Identifying data-driven innovative companies: A web mining approach

Denilton Darold (denilton.d@gmail.com), Fraunhofer ISI

The advances in big data technologies have enabled the development of a wide range of digital services using data as a key resource. Entrepreneurs have seen this as an opportunity to explore and generate new data-intensive products, services, and business models, usually grouped under the term data-driven innovations (DDI). However, the pervasiveness of DDI, spread over a wide range of sectors, makes the identification and measurement of these innovations and the companies behind them ever challenging. One approach is to investigate the adoption of enabling technologies. The issue is that the traditional data sources don't offer the appropriate means of doing that. Surveys become promptly outdated given the speed of technological development. In addition, they have a small scale, undermining the variety of such a diverse topic. The firms' databases provide a large scale, but fail to provide accurate information on the IT technologies, hindering the capacity to assess technology adoption.

Therefore, in this study, we propose the use of web mining to complement the data from firms' databases, enriching the structured information with the text collected from firms' websites. The resulting dataset serves as the source for identifying and classifying datadriven companies, which is done by finding evidence of the adoption of big data technologies and subsequently framing it into a DDI taxonomy.

The goal is to develop and assess a web mining method for firm identification using text information combined with structured and semi-structured data sources. The present use case provides substance to investigate the determinants for DDI through firm characteristics like sector, size, age, funding, legal form - whether subsidiaries or standalone, and founders' background. This firm-level analysis is part of doctoral research and aims to produce large-scale empirical evidence on the structure and success factors of DDI firms in Germany expanding the related studies based on surveys.

Background

Ubiquitous computing and high-speed digitalization have generated massive amounts of data and enabled the development of new products and services based on the insights extracted from these data pools (Geogeghan and Formica, 2018). Traditionally, data analytics plays a supportive role as a decision support system designed to answer a predefined question or improve operational procedures (Wu et al., 2020). In recent years, however, it has more and more started to assume a strategic role in creating new opportunities for innovation (Trabucchi et al., 2018). For this phenomenon, i.e., the

combination of data analytics and innovation, scholars in digital innovation have coined the term data-driven innovation (DDI) (Rizk et al., 2017). The OECD defines data-driven innovation as "the strategic utilization of data and analytics to improve or foster new processes, products, services, and markets" (OECD Publishing, 2015).

The pervasiveness of applications, the heterogeneity of data ecosystems, and the endless business possibilities make the identification of data-driven innovative companies ever challenging. Firms with traditional business models are adding DDI to their portfolio, so there is no clear cut between business sectors. Such complexity requires complementary data sources. To address this problem, a data-driven method is planned, collecting and combining data from several data sources, including structured and unstructured data, through natural language processing (NLP), aiming to increase the scale and consequently the reach and depth of the empirical analysis, to the present date done via surveys. The focus of this study is to assess the web mining data, collecting data from firms' websites.

Web mining is conceptualized as the application of data mining techniques to discover relevant data patterns and relationships from unstructured web data. This technique has been deemed applicable in many fields of research (Askitas & Zimmermann, 2015)

The text collected from firms via web mining or web scraping is a rich source of information for research, as they are used to publish information about their products and services and highlight their innovations (Gök et al., 2015). Moreover, it is possible to spot differences according to firm characteristics based on the number of subpages, hyperlinks, text volume, and language (Kinne and Axenbeck, 2019).

Compared to traditional methods applied to innovation studies, web text presents great potential. The usual indicators like patents and publications have limited coverage as they narrow the population to firms with the means to seek legal protection or to sponsor top-notch academic production. Kinne and Axenbeck (2019) list the shortcomings of traditional indicators as follows:

- Coverage: They cover only part of the firm population;
- Granularity: They lack sectoral, technological, and geographical granularity; Timeliness: They present an outdated version of STI; Cost: High cost for data collection.

Therefore, considering these shortcomings and the magnitude of web data, especially compared to surveys, web scraping offers a promising alternative to identify DDI companies.

Methods

The research steps to accomplish the task comprehend the data collection, processing, and integration, combining the data sources. Once the data sources are joined, the analysis takes place, aiming to group the companies according to the technology groups based on the reference model for Big Data technologies (Curry et al., 2021), namely: Data Visualization, Data Analytics, Data Processing Architecture, Data Protection, and Data Management. The identification and framing are made via keyword selection. The result is an overview of technology adoption of data-related technologies. In Figure 1 an overview of the workflow is shown.

