2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8333

Validation of innovation indicators from companies' websites

Héroux-Vaillancourt, Mikaël and Beaudry, Catherine
Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Canada

Abstract

In this exploratory study, we use a web mining technique to source data in order to create innovation indicators of Canadian nanotechnology and advanced materials firms. 79 websites were extracted and analysed based on keywords related to the concepts of R&D and intellectual property. To understand what our web mining indicators actually measure, we compare them with those from a classic questionnaire-based survey. Formative indices from the surveys variables were built to better represent all the possibilities resulting from the web mining indicators. A MTMM matrix lead us to conclude that the formative indices are a good representation of the web mining indicators. As a consequence, the data extracted via our web mining technique can be used as proxies for the relative importance of R&D and the importance of IP, which would have previously only been measured using conventional methods such as government administrative data or questionnaire-based surveys.

Keywords: *Multi-Traits Multi-Method, construct validity, Web-mining, innovation measurement, nanotechnology and advanced materials*

1. Introduction

The majority of companies working in highly technological areas have an up-to-date website to inform potential customers, potential business partners and investors about their activities. Although the online information is made available by the companies themselves, suggesting the possibility of a strong desirability bias, this source of information can be suitable for the study of technological innovation (Domenech et al., 2015; Gök, Waterworth, & Shapira, 2015). The information obtained is as rich as it is diversified, including products, services, business models, R&D activities, etc. Would it be possible to extract this information and convert it into useful data to research? Moreover, is the information available on the various business websites is reliable and it is sufficient to give a good picture of some characteristics of companies? In other words, can the content of a commercial website be used to identify different innovation characteristics of a business?

When visiting a company's website, recurring themes that emerge from groupings of synonymous words may be noticed. These themes may actually describe factors appearing to be particularly important to the business. This study aims to validate whether the importance of factors emerging from a website is a good representation of the real importance a company actually gives to these factors.

In this study, we analysed and compared 2 sets of measures of innovation of nanotechnology and advanced materials in Canada stemming from two different data gathering techniques: Web scraping/mining and questionnaire-based survey. Comparisons between results from both methods were obtained via correlations. To ensure a convergent and discriminant validation of our results, we performed a Multi-Traits Multi-Method (MTMM) technique.

2. Methodology

2.1 Data acquisition

We started by conducting a classic questionnaire-based survey of which the core is based on the Oslo Manual (OECD & Eurostat, 2005) and explored the following themes: innovation, commercialisation, collaboration and intellectual property. We contacted 2971 Canadian high technological firms from which, 89 subjects were eligible and accepted to participate to our study. In order to build the two factors described above, R&D and intellectual property, we identified all the relevant questions from the questionnaire-based survey and transformed the answers to these questions into different types of variables. In the end, we generated a total of 9 variables pertinent to R&D, and 2 variables measuring intellectual property.

Then, we treated the websites of these 89 companies with the process described in Figure 1, from which we successfully extracted the website content of 79 companies (88%). The keywords related to R&D were selected from the literature (Gök et al., 2015), while the keywords related to intellectual property were identified from our own investigations of the literature. The most relevant keywords of any paper are generally listed on its first page, particularly under the abstract and served as a basis for the list of keywords used for the construction of our factors.

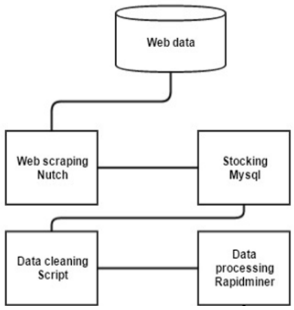

Figure 1. Web mining process

Clustering using keyword frequency analysis with a text mining software enabled us to count the number of occurrences of each keyword for each factor. We transformed these clusters of occurrences into 2 continuous variables. Because the 79 companies are different in structure and size, and therefore present different amounts of information in their websites, we standardized each variable by dividing all occurrences by the total number of words appearing in their website and multiplied the resulting value by 1,000. For each continuous variable, we calculated the Kurtosis and Skewness measures to determine whether they were following a normal distribution. None of our variables followed a normal distribution and were thus transformed by applying a natural logarithm (LN) or an inverse function (INV).

