2019). The webmasters regularly add new products, modify textual content, include new photos and insert and remove links and webpages, etc. (Llopis et al., 2010). But in addition, web content can change because of the spontaneous interactions of the users (Calzarossa and Tessera, 2018). Besides, webpages can appear and disappear on the Web all the time (Calzarossa and Tessera, 2018). Thus, changes in corporate webpages represent changes in the firms.

These changes on webpages of corporate websites are of varying significance at the level of business activity. For example, a structural change in the organization chart page does not have the same significance as the day-to-day addition of new product pages. To detect the types of changes on corporate webpages and to evaluate their business activity significance, it is important to generate a classification of today corporate webpages.

To monitor the changes in corporate websites, the aim of this paper is to propose a theoretical classification of firm webpages to associate changes on them with the regular activity of the firms, and to evaluate the possibility of an automatic categorization using classification models.

2. Literature review

This section reviews research papers on the dynamics of web content changes and classifications for web content.

Today's corporate websites are unstructured and non-traditional sources of social and economic Big Data. They are also dynamic over time. That is, during its finite lifespan —

an average of 2 years and 7 months, according to Crestodina (2017) —, they undergo structural, content, technological and other changes.

To cope with the highly dynamic behavior of corporate websites, it is necessary to predict how often and to what extent their content changes (Calzarossa and Tessera, 2018). That is, it is necessary to identify and classify the changes occurring on the corporate webpages to convert them into knowledge about the health and behavior of firms and institutions and ultimately to make economic forecasts about the firms. In this way, Calzarossa and Tessera (2018) presents a methodological framework — based on time series analysis — for modeling and predicting the dynamics of the web content changes.

Focused in understanding temporal dynamics and evolution of topics on the Web, Santos et al., (2016) developed a methodology to monitor webpages that belong to a same topic. Their results show, among other things, that distinct topics have different change patterns.

This is in line with Radinsky and Bennett (2013), who state that webpages are dynamic channels whose contents change with time, in their work on predicting content change on the Web.

Regarding the classification of the content of the web content of webpages, Zhou and Sun (2014) distinguished that the web content often comes in two camps: Evergreen content frequently don't change with time, whereas ephemeral content easily become dated.

From a more general view of corporate sites, based on users' expectations and the direct purpose of the websites, Cebi (2013) classifies the websites into (i) commercial websites, to make money selling products or services; (ii) service websites, which present various services free of charge; and (iii) mixed-type websites which present two or more purposes at the same time.

Finally, in relation to how to rescue and treat information extracted from non-traditional, non-structured Big Data sources, Blasquez and Domenech (2018) propose a flexible Big Data architecture for nowcasting and forecasting social and economic changes, applicable to data sources such as corporate websites.

Based on this review, it was decided to observe and analyze a significant number of today corporate websites and generate a theoretical classification for today corporate webpages.

3. Proposed classification of corporate webpages

As websites have become more massive and evolved, stereotypical structures have emerged from their use, i.e., stable and recognizable content configurations have emerged for visitors. Although there is a wide diversity of webpages within a single website, it is possible to identify certain typologies that are repeated in many websites. Thus, based on

the observation of a significant number of today corporate websites at different points in their lifespan, a theoretical classification for corporate webpages has been generated.

Corporate: They present content related to the firm itself. These webpages have a long-life expectancy, i.e., they are likely to remain enabled throughout the lifespan of the website. The content of these pages undergoes few changes during their lifespan; e.g., corporate pages or data protection policy pages. Normally, changes in the content of this type of page will be related to deep changes, for example, corporate pages will change due to restructuring in the firm/institution.

Post: They present current content for firm's stakeholders, especially for the general public, e.g., a news page or a job offer. These webpages have a long-life expectancy, but their content is likely to grow through spontaneous comments from visitors, e.g., forum posts. The constant appearance of this type of page, added to the constant growth of its content due to user interactions, indicates that there is active communication between the firm and its stakeholders. If no such pages appear during the lifespan of the website, it will be a bad sign of the firm's digital evolution.

Service: They show the products and/or services offered by the firms. These webpages have a short-life expectancy, that is, there is a short time between their creation and their disappearance. In addition, the content of these web pages normally does not change during their lifespan. The constant appearance and disappearance of this type of pages, during the lifespan of a website, indicates that the firm has a regular business activity. An example of this type of pages are Amazon's product pages.

