An intelligent system for retrieving economic information from corporate websites

Josep Domenech, Bernardo de la Ossa, Ana Pont,
Jose A. Gil, Milagros Martinez and Alicia Rubio
Universitat Politecnica de Valencia.
Cami de Vera, s/n. 46022 Valencia (Spain)
jdomenech@upvnet.upv.es; berospe@doctor.upv.es;
{apont,jagil,mimar,alicia}@disca.upv.es

Abstract

The prompt availability of up-to-date economic indicators is crucial to monitor the economy and to steer the design of policies for promoting business innovation and raising firm competitiveness. Economic indicators usually suffer important lags since they are commonly obtained from official databases or from interviews to a sample of agents; thus limiting the representativeness and usefulness of the information. In a context in which the presence of companies in the World Wide Web is almost an obligation to succeed, corporate websites are connected, in some way, to the firm economic activity. On the basis of this relation, this paper proposes an intelligent system that analyzes corporate websites to produce web indicators related to the economic activity of the firms. This system has been successfully implemented and applied to infer company size characteristics from data gathered from corporate websites. Our results show that relatively large companies provide web content in a foreign language and use proprietary web servers.

1. Introduction

The design of programs for promoting business innovation and for raising the competitiveness of the industry requires prompt and accurate indicators to make the appropriate decisions, to allocate resources and to monitor and evaluate their outcomes. Unfortunately, some indicators currently used to this end are expensive to collect, suffer important lags between the gathering process and their analysis, and might not be appropriate and representative enough for a low-level granularity study. Consequently, the prompt availability of economic indicators is of crucial interest for forecasting and steering economic policies, especially in a context where changes continuously happen.

Meanwhile, the World Wide Web has grown at exponential rates. It is massively used now by firms as an open window to the world for showing their activities in a more global range, for capturing new customers or simply as a new way of business. Insofar as the business activities emerge on the Web, their economic activity can be detected and measured.

Therefore, our main goal is to use the data gathered from the WWW to produce a new source of information for the economic studies. To this end, we propose an intelligent system to automatically extract a set of indicators from a given corporate web site. Then, these website indicators are processed in order to produce the economic indicators required for economic research. This papers validates the proposed model by presenting a proof of concept that evidences correlations between some business characteristics (particularly, firm size) and some corporate website characteristics (content and metadata). The focus on firm size of this paper is justified by the fact that it represents one of the main determinants of business innovation efforts [1]. However, this model can be applied to find web indicators for any other aspect of business activity.

2. Background and related work

The first approach to systematically use the web as a source of economic information is the webometrics. These indicators rely on analyzing web page links to compute measures similar to some widespread bibliometric indicators [2, 3]. The main drawback of this ap-

[1]This work has been partially supported by Spanish Ministry of Science and Innovation under grant TIN2009-08201, Generalitat Valenciana under grant GV/2011/002 and Universitat Politecnica de Valencia under grant PAID-06-10/2424.

proach is that the large heterogeneity found in the web hinders the reliability of such indicators [3]. However, this heterogeneity has not been a handicap when the analysis is restricted to the scientific production of universities or nations [4, 5, 6]. Other examples of successful applications include comparisons of scientific disciplines [6] and the collection of indicators for specific scientific fields [7]. However, these approaches might result inadequate for a company-level granularity because they are closely dependent on the science context.

The reports generated by Google Trends (GT) are another interesting approach to obtain economic indicators from the web. This tool provides up-to-date reports on the volume of web search queries with some specific text. First introduced as economic indicator by Choi and Varian [8], GT data were successfully used for improving some prediction models on micro and macro indicators: claims for unemployment benefits, retail sales, automotive sales, home sales and travels. Since then, GT has been applied to a number of situations [9, 10, 11]. These research works show that GT can provide useful hints on the economic activity. However, its ability for characterizing the supply-side of the market (i.e., what firms offer) is limited since it only provides data about the demand-side of the market (i.e., what users demand).

There are also some attempts that use web information to analyze some firm characteristics [12, 13]. Unfortunately, research at firm level has been only performed by following a non-automated analysis, thus limiting the scale of applicability and the promptness of the methodology. Although there are some other works that analyze corporate web sites at a larger scale [14, 15], they require a clear strategy for manual coding to avoid biases in the retrieval of the information. Doing this huge task in a non-automated way can only be carried out by reducing the analysis of the corporate web site to a single section [14] or even to just the home page [15].

Consequently, this research is relevant to address current economic and innovation policy questions in a general way. This fact motivated us to develop an intelligent system to automatically find web indicators for firm economic activities.