2.2 Construction of formative indices and validation construct

Given the vast field of words used to construct the web variables, treating each questionnaire-based variable individually may not be appropriate. To illustrate the large lexical field of possible words related to the factors studied, it is conceptually sound to build one single measure, one formative index with all the questions related to R&D and IP. Moreover, Principal Component Analysis (PCA) was performed on all the items related to R&D and to IP and did not produce any significant K-M-O and Cronbach's Alpha

measures. This situation suggests the use of a formative index comprising several sub-elements explaining our R&D and IP factors.

Partial Least Square (PLS) regressions were estimated to determine whether it is possible to create valid formative indices for these two factors, R&D and IP. In order to use PLS regressions, the methodology requires only the use of complete data sets (Nelson, Taylor, & MacGregor, 1996). Non-response is usually treated either by weight adjustment, i.e. delete incomplete data entry and weigh remaining respondents to compensate for the deletion, or by imputation, i.e. adding artificial values based on average by classes and editing methods (Särndal, Swensson, & Wretman, 1992) to replace the missing values (Haziza & Beaumont, 2007). Since our sample size for IP is already low for one of the items (39 for the number of patents), we could not afford to treat the missing data with a weight adjustment. Thus, we replaced the missing data with their imputation class based on control variables. We sorted by sector, then by the number of employees and then by revenue. Depending on the situation, we used the mean of the class or the most conservative nearest-neighbour, a method commonly used in the literature (Haziza & Beaumont, 2007; Little, 1986; Thomsen, 1973).

Since not all the items shared the same scale, we transformed each variable into a Z-score. PLS regressions were then estimated using the WarpPLS 5.0 software with the following settings: MODEL B BASIC Warp3 Stable 3 and MODEL B BASIC Linear Stable 3. The two different settings produced the same conclusions. The details of the construct comparing the Web mining technique and the questionnaire-based survey are shown in Table 1.

All weights are significant (*p-value* < 0.01), indicator weight-loading signs are all positive, variance inflation factors (VIF) are all very low (<1.5) and the Effect sizes (ES) are all greater than 0.02. All the criteria are met to indicate that the indices generated are valid (Cenfetelli & Bassellier, 2009; Cohen, 1988; Diamantopoulos, 1999; Diamantopoulos & Siguaw, 2006; Diamantopoulos & Winklhofer, 2001; Petter, Straub, & Rai, 2007). For each factor, the sum of each weighted variable generated both indicator RD_INDEX and IP_INDEX.

Table 1. The validation construct

Concepts	Web Mining	Questionnaire	
R&D	LN_WEB _RD (Continuous, normal)	RD_INDEX (Continue, normal)	Z_NUMBER_RD (Continue, normal) Z_INT_INTERN_INFO_RD (Continue, normal) Z_INT_EXT_INFO_RD (Continue, normal) Z_INT_CONT_RD (Continue, normal) Z_INT_PROV_RD (Continue, normal) Z_TIME_RD (Continue, normal) Z_PROP_RD (Continue, normal)
Intellectual property	INV_WEB _IP (Continuous, normal)	IP_INDEX (Continue, normal)	Z_SUM_IP (Continuous, normal) Z_NB_PATENT (Continuous, normal)

3. MTMM Analysis results

First introduced by Campbell & Fiske (1959), the Multi-Trait Multi-Method (MTMM) allows for the convergent and discriminant validation of a construct where a set of t traits (interchangeable with factors in our case) are measured with m different methods.

This MTMM matrix includes the two data mining indicators along with RD_INDEX and IP_INDEX (see Table 2). The reliability diagonal will be neglected in our analysis since the measures are made with single items from the web method and with formative indices from our questionnaire. The monotrait-heteromethod diagonal shows high and significant

correlations for R&D ($r = 0.419$; *p-value* < 0.01) and for IP ($r = 0.52$; *p-value* < 0.01), which hints at strong convergent validity. The heterotrait-monomethod value is low and not significant for the Web mining method ($r = 0.182$; *p-value* > 0.05) but the questionnaire-based survey method value is high and significant ($r = 0.32$; *p-value* < 0.01). However, the monotrait-heteromethod value is much higher than the heterotrait-monomethod values (< 0.419 > 0.182 and 0.52 > 0.32 for R&D and IP respectively), which indicates good nomological validity and that there are no mono-method biases. The first heterotrait-heteromethod value is low and not significant ($r = -0.17$; *p-value* > 0.05) while the other is moderate and significant ($r = 0.294$; *p-value* < 0.05). However, and more importantly, the correlations are lower than the corresponding values found in the validity diagonal, which shows good discriminant validity. All the conditions are satisfied under the original guidelines proposed by Campbell and Fiske (1959), and therefore, no risk of potential biases is induced within the methods, the traits or a combination of both. The results based on this methodology suggests that our web mining indicators reflect the importance given to innovation factors such as R&D and Intellectual property.