Catalogue: They conglomerate content linked to the Service and Post type pages. These webpages have a long-life expectancy but their content change constantly as new products or posts are added to the website, e.g., Amazon's homepage (https://www.amazon.es). The constant updating of content of this pages indicates that the firm behind the website has a regular business activity. If a page of this type does not undergo changes in content during its lifespan, it will be a bad sign of the firm's evolution.

4. Data and methods

4.1. Dataset description

To carry out the experiments, a dataset was defined and generated four times, once for each defined target class, as follows: A sample of 999 Spanish firms was extracted from the Bureau Van Dijk's SABI (Sistema de Análisis de Balances Ibéricos) database using stratified sampling with a uniform fixation to three strata: large firms, medium-sized firms and small firms according to the EU definition of SMEs. Then, a subset of 100 corporate websites was randomically extracted from the sample of Spanish firms. After that, 1005

webpages from the subset of corporate websites were visited and, after analyzing them structurally, assigned each one a label according to the classification defined above making sure that the dataset classes were balanced. Thus, 250 pages of Corporate type, 259 of Post type, 248 of Service type and 248 of Catalogue type were labeled. Then the Scrapy framework was used to automatically extract from the 1005 labeled webpages a series of relevant features, common at the source code level, that allowed to classify the webpages. Finally, the features extracted from each page were processed and converted them into categorical values.

Table 1. Summary of types of webpages defined, the changes that typically occur on them and the link of these changes to the firm's regular activity.

Type	Changes that occur	Link to business activity
Corporate	Specific changes in content or sudden appearance/disappearance	Structural changes in the firm or institution
Post	Content grows due to user interactions	There is active communication between the firm and its stakeholders
Service	Constant appearance and disappearance	The firm or institution has a regular economic activity
Catalogue	Constant changes in content	

4.2. Description of variables

As a result of the process described above, the dataset is made up of seven quantitative variables (HTML_size, Text_size, Tags_number, Links_number, Words_number and Images_number and Modified), and three sets of categorical variables (Files_extensions, HTML_tags and NL_words) that together form more than 37,000 independent variables that allowed to classify the webpages according to the typologies defined. The sets of variables Files_extensions, HTML_tags and NL_words contain the frequency of occurrence of the file extensions, HTML tags and Natural Language (NL) words of the scraped pages. It should also be noted that, the values of the HTML_size, Text_size, Tags_number, Links_number, Words_number and Images_number quantitative variables are transformed into quantiles (categorical variables) using the Pandas library.

4.3. Classification models used

From the literature review and after performing several preliminary tests with the classification models offered by Scikit-learn, for this study four of the most widely used classification models in the literature were used: Logistic Regression (aka logit, MaxEnt), Linear Support Vector Classification (Linear SVM), Neural network model Multi-layer

Perceptron classifier and AdaBoost classifier. The models were used with the default Scikit-learn configuration.

Table 2. Description of quantitative sets of categorical variables that make up the dataset.

Variable	**Description of quantitative variable**
Type	Target variable of the classification (Corporate, Service, Catalogue or Post).
HTML_size	Size in bytes of the body of the scraped pages.
Text_size	Size in bytes of the text content within the body of the scraped pages.
Tags_number	Number of HTML tags within the source code of the scraped pages.
Links_number	Number of links in HREF attribute of Anchor tags of the scraped pages.
Words_number	Number of natural language words within of the scraped documents.
Images_number	Number of images in SRC attribute of IMG tags of the scraped pages.
Modified	Last modification date of the of the scraped pages.
Variable	**Description of set of categorical variables**
Files_extensions	All file extensions present within the source code of the scraped pages.
HTML_tags	All HTML tags present within the source code of the scraped pages.
NL_words	All NL words present within the source code of the scraped pages.

5. Results and discussion

The performance of the classifiers was evaluated on a class-by-class basis, so the results are presented below organized according to the target classes.

Figure 1 shows AUC values of 5-fold Stratified cross-validation obtained for the Logistic Regression, Linear SVM, Multi-layer Perceptron and Ada_Boost classifiers for Post target class (which obtained the highest AUC values).

The highest AUC values were obtained when classifying the Post class using the Logistic Regression and Linear SVM classification models, both with values of 0.96 AUC and ± 0.01 of Standard Deviation. Logistic Regression was the model that presented the best performance in classifying each of the 4 target classes with respect to the other models used, also obtaining the highest AUC values when classifying the Corporate (0.90 ± 0.02), Service (0.90 ± 0.01) and Post Catalogue (0.84 ± 0.03) classes. The lowest AUC values were obtained when classifying the Catalogue class using the Multi-layer Perceptron (0.79

± 0.04) and AdaBoost (0.78 ± 0.04) models. Overall, the Catalogue class appears to be a difficult class for the models to classify.