3. Intelligent system for retrieving economic information

This section proposes a model for an ingelligent system that automatically produces economic indicators from the analysis of corporate websites. Then, an implementation as a proof of concept is presented.

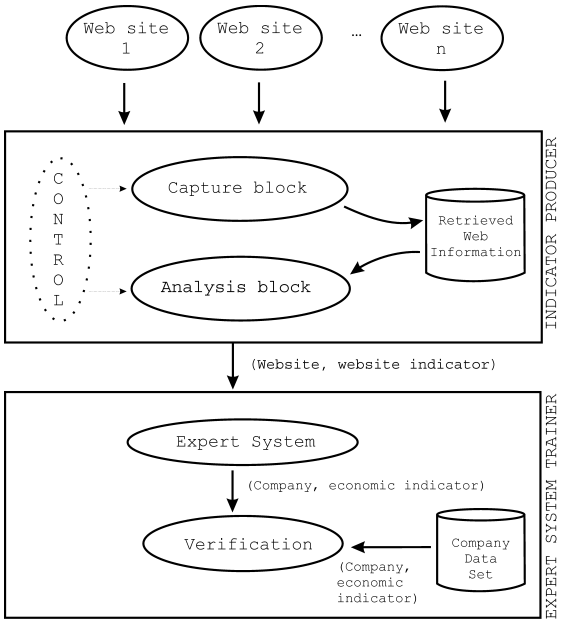

Figure 1. Model for an intelligent system for retrieving economic information.

3.1. Model

The system for retrieving economic indicators has been divided into two main parts, each one performing a specific task: the production of website indicators and the production of valid economic indicators. A representation of this model is depicted in Figure 1.

Indicator producer The objective of this part of the system is to obtain web indicators that could be potentially related to business activities. To do so, we defined three main functional blocks: the website capture block, the analysis block and the control block.

The capture block (see Fig. 1) represents a robot that downloads (part of) the corporate websites whose URI were given as input and stores them in a local copy to allow future analysis. When capturing websites, both content and metadata (i.e., HTTP headers) are stored since both can provide useful insights on the activity of the company.

The goal of the analysis block is to produce indicators related to corporate websites by analyzing the local copies downloaded before. This block is composed of several sub-blocks, each one examining a given aspect of the website (e.g., language, server technology) to produce related indicators. This allows the use of different analysis strategies depending on the indicator to be produced.

Finally, the control block is in charge of managing the data flow between the capture and the analysis blocks. This part of the system is specially relevant for parallelizing the different tasks that are performed for each corporate website.

Expert system trainer The second part of the model pursues to produce valid business activity indicators from the website indicators generated before. To do so, our model includes two functional blocks: an expert system and a validation block

The expert system employs a knowledge base to infer some firm economic characteristics (i.e, economic indicators) from a given set of corporate website indicators. The training process of the expert system is performed by evolving the knowledge base.

The validation block evolves the knowledge base by comparing the output of the expert system to the data that other economic databases provide. Once the rules used by the expert system are validated, it is possible to obtain economic indicators for those companies not included in databases or for those whose economic information is obsolete.

3.2. Proof of concept

With the purpose of validating the feasibility of the proposed intelligent system, we have implemented a proof of concept to find the relations between corporate web characteristics and firm sizes. To this end, some web indicators that could be related to this business characteristic are computed and then compared to the company size registered in a business database.

The capture block of the system was implemented as a modified version of HTTrack [16]. This robot captures web content and HTTP headers from a given website by parsing recursively the links found in the provided URI, including most references included in javascript code and Flash objects. To avoid overloading web servers, limitations on time and bandwidth usage when downloading a site were defined.

The analysis block was implemented as a set of three scripts that were sequentially run. Each script computed one or more related website indicators. The first script produces an indicator related to the web server software, classifying it as open source (i.e., Apache, lighttpd or Nginx) or proprietary (i.e., Microsoft IIS or Oracle). This is done by checking the `Server` header in the HTTP response of the corporate home page.

The second script generates an indicator about the top-level domain of the corporate website. Its implementation is as simple as parsing the URI of the corporate home page.

The third script produces indicators related to the language in which the corporate web site is written. To do so, a spell checker is used on each HTML to find the most likely language of the text content. The generated website indicators are the number of resources written in each language.

The control block was implemented as a queue in which several tasks are processed in parallel. Although processing each corporate web site has two sequential tasks (capture and analysis), parallelization is possible because the indicators for each website are independent. Hence, the control block parallelized the process of different websites so that all cores in our servers were busy.