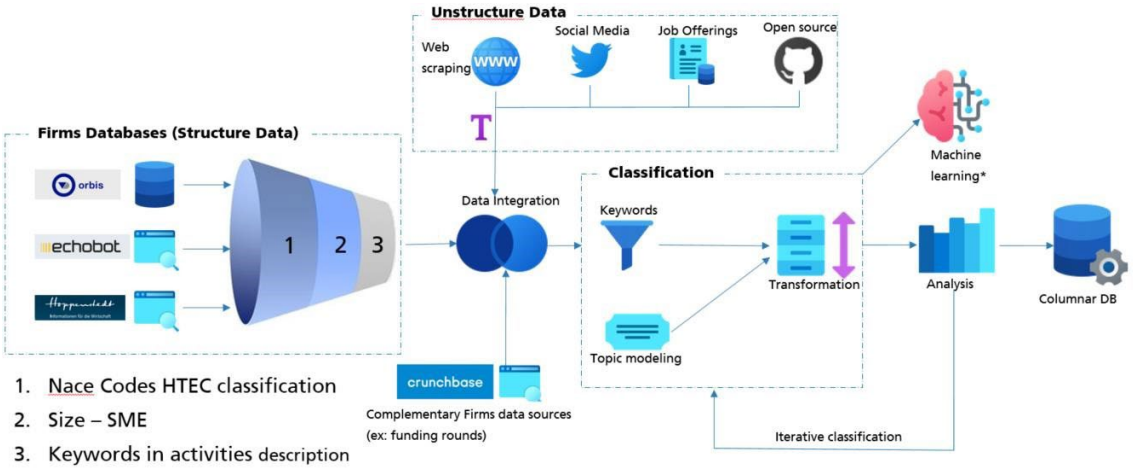

Figure 1 - Overview of methodology

Source: own compilation.

The data collection encompasses data from companies' data vendors and also unstructured data, enabling additional sources to identify innovation activities related to DDI. For the structured data, companies' information will be collected from the ORBIS database (complemented by Echobot, Hoppenstedt and Crunchbase) for a period of ten years. After that, further delimitations like the NACE code are applied to narrow down the data volume, considering only knowledge-intensive firms, according to a classification by Eurostat[1].

[1] HTECH. https://ec.europa.eu/eurostat/cache/metadata/en/htec_esms.htm

The unstructured data, the web scraped data from the firm's websites, is collected using the python library name scrawler[2]. Given the scale, only a few pages per firm will be collected. The number of firms in the preliminary queries amounts to hundreds of thousands of companies. Therefore, the tool is parameterized to scrape only five subpages of each domain. The depth, i.e., how many levels (domains and subdirectories) the crawler digs into the domain discovering internal URLs, is set to three. As the firms' websites are quite diverse, containing unpredictable layouts, data formats, and media files, additional controls are implemented to improve the quality of the data collected and ensure the efficient use of computational resources, avoiding unnecessary scraping.

The data collected, one CSV file per firm, is then aggregated using the python library *pandas*[3]. Subsequently, several NLP techniques are applied to clean and transform the text, splitting it into sentences. Then, the text passes through stemming and normalization for the keyword queries to improve the assertiveness and the number of matchings. Next, firms are classified according to the aforementioned five big data technology groups through this matching procedure. Finally, the sentences considered relevant in the web scraped data go through a summarization procedure using BERT[4] in order to produce a succinct description of firms based on their websites.

Results

The goal is to enrich the existent firms' structured data with text content from their websites via web scraping. The methodological challenges go from collecting the data to properly selecting the relevant content to achieve the ultimate goal of identifying DDI companies in Germany. The final dataset will contain a comprehensive list of indicators along with text information scraped from firms' websites. Such a dataset provides substance to a throughout analysis of the determinants of adoption as well the structure and success factors of DDI companies. The contribution regarding the methodology lies in the identification of companies via technology adoption using web scraped through data science techniques, combining diverse data sources, and increasing the reach and variety of the empirical analysis.

---

[2] Scraper tool: https://github.com/dglttr/scrawler/tree/main/scrawler

[3] Data analysis tool: https://pandas.pydata.org/

[4] Summarizer: https://pypi.org/project/bert-extractive-summarizer/

Bibliography

Askitas, N., & Zimmermann, K. F. (2015). The Internet as a data source for advancement in social sciences. International Journal of Manpower, 36(1), 2–12. https://doi.org/10.1108/IJM-02- 2015-0029

Curry, E., Zillner, S., Metzger, A., Berre, A.J., Auer, S., Walshe, R., Despenic, M., Petkovic, M., Roman, D., Waterfeld, W., Seidl, R., Hasan, S., ul Hassan, U., Ojo, A., 2021. Technical Research Priorities for Big Data. Elem. Big Data Value 97–126. https://doi.org/10.1007/978-3-030-68176-0_5

Geogeghan, M., Formica, P., 2018. Data-Driven Innovation 123–127. https://doi.org/10.1007/978-3-319-62878-3

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

Kinne, J., Axenbeck, J., 2019. Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany. ZEW Discuss. Pap.

Nickerson, R.C., Varshney, U., Muntermann, J., 2013. A method for taxonomy development and its application in information systems. Eur. J. Inf. Syst. 22, 336–359. https://doi.org/10.1057/ejis.2012.26

OECD Publishing, 2015. Data-Driven Innovation: Big Data for Growth and Well-Being. OECD Publishing, Paris. https://doi.org/10.1787/9789264229358-en

Rizk, A., Bergvall-Kåreborn, B., Elragal, A., 2017. Digital service innovation enabled by big data analytics - A review and the way forward. Proc. Annu. Hawaii Int. Conf. Syst. Sci. 2017-Janua, 1247–1256. https://doi.org/10.24251/hicss.2017.149

Trabucchi, D., Buganza, T., Dell'Era, C., Pellizzoni, E., 2018. Exploring the inbound and outbound strategies enabled by user generated big data: Evidence from leading smartphone applications. Creat. Innov. Manag. 27, 42–55. https://doi.org/10.1111/caim.12241

Wu, L., Hitt, L., Lou, B., 2020. Data analytics, innovation, and firm productivity. Manage. Sci. 66, 2017–2039. https://doi.org/10.1287/mnsc.2018.3281