Table 2. MTMM matrix for RD_INDEX and IP_INDEX

	Traits	**Method 1 (Web)**		**Method 2 (Questionnaire)**	
		RD	IP	RD (RD_INDEX)	RD (IP_INDEX)
Method 1 (Web)	RD	N/A[a]			
	IP	-0.182	N/A[a]		
Method 2 (Questionnaire)	RD (RD_INDEX)	0.419**	0.294*	N/A[a]	
	IP (IP_INDEX)	-0.17	0.52**	0.32**	N/A[a]

Note: All traits from Method 1 are measured by single items, there are no reliability statistic that can be calculated. All traits from Method 2 are measured by a formative index and thus, reliability statistic is irrelevent.

$p < .05$.

$p < .01$.

4. Limitations and future research

Obviously, more data would allow our research to be more robust. Another limitation of our methodology is the fact that we did not take into account the context around our keywords, possibly leading to multiple false positives. The addition of machine learning techniques, such as Recurrent Neural Network or Natural Language Processing or Bag-of-Words model, is a promising avenue to improve the level of precision by adding to the method the necessary context around keywords. Moreover, we started with theoretical factors for the conceptual framework, then identified the keywords related to these factors, and finally mined the website for these specific keywords. An interesting alternative would be to do this the other way around, i.e., to start with the website content and to identify the factors that can be naturally found via unsupervised machine learning algorithms. The term frequency inverse document frequency technique (TF-IDF) could be used to provide insight into the importance of keywords relative to the rest of a document.

In a nutshell, our methodology seems it can be used as a valid approach to provide data for future innovation and technology management studies for the relative importance given to a factor such as R&D and IP, and to test the validity of the measures thus created. In most questionnaire-based surveys, that information is gathered using 1 to 7 Likert scale questions. If the goal of a study is to determine the degree of importance of core factors such as R&D or IP for a firm, the use of Web mining indicators is reasonable. However, if the goal is to gather more specific information, such as the precise actions undertaken by a firm, these web mining indicators may lack the necessary context to behave as expected. The importance given by a company to certain types of activities represents which activities that are supported and encouraged by the culture of the company (Herzog, 2011). Therefore, it is possible that our methodology suggest a novel way to measure quantitavely innovation culture.

Of course, company websites are willingly structured in a cooperative and agreeable manner toward whomever is seeking information concerning products, services, activities, and so on. The self-reporting bias induced by this methodology is inevitable. However, it is important to note that questionnaire-based surveys and most national official public directory are all subject to self-reporting biases as well. Fortunately, the bias induced by the web mining technique is as much a quality as a flaw, in that it provides insight on how the company wants to be perceived. Indeed, companies write on their websites about what they care about, what is important for them and who they are as a company. This qualitative information represents the essence of the company. Future research is needed to determine whether this qualitative information could be used as a proxy to understand a company's culture for instance. Furthermore, future research will be performed to assess how these indicators can be used in actual regressions to understand innovation patterns. It will be especially interesting to assess whether these web indicators tend to be substitutes or

complements to the traditional measures use in innovation management studies. This will be performed in the coming months with a sample of 1700 companies.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly*, *33*(4), 689–707.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Domenech, J., Rizov, M., and Vecchi, M. (2015). *The Impact of Companies' Websites on Competitiveness and Productivity Performance* (Conference Paper: First International Conference on Advanced Research Methods and Analytics).

Diamantopoulos, A. (1999). Viewpoint – Export performance measurement: reflective versus formative indicators. *International Marketing Review*, *16*(6), 444–457. https://doi.org/10.1108/02651339910300422

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management*, *17*(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, *38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Haziza, D., & Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, *75*(1), 25–43.

Herzog, P. (2011). Innovation culture. In *Open and Closed Innovation* (pp. 59–82). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-8349-6165-5_3

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review / Revue Internationale de Statistique*, *54*(2), 139–157. https://doi.org/10.2307/1403140

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

OECD, & Eurostat. (2005). *Oslo Manual*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264013100-en

Petter, S., Straub, D., & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, *31*(4), 623–656.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling Springer. *New York*.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, *4*, 278–283.