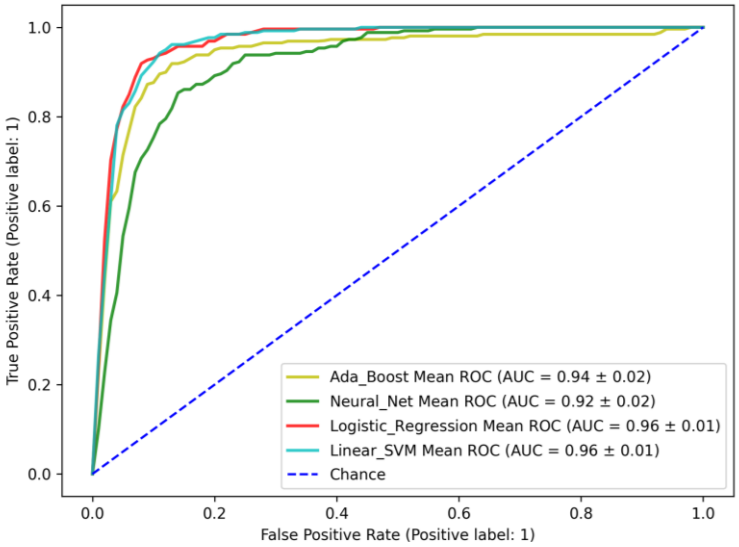

Figure 1. AUC values for all classification models for Post target class.

Table 3. AUC and Standard Deviation for all models, sorted by target class.

Target class	Logistic Regression	Linear SVM	ML Perceptron	AdaBoost
Catalogue	0.84 ± 0.03	0.83 ± 0.03	0.79 ± 0.04	0.78 ± 0.04
Corporate	0.90 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.84 ± 0.03
Post	0.96 ± 0.01	0.96 ± 0.01	0.94 ± 0.02	0.92 ± 0.02
Service	0.90 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.84 ± 0.03

5. Conclusions

This study, on the one hand, set out a theoretical classification of corporate webpages to associate their changes with the regular activity of firms. Through the observation of a significant number of today corporate websites at different points in their lifespan, four types of corporate webpages were defined: Catalogue, Corporate, Post and Service. On the other hand, the study also evaluated the possibility of an automatic categorization of corporate webpages using classification models. Logistic Regression presented the best performance in classifying each of the 4 target classes followed by Linear SVM. The best classified class was the Post class. Finally, it is necessary to emphasize that an automatic

categorization of today corporate webpages is important to firms to associate changes on their websites with their business activity and to evaluate its significance of these changes to the business. This could make a positive difference for a particular firm or organization.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Blazquez, D., and Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*, 99-113.

Calzarossa, M. C. and Tessera, D. (2018). Analysis and forecasting of web content dynamics. In 32nd *International Conference on Advanced In-formation Networking and Applications Workshops (WAINA)*, pages 12–17.

Cebi, S. (2013). Determining importance degrees of website design parameters based on interactions and types of websites. *Decision Support Systems*, 54(2), 1030 – 1043. 4.

Crestodina, A. (2017). What is the average website lifespan? 10 factors in website life expectancy.

Crosato, L., Domenech, J., and Liberati, C. (2021). Predicting sme's default: Are their websites informative? *Economics Letters*, 204, 109888

Han, S., Brodowsky, B., Gajda, P., Novikov, S., Bendersky, M., Najork, M., Dua, R., and Popescul, A. (2019). Predictive crawling for commercial web content. In *Proceedings of the 2019 World Wide Web Conference*, pages 627–637.

Llopis, J., Gonzalez, R., and Gasco, J. (2010). Web pages as a tool for a strategic description of the spanish largest firms. *Inf. Process. Manage*.,46, 320–330.

Llopis, J., Gonzalez, R., and Gasco, J. (2019). The evolution of web pages for a strategic description of large firms. *Economic Research-Ekonomska Is-tra ̌zivanja*,0(0), 1–21.

Santos, A., Pasini, B., and Freire, J. (2016). A first study on temporal dynamics of topics on the web. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 849–854.

Radinsky, K. and Bennett, P. N. (2013). Predicting content change on the web. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 415–424, New York, NY, USA. Association for Computing Machinery.

Zhou, E. and Sun, L. (2014). Evergreen or ephemeral: Predicting webpage longevity through relevancy features. 5.