Once the website indicators are generated, they can be compared to the economic indicators found in business databases to design and validate the expert system.

4. Evaluation of web indicators

This section presents some web indicators for firm size computed after implementing the system proposed above. To do so, we firstly describe how the dataset was built, putting economic activity together with corporate website information. Then, these data are analyzed to find connections between both company perspectives.

4.1. Company economic activity dataset

To build a representative dataset of the economic activity carried out by real companies, a selection of firms were randomly retrieved from SABI, which is a database widely used in economic research. It contains accounting information for about 1.6 million companies established in Spain and Portugal.

Our sample includes 10,000 firms headquartered in the province of Valencia, in Spain. This sample is highly representative of the local economy as it represents about one fifth of the companies based in this province available in SABI. Although this database offers a wide variety of company information, we are only interested in three fields: URL of the corporate website, firm's last year revenue and headcount. From these 10,000 firms, only those with corporate website were considered for further study. This reduced the sample to 1,033 companies.

As in many other regions in the world, Valencia economic activity is mainly dominated by small and medium enterprises (SME), being large firms very reduced in number. It is well known that large companies have special characteristics that make them different from SMEs. For this reason, all firms with more than 1,000 employees or with yearly revenue greater

Table 1. Main characteristics on the corporate website of the analyzed firms

	Mean	Std dev
HTMLs	538.97	1 450.80
Images	390.09	1 531.71
Javascripts	5.57	12.71
Proprietary webserver	.27	.44
ES version	.93	.26
EN version	.65	.48
FR version	.19	.39

than 100 million EUR were excluded from the sample. This filtered out eight companies.

4.2. Corporate website dataset

The dataset with economic information was extended to include some indicators extracted from each firm corporate website. To do so, all the corporate websites were captured and processed according to the implementation of the model proposed above. Then, web indicators were computed.

During the process of capturing websites we found that many companies had their website under construction or they provided so little content that make it impossible to infer some economic information from them. For this reason, a threshold on website size was considered for being able to infer economic information from a corporate website. That is, firms whose website had fewer than four HTML resources were considered as not having corporate website.

Table 1 summarizes the main characteristics of the analyzed corporate websites. The distribution of website size, measured in number of HTMLs, is positive skewed (heavy-tailed) since the average is 538.97 HTML resources, with a median of 55. Similarly, the distribution of images and javascript resources is also heavy-tailed. Websites had on average 390 images (median = 81) and 5 javascript (median = 2) resources.

Corporate websites were hosted predominantly on open source web servers (mainly Apache). Websites hosted in a proprietary web server (mainly IIS and Oracle) represented 27% of firms. Most websites (93%) had content written in Spanish, while 65% of them had some content in English. French is the third language, being present in 19% of websites.

4.3. Data analysis

This subsection relates firm economic data with some indicators extracted from their corporate website. For the sake of clarity, web indicators were grouped into two types: server-based indicators and language-based indicators.

4.3.1. Server-based indicators.

Web server software As described above, some companies host their website on open source servers, namely Apache; while some others prefer to use a proprietary web server, namely Microsoft Internet Information Services.

Figure 2 compares firm size according to what the technology of the web server software is. As one can observe, the size of companies whose website is hosted in a proprietary server is, on average, larger than those using an open source web server. This difference can be appreciated both in mean headcount (28 vs 40 employees) and in mean yearly revenue (4,257k vs 7,599k EUR).

Top-level domain The top-level domain is many times an essential part of a company's branding. For this reason, it is expected that it is connected to the firm economic activity.

The most common top-level domain across the considered companies is `.com`, which accounts for 65% of them. The second most used TLD is `.es`, employed by 31% of firms. Remaining domains are used only by 4% of companies.

Figure 3 shows the relation between the corporate website TLD and the firm size. Results on Fig. 3(b) evidence that firms with a website under `.com` domain have significantly less revenue than those under `.es` (4,306k vs 7,228k EUR). However, this difference is not significant when comparing headcount (see Fig. 3(a)). Although this relation is the opposite that one could expect, it can be explained by the huge restrictions and strict requirements that applied for registering a domain under `.es` until 2005. This made many businesses decide to register their domain under `.com` instead, particularly those with fewer resources to meet all the requirements.

Other TLDs are less common across the analyzed websites. Firms under these domains are more heterogeneous, making it large the confidence interval for other TLDs (as Fig. 3 shows). Therefore, no significant differences are found for firms under domains other than `.com` and `.es`.

