GI_Forum 2018, Issue 1
Page: 82 - 89
Short Paper
Corresponding Author:
bernd.resch@sbg.ac.at
DOI: 10.1553/giscience2018_01_s82

Generating Big Spatial Data on Firm Innovation Activity from Text-Mined Firm Websites

Jan Kinne[1] and Bernd Resch[2]

[1]Centre for European Economic Research, Germany
[2]University of Salzburg, Austria

Abstract

Innovation is one of the major drivers of economic growth, where spatial processes of knowledge spillover play a vital role. Current practices in assessing firms' innovation activity, including patent analysis and questionnaires, suffer from severe limitations. In this paper, we propose a novel approach to estimate firms' innovation activity based on the texts on their websites. We use an automated web-scraper to harvest text from the websites, then extract semantic topics in a self-learning, generative topic-modelling approach, and finally analyse these topics using an Artificial Neural Networks (ANN) method to assess each firm's level of innovation. This procedure results in a large-scale dataset that will be used for further spatial economic analysis of the distribution of innovative firms and the processes that drive the development of innovation in firms.

Keywords:

firm location, microgeography, innovation, web scraping, Big Spatial Data, text mining, topic modelling, neural networks

1 Introduction

Innovation is considered one of the main drivers of economic growth. The disruptive force of radical innovations resets the economy and paves the way for new periods of long-term economic growth, while incremental innovations lead to continuous change. An innovation is the implementation of a new or significantly improved product or process, characterized by its degree of novelty (innovations that are new to the firm, the market, the industry or the world) and type of innovation (product, process, marketing or organizational) (OECD & Eurostat, 2005). The spatial processes of knowledge exchange and collective knowledge growth known as knowledge spillovers are assumed to be one of the most important drivers of the development of innovation (Audretsch & Feldman, 2004; Florida, Adler, & Mellander, 2017). Knowledge spillovers are stimulated by high density and diversity of people, firms and institutions, which offer opportunities for interaction, cooperation and mutual learning. Economic actors from various backgrounds are brought together, often by chance, which allows them to exchange ideas and to recombine them into new and productive forms,

driving the development and diffusion of innovation (Helbing, Ku, West, & Bettencourt, 2007; Nelson, 2009).

Geographically detailed studies indicate that knowledge spillovers operate at a fine spatial scale and decay rapidly within a few hundred metres (Arzaghi & Henderson, 2008; Carlino & Kerr, 2015; Catalini, 2012; Jang, Kim, & von Zedtwitz, 2017; Kabo, Cotton-Nessler, Hwang, Levenstein, & Owen-Smith, 2014; Kerr, Duranton, Glaeser, & Henderson, 2014). Ideally, the analysis of these microgeographic processes requires comprehensive and non-aggregated geographic data on firms and conditional location factors. Geographic data on innovation-related infrastructure has become available only recently with the emergence of Volunteered Geographic Information (VGI) and the increasing availability of open geodata from government agencies (Elwood, Goodchild, & Sui, 2012; Goodchild & Longley, 2014; Sui & Goodchild, 2011). Such data have been used for microgeographic analyses of firm location patterns (Ahlfeldt, 2013; Ahlfeldt & Richter, 2013; Kinne & Resch, 2018; Möller, 2014; Rammer, Kinne, & Blind, 2016). However, traditional firm-level innovation geodata (i.e. patents owned by firms, and indicators from questionnaires) is scarce, as it covers only a fraction of the firm's population or is restricted to geographic areas such as single cities (see Figure 1). A microgeographic analysis of innovation processes requires a dense (complete) geographic pattern of firm locations with related information on the firms' innovation activity. This lack in data is associated with a gap in the understanding of the micro-foundations of the development and diffusion of innovation.

