Mining corporate web sites for innovation measures: Literature review and framework

R. Sandra Schilo schillo@telfer.uottawa.ca Telfer School of Management, University of Ottawa

Mohammadreza Seifollahim seif087@uottawa.ca Telfer School of Management, University of Ottawa

Diana Inkpen diana.inkpen@uottawa.ca School of Electrical Engineering and Computer Science, University of Ottawa

Introduction

The measurement of innovation among companies is of broad interest to corporate managers and policy developers alike, yet measuring innovation reliably remains a challenge. Most research to date has made use of company surveys, which results in limitations relating to response and recollection biases, sample size limitations due high data collection efforts, and the type of information that can be gathered. Researchers are increasingly exploring web-based data collection methods to counter some of these limitations and derive insights from freely accessible data that is frequently updated by companies in the course of their operations. This paper provides a review of existing research using corporate web sites to identify innovation-related indicators and proposes a framework for future work.

Review of Measures and Methods

Compared to more traditional methods of measuring corporate innovation including bibliometrics and patent analysis, there are still very few publications applying web-based measures to the context of corporate innovation and R&D. Previous research has typically utilized either web content analysis – mostly in the form of keyword-based analyses – or web structure analyses, but very few studies combine web structure and web content analysis.

The most common methods in content analyses are based on keyword analyses. Some studies employ simple reporting of frequencies and normalized frequencies. Most studies, however, follow up keyword analyses with factors analyses (Libaers et al., 2010), principal component analyses and/or cluster analyses (Arora et al., 2013; di Tollo et al., 2015). Co-occurrence of keywords is beginning to be employed, either through logical combination of simple keyword analyses (di Tollo et al., 2015; Gök et al., 2015) or more advanced considerations, such as on-screen visibility (Jianhua et al., 2016) and co-word networks and emotion analysis using corporate annual reports (Garechana et al., 2017). The use of internet archives is quite common, e.g. to identify corporate status as active or inactive (Blazquez and Domenech, 2017a) or to identify changes in the product and service offering of companies and their financing (Youtie et al., 2012).

More complex, theoretically aligned studies have combined several web-based measures to identify commercialization models of companies (Libaers et al., 2010), business strategy(Arora et al., 2015; Youtie et al., 2012) , and most recently export orientation (Blazquez and Domenech, 2017b). While not all of these studies made an explicit link to corporate innovation, these concepts have been considered in the innovation literature, and future innovation studies should consider these methodologies.

Innovation Dimensions

The review of the literature on web-based innovation measures suggests that the literature is not yet at the stage of theory-driven web scraping (Landers et al., 2016). We thus suggest in this paper that web-based innovation measures should aim to reflect established innovation measures. The most widely applied innovation measurement framework in the traditional innovation literature are the various versions of the OECD's Oslo Manual (OECD, 1997 and onwards). This manual contains a range of questions relating to companies' innovation strategies and innovation activities, including activities undertaken with regards to goods and service innovation, process innovation, organizational innovation and marketing innovation.

Innovation strategy

Surveys following the Oslo Manual typically contain at least one, often more questions concerning corporate innovation strategies. As Libaers et al. (2010) note, most web sites include sections on the company's mission or strategy. Keyword analyses have allowed researchers (Libaers et al., 2010) to identify commercialization strategies among innovative companies, and our framework proposes to focus specifically on innovation strategies as captured in the Oslo Manual.

Innovation activities

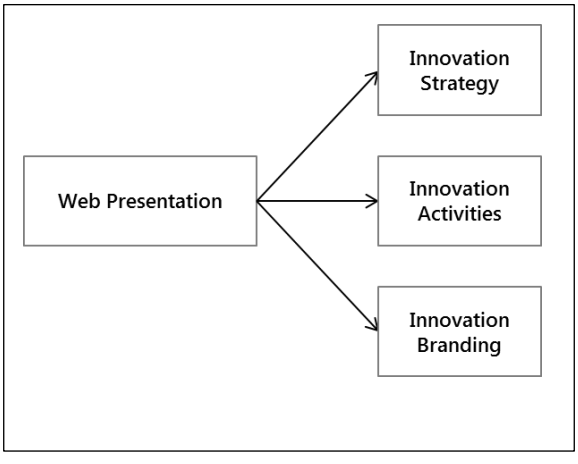
In order to measure activities on web sites that display static information, we suggest two approaches. Firstly, companies may choose to mark new products and services, or even some process or organizational innovations as new on their web sites, especially in sections that contain press releases or information about specific products and services. Keyword analyses are likely to identify some of these innovations. Secondly, the use of the internet archives allows for comparison between versions of the web sites, which may allow to identify new products and services offered (Youtie et al., 2012), but also some changes in organizational structures.

Innovation Branding

Focusing on survey methods, the Oslo Manual does not collect data on the marketing of innovations, i.e. whether a company presents itself as innovative in its marketing and branding efforts. However, company web sites are specifically designed for marketing purposes (Blazquez and Domenech, 2017b), and as such it can be expected that some presentations of innovativeness on company websites should be considered branding, rather than indications of substantive innovation. Web-based methods need to consider this aspect and develop measures to differentiate between these representations.

Innovation Web-scraping Framework

In conclusion, our paper presents a framework designed to match the theoretical and empirical underpinnings of the Oslo Manual (see Figure 1). It focuses on two dimensions covered in the Oslo Manual – innovation strategies and activities – and complements them with the consideration of web branding. We also discuss control variables such as company size (Domenech, 2012).



Application of this framework will allow validation against the large Oslo Manual data sets available to for many countries and opens opportunities for large scale web-based innovation research.

Figure 1: Innovation Web-Scraping Framework

Arora, S.K., Li, Y., Youtie, J., Shapira, P., 2015. Using the wayback machine to mine websites in the social sciences: a methodological resource. Journal of the Association for Information Science and Technology.

Arora, S.K., Youtie, J., Shapira, P., Gao, L., Ma, T., 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms. Scientometrics 95, 1189-1207.

Blazquez, D., Domenech, J., 2017a. Is web data capable of detecting firms' activity status?, NTTS Conference, Brussels, Belgium.

Blazquez, D., Domenech, J., 2017b. Web data mining for monitoring business export orientation. Technological and Economic Development of Economy, 1-23.

di Tollo, G., Tanev, S., Liotta, G., De March, D., 2015. Using online textual dta, principal component analysis and arificial neural networks to study business and innovation practices in technology-driven firms. Computers in Industry 74, 16-28.

Domenech, J., 2012. An intelligent system for retrieving economic information from corporate websites, Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, pp. 573-578.

Garechana, G., Río-Belver, R., Bildosola, I., Salvador, M.R., 2017. Effects of innovation management system standardization on firms: evidence from text mining annual reports. Scientometrics, 1-13.

Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102, 653-671.

Jianhua, L., Porter, A., Zhixiong, Z., Hongmei, G., 2016. Comparison of different "window-size" key phrase co-occurrence for knowledge representation, Global TechMining Conference.

Landers, R.N., Brusso, R.C., Cavanaugh, K.J., Collmus, A.B., 2016. A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological methods 21, 475.

Libaers, D., Hicks, D., Porter, A., 2010. A taxonomy of small firm technology commercialization. Industrial and Corporate Change 25, 371-405.

OECD - Organisation for Economic Co-operation Development, 1997. The measurement of scientific and technological activities: proposed guidelines for collecting and interpreting technological innovation data: Oslo manual. OECD.

Youtie, J., Hicks, D., Shapira, P., Horsley, T., 2012. Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. Technology Analysis & Strategic Management 24, 981-995.