4.3.2. Language-based indicators. The language in which a corporate website is written gives information on the intended customers of the company, so this should be connected to its economic activity. In this

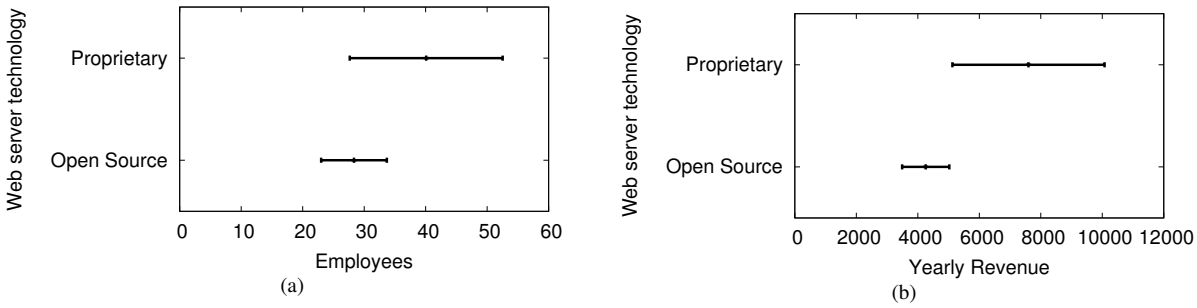

Figure 2. Ninety-five percent confidence interval for the average firm size classified by the web server software.

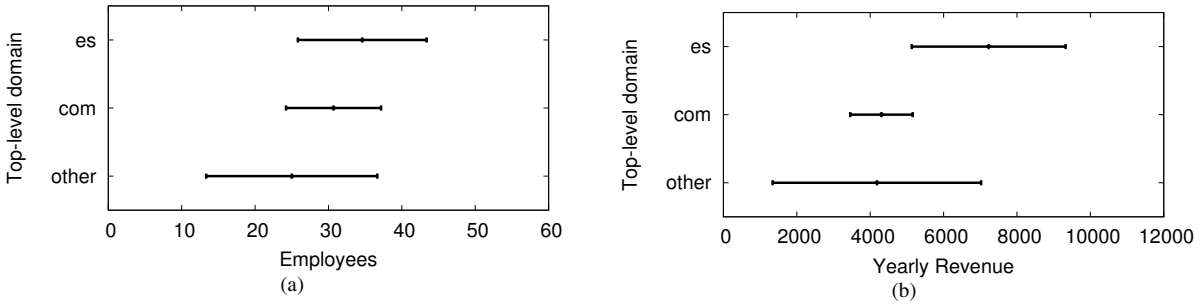

Figure 3. Ninety-five percent confidence interval for the average firm size classified by the top-level domain of the corporate website.

context, we considered as indicator that a website have some content written in English or in French. These languages were considered because they are the two most commonly used languages in the analyzed corporate websites, apart from Spanish.

English Figures 4(a) and 4(b) compare company average size depending on whether corporate website have some content written in English. As expected, firms whose website has some English content hire consistently more employees (36) and have larger incomes (6,067k EUR) than those which do not (23 employees and 3,595k EUR).

French Compared to English, French is less often used in corporate websites. This fact increases the differences in firm size between those corporate websites with some content in French and those without. Figure 4(c) shows this from the number of employees perspective. Firms with some French content in their website hire on average 54 workers, while those without French content have a mean of 26 employees. Similarly, Fig. 4(d) illustrates the differences in yearly revenue (9,853k vs 4,100k EUR).

5. Conclusions

Links between the economic activity and what is going on in the web are becoming stronger in the last years. In this context, this paper proposed a model for studying these links from a still unexplored point of view. That is, how firm economic activity is related to its corporate website information.

The proposed system for retrieving economic information from corporate websites has been implemented and successfully applied for finding some interesting relations between the economic activity of a sample of companies based in Valencia (Spain) and their corporate websites. Web economic indicators for this sample were related to the web server used and to the language employed for writing the content. These indicators allow us to monitor and to infer economic characteristics of Valencian firms by only analyzing their website.