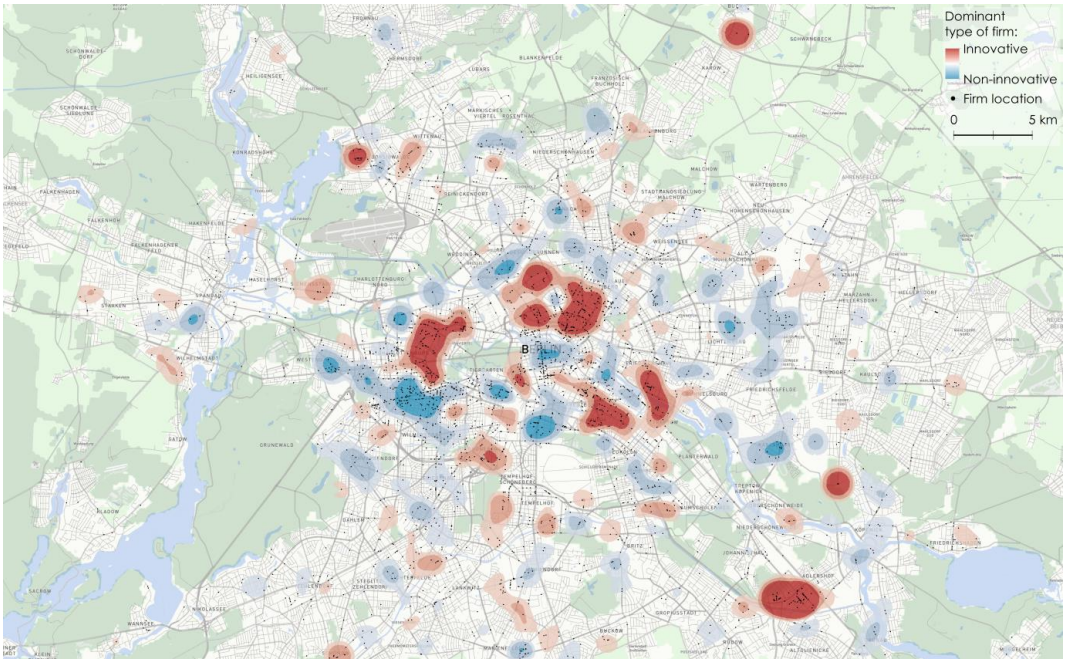

Figure 1: Micro-location patterns of innovative and non-innovative firms in Berlin 2011–2015. Data: ZEW 2018; Basemap: Mapbox/OSM. [Adapted from Rammer, Kinne, & Blind, 2016]

We propose a novel approach to generate big spatial data on firm-level innovation activity: *innovation indicators from firms' websites*. For this purpose, firm websites are crawled and texts on these websites are downloaded using an automated web-scraper (similar to Gök et al., 2015). In doing this, we leverage the fact that almost all (significant) firms have websites nowadays. We assume that firms use their websites to provide up-to-date information on their products and services, highlighting their new and innovative features. This information is publicly available and is digitally codified (i.e. it is *codified knowledge*). Innovation indicators based on these frequently-updated website texts have the potential to provide accurate information on firm-level innovation activity.

The major challenge for this research is the identification and extraction of the bits of information on innovation activities from the overall text corpus. To achieve this, we leverage the recent development in text data-mining (e.g. Latent Dirichlet Allocation (LDA)) (Blei, Ng, & Jordan, 2003), which has also been applied successfully in GIScience (Resch, Usländer, & Havas, 2017; Steiger, Resch, & Zipf, 2016). We also use (deep) neural networks (see e.g. Grentzkow et al., 2017; Mennis & Guo, 2009), which have likewise been applied successfully, in urban (economic) geographical analysis (Steiger et al., 2016).

In this paper, we propose a methodology for automatically scraping firms' websites and estimating their innovation activity using a combination of a topic-modelling method (LDA) and an Artificial Neural Network (ANN) trained on a set of established, labelled (categories innovative vs. non-innovative), innovation indicator data (see Section 3, Methods, for more details).

2 Data

The Mannheim Enterprise Panel

The *Mannheim Enterprise Panel* (MUP) is a database which covers the total number of firms located in Germany. It contains about three million firms which are updated on a semi-annual basis. In 2017, the MUP included about 2.97 million active firms in Germany. The data covers their characteristics, such as the branch of industry (using NACE codes, a classification of economic activities in the European Union), as well as their postal addresses and URLs (Bersch, Gottschalk, Müller, & Niefert, 2014).

The Mannheim Innovation Panel

The Mannheim Innovation Panel (MIP) is based on a questionnaire-based innovation survey which covers the areas of mining, manufacturing, energy, construction, producer services and distributive services. It provides information about the introduction of new products, services and processes, and the expenditures on innovations. The annual MIP survey covers about 10,000 firms (0.3% of the total German firm population).

PATSTAT Patent Database

The PATSTAT patent database maintained by the European Patent Office contains bibliographical information and data on the legal status of patents relating to over 100 million patent documents from European and non-European countries (European Patent Office, 2016). We merged the PATSTAT database with the Mannheim Enterprise Panel to extend the latter with another established innovation indicator which can be used in our proposed framework.

3 Methods

The overall methodology of our approach is illustrated in Figure 2. The workflow consists of (1) filtering website addresses based on information in the MUP; (2) web scraping the text content from the firms' websites; (3) selecting an appropriate classification model based on MUP metadata (e.g. industrial sector), classifying the scraped website texts, and subsequently generating indicators of innovation; (4) storing the results; (5) querying the data for geographical analysis.

