Could the organisation's websites be a valid data source for research? - An analysis of the complementary nature between web-based indicators and traditional indicators in innovation studies

1. Background

Most tech companies developing new innovations have a website to inform potential customers, potential business partners and investors about their activities. The possibility to use this information for research in innovation management is enticing. Indeed, the information is abundant and available for free. However, this gold mine of information represents many methodological challenges due to the unstructured, unstandardized, decentralised nature of websites. Furthermore, since the information is made available by the companies themselves, firm representatives can pick and choose the information they want to display suggesting a strong self-reporting bias. Indeed, companies can decide to not disclose relevant information because they can judge it to be sensitive, or simply irrelevant for the context of their website. Despite these limitations, some web content analyses were performed in innovation studies.

Gök et al., (2015) proposed a web content analysis based on keywords frequency analysis to assess R&D activities of UK-based green goods SMEs. The results showed that the web-based indicator did not correlate significantly with the non-web-based indicator, which hints that these two indicators did not reflect the same concept. This result raised two important questions: Could web-based indicators be complementary to non-web-based indicators? If this is the case, what do they actually reflect? In this exploratory study, we aim to provide some insights regarding these two questions with the use of a similar web content analysis method in order to analyse the innovation of technological firms.

2. Methods

We perform a classic regression study that explains firms' innovation performance with well-known innovation determinants such as R&D, intellectual property (IP), collaboration and external financing. From the different surveys, the innovation performance indicators that were included in our analysis are the degree of innovativeness, the number of innovations, the time to market and the percentage of revenue coming from latest innovation.

In order to build these indicators, we used a dataset of 1,570 technological companies that answered innovation intermediaries impact surveys collected by a consultancy firm that specialises in the evaluation of innovation intermediaries. We extracted the text content of these companies' websites and created indicators using keyword frequency analysis with a text mining software that counts the occurrences of each keyword for each factor. Because the websites 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.

Models including questionnaire-based indicators and web-based indicators were included to test the complementary nature of both indicator types. In our regression models, the first block is allocated to control variables, while the next block comprised of all classic questionnaire-based

variables. This allows us to see the model baseline in terms of explanatory power (adjusted R-squared). Each subsequent block input a new web-based variable to the model. This allows us to see the progression in terms of the explanatory power and to better understand the impact of the web-based variables on the different models.

3. Preliminary results

We scraped and analysed 965 firms (61% of the technological firms) and are currently processing the remaining firms. Our first observation suggest that for each concept, each pair of measures taken from both methodologies does not have a high and significant correlation, hence that the web-based indicators effectively have the potential to be complementary rather than substitutable, which reproduces Gök et al. (2015) findings. We then tested all the possible combinations of independent variables that did not have any high correlation (below 0.4). We estimated many significant regression models including our web-based indicators. For the purpose of this paper, we selected the best model obtained explaining the influence of the degree of innovation perceived by the firm leaders. Our results show that the firms that invest the biggest proportion of their revenue in R&D and the firms that receive the most investment consider themselves more innovative. Second, we find that the firms mentioning IP-related concepts on their website increase the degree of innovation perceived by the firm leaders. Third, mentions about R&D-related concepts follow a U-shape relationship, meaning that the companies that either did not mention R&D at all or those that did mention R&D the most yield the best performance. Fourth, we tested the interaction between our web-based variables, but these did not affect the models significantly. Finally, our web-based indicators improved this model significantly from an adjusted R-squared of 0.229 they contributed to increasing it to 0.281.

4. Discussion

Significant models with the four factors measured with the web-based indicators were found, which means they offered additional, complementary information and contributed to our models in explaining innovation performance.

Although this study did not assess quantitatively the nature of these web-based indicators, we offer a conceptual proposition of what they could actually mean. When visiting a company's website, recurring themes that emerge from groupings of synonymous words may actually describe factors appearing to be particularly important to the business. Firms self-report the content on their website by revealing to the world who they are and what is important to them. This subjective information communicated from the firm to the world is induced by cultural artefacts. In future studies, researchers could use web content analysis to better understand a firm culture.

It is important to mention that the keyword frequency clustering analysis method itself shows some limits. Obviously, it is not because a word is mentioned on a website that it is used within an innovation-related mindset. This can lead to multiple false positives. Other machine learning techniques, such as Natural Language Processing could add the proper context into these types analysis and it will be tested in the upcoming months.

5. Reference

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