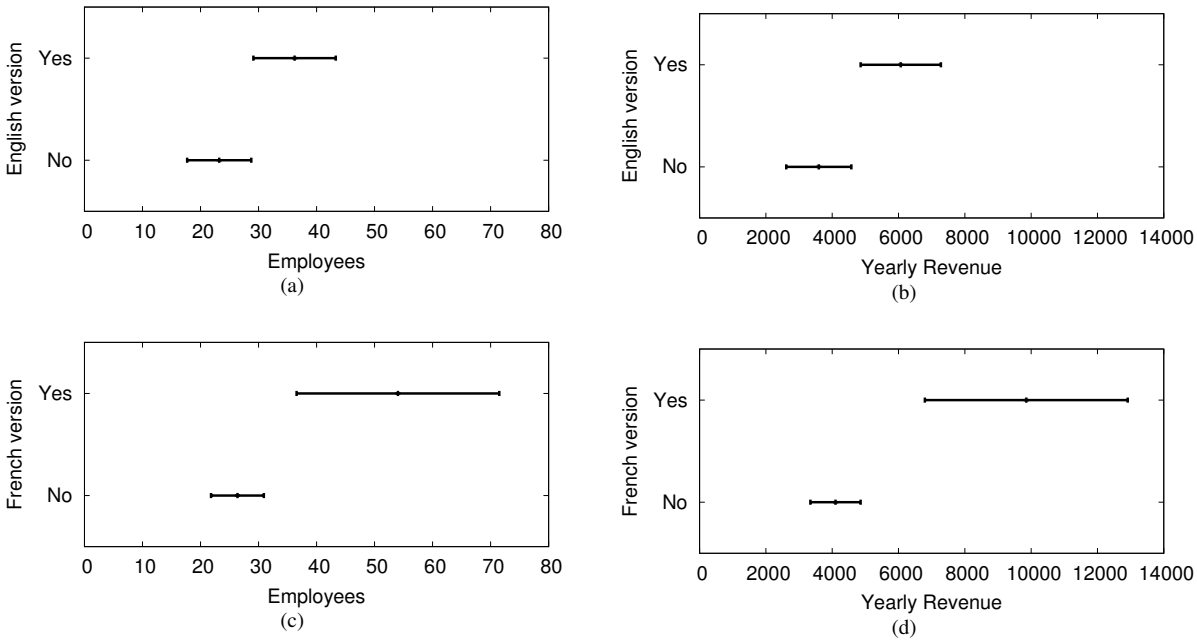

Figure 4. Ninety-five percent confidence interval for the average firm size depending on whether there is some content written in a foreign language (English or French).

References

[1] Z. J. Acs and D. B. Audretsch, "Innovation, market structure, and firm size," *The Review of Economics and Statistics*, vol. 69, p. 567–574, 1987.

[2] P. Ingwersen, "The calculation of web impact factors," *Journal of Documentation*, vol. 54, pp. 236 – 243, 1998.

[3] A. G. Smith, "A tale of two web spaces: comparing sites using web impact factors," *Journal of Documentation*, vol. 55, pp. 577–592, 1999.

[4] M. Thelwall and G. Harries, "Do the web sites of higher rated scholars have significantly more online impact?" *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 149 – 159, 2003.

[5] D. Wilkinson, G. Harries, M. Thelwall, and L. Price, "Motivations for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication," *Journal of Information Science*, vol. 29, pp. 49 – 56, 02/2003 2003.

[6] G. Heimeriks, P. van den Besselaar, and K. Frenken, "Digital disciplinary differences: An analysis of computer-mediated science and 'mode 2' knowledge production," *Research Policy*, vol. 37, 2008.

[7] M. Thelwall, A. Klitkou, A. Verbeek, D. Stuart, and C. Vincent, "Policy-relevant webometrics for individual scientific fields," *Journal of the American Society for Information Science and Technology*, vol. 61, 2010.

[8] H. Choi and H. Varian, "Predicting initial claims for unemployment benefits," http://research.google.com/archive/papers/initialclaimsUS.pdf, 2009.

[9] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 17 486 – 17 490, 2010.

[10] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. google trends," *Journal of Forecasting*, vol. 30, pp. 565 – 578, 09/2011 2011.

[11] M. Dzielinski, "Measuring economic uncertainty and its impact on the stock market," *Finance Research Letters*, 2011.

[12] M. Overbeeke and W. E. Snizek, "Web sites and corporate culture: A research note," *Business & Society*, vol. 44, pp. 346 – 356, 2005.

[13] J. Llopis, R. Gonzalez, and J. Gasco, "Web pages as a tool for a strategic description of the spanish largest firms," *Information Processing & Management*, vol. 46, pp. 320 – 330, 2010.

[14] P. J. Kim, S. and E. K. Wertz, "Expectation gaps between stakeholders and web-based corporate public relations efforts: Focusing on fortune 500 corporate web sites," *Journal of Information Science*, vol. 36, 2010.

[15] M. S. J. Hwang, J. S. and G. Lee, "Corporate web sites as advertising: An analysis of function, audience and message strategy," *Journal of Interactive Advertising*, vol. 3, pp. 10–23, 2003.

[16] "Httrack," http://www.httrack.com.