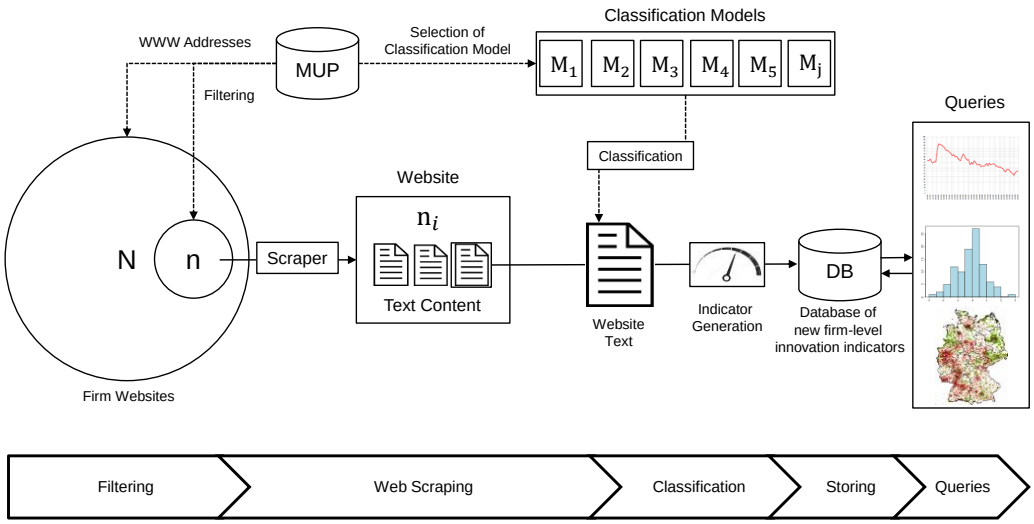

Figure 2: Workflow: web scraping, text classification, innovation indicator generation, storing and querying

Web Scraping

We programmed a web-scraper (*ARGUS*; Kinne, 2018) which is able to perform so called *broad crawls*, i.e. the scraper is not restricted to and specialized for a single website or type of website, but is able to crawl a broad range of different websites. The web-scraper was built using the Scrapy Python framework (Scrapy Community, 2008) and is free to use. An

ARGUS broad crawl is based on the URLs of the firms' websites included in the Mannheim Enterprise Panel (MUP) and proceeds as follows:

1. The firm's website is requested using the URL in the MUP.
2. A collector item is instantiated, which is used to collect the website's text, meta-data (e.g. timestamps, number of scraped URLs etc.), and a so-called URL stack.
3. The firm's website main page is processed:
   a. The main page's texts are extracted and stored in the collector item.
   b. URLs which refer to subpages of the same firm website (i.e. domain) are extracted and stored in the collector item's URL stack.
4. The algorithm continues to request URLs from the URL stack using a simple heuristic which gives higher priority to URLs of shortest length and those which refer to webpages in a predefined language.
   a. Texts and URLs are collected from the subpage and stored in the collector item.
   b. The next URL in the URL stack is processed.
5. The algorithm stops processing a domain when all subpages have been processed or as soon as a predefined number of subpages for the domain have been processed.
6. The collected texts are processed (i.e. cleaned) and written to an output file.
7. The next firm's main page is requested, and the process is repeated until all firm website addresses have been processed.

Estimating Firms' Innovation Acivities using LDA Topic-Modelling and Artificial Neural Networks

Figure 3 illustrates the methodology for estimating the innovation activity of a firm. We train a neural network to identify innovative firms based on web-scraped texts from their websites. The texts are first transferred to LDA topic probability vectors and then used as input to a neural network. As labelled training data for the neural network, we use web-scraped texts of firms for which established firm-level innovation indicators are available (patents, indicators from questionnaires) (see Section 2, Data). Using the trained neural network, we will be able to estimate the innovation activity of the entire firm population.

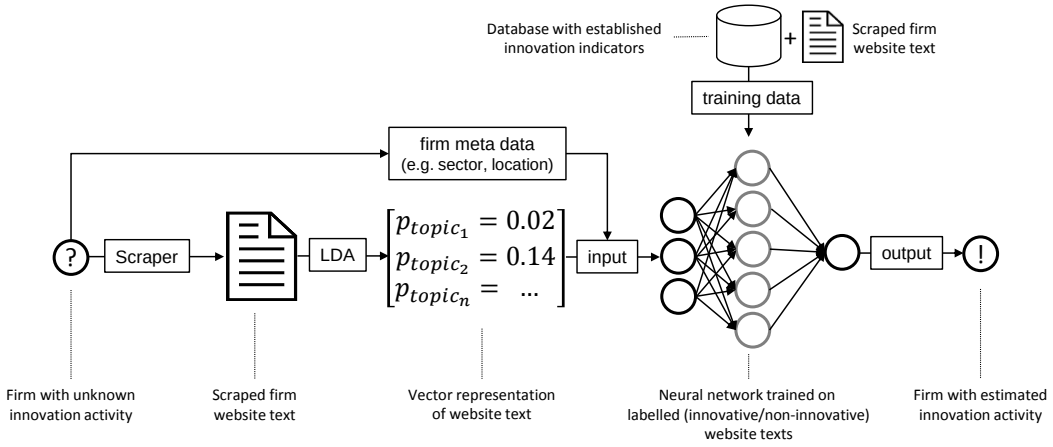

Figure 3: Proposed model to estimate the innovation activity of firms based on their website texts

LDA assumes that each document *d* of a set of documents *D* contains one or more topics z, which is (or are) again defined by a probability distribution of single words *w*, the only observed variable in the model. The latent variable ϕ represents a multinomial distribution of words within a topic. The other latent variable, θ, constitutes a multinomial distribution of topics in a document. a and β are two concentration parameters: a represents prior knowledge about the distribution of topics in a document, whereas β contains prior knowledge about the distribution of words in a topic. A higher value for a leads to a more smoothed distribution of topics over a document; a lower value, especially lower than zero, leads to a higher concentration of topics. ϕ, θ and z are latent and therefore unobserved variables, which are generated when the process is running (Griffiths & Steyvers, 2004). Ultimately, LDA generates topics that are defined by a number of particular words.

4 Preliminary Results and Conclusion

Figure 4 shows an extract of our firm database that was generated by the ARGUS web-scraper. We were able to extract the website texts of 2.3 million firm websites, which can be used in the subsequent text analysis. Using a conventional office-grade PC, the web-scraping took only about six days, making our ARGUS web-scraper a suitable tool for regular and frequent massive web-scraping. With the successful extraction of texts from a large number of diverse websites, the first step in this long-term study has been completed.

We identified a number of possible issues concerning the subsequent text analysis. First, the generative LDA topic model cannot handle multiple languages at once. Thus, we need to apply language-detection techniques carefully in a pre-processing step to produce 'pure' and non-skewed topics. We implemented a simple language-detection heuristic in our ARGUS web-scraper, which enables us to restrict the web-scraping to certain languages. Furthermore, no automated, quantitative validation method has been defined yet. Thus, we still rely on experts in the field to rate our analysis results, which are not scalable to large datasets. Finally, we cannot be sure that our planned text analysis will work out as desired. For

example, our training data sets with established innovation indicators may be too small to successfully train our neural network for example.

In the near future, we will work on the text analysis algorithms and a rigorous quantitative validation procedure. We will then feed the results into our actual research – i.e., assessing the distribution of innovative and non-innovative firms on a fine spatial scale, together with analysing a variety of microgeographic co-variates for the development of innovation in firms.

ID	dl_rank	dl_slot	error	redirect	start_page	text	timestamp	
313455	0	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://ww
313455	1	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.z
313455	2	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.
313455	3	zew.de	None	False	https://www.zew.de/	Das Zentrum für Europäische Wirtschaftsforschu...	Wed May 23 13:06:30 2018	https://www.zew.de
313455	4	zew.de	None	False	https://www.zew.de/	Unser informiert Sie direkt über Neuigkeiten a...	Wed May 23 13:06:30 2018	https://www.zew.de/d
313455	5	zew.de	None	False	https://www.zew.de/	Tel: 0621 1235-132 Fax: 0621 1235-255 E-Mail: ...	Wed May 23 13:06:30 2018	https://www.zew.de/d
313455	6	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de
313455	7	zew.de	None	False	https://www.zew.de/	Das ZEW ist ein gemeinnütziges wirtschaftswiss...	Wed May 23 13:06:30 2018	https://www.zew.de/de
313455	8	zew.de	None	False	https://www.zew.de/	Wenn Sie den Hauptbahnhof verlassen haben, übe...	Wed May 23 13:06:30 2018	https://www.zew.de/d

Figure 4: Exemplary extract of the database containing the scraped website texts of a single firm.

References

Ahlfeldt, G. M. (2013). *Urbanity* (SERC Discussion Paper No. 136). *SERC Discussion Paper*. London.

Ahlfeldt, G. M., & Richter, F. J. (2013). *Urban Renewal after the Berlin Wall* (Serc Discussion Paper No. 151). *Serc Discussion Paper 151*. London. https://doi.org/10.1093/jeg/lbw003

Arzaghi, M., & Henderson, J. V. (2008). Networking off Madison Avenue. *Review of Economic Studies*, *75*(4), 1011–1038. https://doi.org/10.1111/j.1467-937X.2008.00499.x

Audretsch, D. B., & Feldman, M. P. (2004). Knowledge spillovers and the geography of innovation. *Handbook of Regional and Urban Economics*, *4*(December 2002), 2713–2739. https://doi.org/10.1016/S1574-0080(04)80018-X

Bersch, J., Gottschalk, S., Müller, B., & Niefert, M. (2014). *The Mannheim Enterprise Panel (MUP) and firm statistics for Germany*. *ZEW Discussion Paper*. https://doi.org/10.2139/ssrn.2548385 M4 - Citavi

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Carlino, G., & Kerr, W. R. (2015). Agglomeration and Innovation. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics* (Vol. 5, pp. 349–404). Amsterdam: Elsevier North-Holland. https://doi.org/10.1016/B978-0-444-59517-1.00006-4

Catalini, C. (2012). *Microgeography and the Direction of Inventive Activity*. *Rotman School of Management Working Paper* (Vol. 2126890). https://doi.org/10.1287/mnsc.2017.2798

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American*

Geographers, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657
European Patent Office. (2016). The PATSTAT product line. Munich: European Patent Office.
Florida, R., Adler, P., & Mellander, C. (2017). The City as Innovation Machine. *Regional Studies*, *51*(1), 86–96. https://doi.org/10.1080/00343404.2016.1255324
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
Goodchild, M. F., & Longley, P. A. (2014). The Practice of Geographic Information Science. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of Regional Science* (pp. 1107–1122). Berlin, Heidelberg: Springer.
Grentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as Data* (NBER Working Paper Series No. 23276). Cambridge, Massachusetts.
Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.
Helbing, D., Ku, C., West, G. B., & Bettencourt, L. M. A. (2007). Growth, innovation, scaling, and the pace of life in cities. *PNAS*, *104*(17), 7301–7306.
Jang, S., Kim, J., & von Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research*, *78*(June), 143–154. https://doi.org/10.1016/j.jbusres.2017.05.017
Kabo, F. W., Cotton-Nessler, N., Hwang, Y., Levenstein, M. C., & Owen-Smith, J. (2014). Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy*, *43*(9), 1469–1485. https://doi.org/10.1016/j.respol.2014.04.007
Kerr, W. R., Duranton, G., Glaeser, E., & Henderson, V. (2014). Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics*, *96*(3).
Kinne, J. (2018). ARGUS - An Automated Robot for Generic Universal Scraping. Mannheim: Centre for European Economic Research. Retrieved from https://github.com/datawizard1337/ARGUS
Kinne, J., & Resch, B. (2018). Analyzing and Predicting Micro-Location Patterns of Software Firms. *ISPRS International Journal of Geo-Information*, *7*(1), 26. https://doi.org/10.3390/ijgi7010001
Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery-An introduction. *Computers, Environment and Urban Systems*, *33*(6), 403–408. https://doi.org/10.1016/j.compenvurbsys.2009.11.001
Möller, K. (2014). *Culturally clustered or in the cloud? Location of internet start-ups in Berlin* (SERC Discussion Paper No. 157). *SERC Discussion Paper* (Vol. 157). London: London School of Economics.
Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005. https://doi.org/10.1016/j.respol.2009.01.023
OECD, & Eurostat. (2005). *Oslo Manual: Guidlines for Collecting and Interpreting Innovation Data*. *Communities* (Vol. Third edit). OECD. https://doi.org/10.1787/9789264013100-en
Rammer, C., Kinne, J., & Blind, K. (2016). *Microgeography of innovation in the city: Location patterns of innovative firms in Berlin* (ZEW Discussion Paper No. 16–080). Mannheim.
Resch, B., Usländer, F., & Havas, C. (2017). Combining Machine-learning Topic Models and Spatio-temporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment. *Cartography and Geographic Information Science*, https://doi.org/10.1080/15230406.2017.1356242.
Scrapy Community. (2008). Scrapy. Scrapinghub Ltd. Retrieved from https://github.com/scrapy/scrapy
Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographic Information Science*, *30*(9), 1694–1716.
Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. https://doi.org/10.1080/13658816.2011.604636