From university research to innovation Detecting knowledge transfer via text mining

Woltmann, Sabrina; Clemmensen, Line Katrine Harder; Alkærsig, Lars

Published in:
Proceedings of the 21st international conference on Science and Technology Indicators (STI 2016)

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

From university research to innovation
Detecting knowledge transfer via text mining

Sabrina Woltmann*, Line H. Clemmensen** and Lars Alkærsig*

*swol@dtu.dk; lalk@dtu.dk
Management Engineering, Technical University of Denmark, Centrifugevej 372, 2800 Kgs. Lyngby, (Denmark)

** lkhc@dtu.dk
Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, (Denmark)

ABSTRACT
Knowledge transfer by universities is a top priority in innovation policy and a primary purpose for public research funding, due to being an important driver of technical change and innovation. Current empirical research on the impact of university research relies mainly on formal databases and indicators such as patents, collaborative publications and license agreements, to assess the contribution to the socioeconomic surrounding of universities. In this study, we present an extension of the current empirical framework by applying new computational methods, namely text mining and pattern recognition. Text samples for this purpose can include files containing social media contents, company websites and annual reports. The empirical focus in the present study is on the technical sciences and in particular on the case of the Technical University of Denmark (DTU). We generated two independent text collections (corpora) to identify correlations of university publications and company webpages. One corpus representing the company sites, serving as sample of the private economy and a second corpus, providing the reference to the university research, containing relevant publications. We associated the former with the latter to obtain insights into possible text and semantic relatedness. The text mining methods are extrapolating the correlations, semantic patterns and content comparison of the two corpora to define the document relatedness. We expect the development of a novel tool using contemporary techniques for the measurement of public research impact. The approach aims to be applicable across universities and thus enable a more holistic comparable assessment. This rely less on formal databases, which is certainly beneficial in terms of the data reliability. We seek to provide a supplementary perspective for the detection of the dissemination of university research and hereby enable policy makers to gain additional insights of (informal) contributions of knowledge dissemination by universities.

INTRODUCTION
Universities are facing increasing demands for active dissemination of their research results and their contribution to knowledge development in their socioeconomic environment (Jongbloed, Enders & Salerno 2008); commonly referred to as the third mission. Since knowledge is a crucial aspect for innovation processes, its transfer has become a new policy priority and is often directly targeted by public funding (Ramos-Vielba, Fernández-Esquinas & Espinosa-de-los-Monteros 2009, Huggins & Johnston 2009). This study covers the extent

of knowledge transfer of university research within technical sciences, as these are key drivers for innovation.

Current empirical research focuses primarily on the analysis of formal interactions between universities and their company partners (Broström 2010), relying on indicators such as patents, collaborative publications, contracts and license agreements (Drucker & Goldstein 2007). These well-developed empirical approaches somewhat capture the success of knowledge dissemination and commercialization of university driven innovations. However, these studies bear some deficiencies, as they often fail to include indirect impacts by focusing on formal cooperation and knowledge exchange. Additionally, most empirical studies frequently require complex adjustments for each unique case. Moreover, their key indicators often depend on formal databases with varying quality and accessibility and they require long-term assessments, which delays the outcomes and limit comparability (Vincett 2010).

In this study, we use modern computational methods to expand the empirical framework by introducing specific data mining approaches and testing these on the Technical University of Denmark (DTU). To complement the current scope, we focus in particular on the application of text mining and pattern recognition tools. These tools capture occurrences where knowledge is used without a statement about its origin. Our data sources include the online presence of companies in regional proximity to the university, including social media sites, company websites and annual reports.

The study's intent is to counteract certain empirical challenges, by detecting knowledge transfer without focusing on formal cooperation channels and develop additional indicators also capturing informal contributions. Compared to traditional assessments, the main advantages are that the measure is instantaneous, resulting in reduced time delay, and that it relies less on formal databases.

METHODOLOGY

We seek to generate a complementary perspective by applying novel computational methods and embedding them in the current impact assessment framework of public research, therefore seizing the widely agreed potential of those applications by adapting them to our specific purpose. We capture, identify and verify the existence and extent of knowledge transfer to the economic surrounding of the DTU. To identify research outcomes, which can be attributed to the university, this study uses text-mining methods.

To implement these measures we follow systematic and distinct actions, including the sample generation, data collection, pre-processing and the application of statistical correlation measures.

Sample generation

Assuming that the private economy is an essential beneficiary of knowledge exchange, we included relevant private companies. Defined as companies

- with direct relations to DTU defined by hyperlinks on the DTU website (first-degree partners) ;
- with indirect relations to DTU defined by hyperlinks on partner websites (second-degree partners);

- with regional facilities near by the DTU (in the national context of Denmark, indicated via a Danish VTA registration).

Data collection
This study uses company websites, which consist of unstructured text-data, to identify company knowledge, products and expertise. Thus, we gathered these texts, which are available in form of online publications released by the companies themselves, and extracted them as HTML files. Associated social media entries will be included at a later stage of the project, as social media content requires specific treatment due to their special linguistic composition. The collected HTML files are pre-processed and transformed into unstructured raw text, maintaining only content and semantically relevant information. We implemented language identification parameters to extract exclusively English texts (Palmer 2010).

Text mining
To analyze the data, we apply methods from the field of natural language processing (NLP), as it provides tools for simple and advanced text analytical procedures. Text mining requires text corpora containing the relevant text fragments in form of tokens. In our case, we developed one text corpora derived from the raw text files of the company websites and a second 'reference' corpus containing an extensive sample of research publications. The university online publication database ORBIT provided the texts for the reference corpus, as this database comprises almost all publications including patents, projects, etc. made by DTU employees[1].

Pattern recognition algorithms and machine learning methods provide in-depth comparisons between the reference and the company corpus (Bird, Klein & Loper 2009). To extrapolate the important patterns, including correlations, semantic compositions and outlier comparison, this study uses various available text mining methods. These include term-based methods, phrase-based methods, etc., which provide a variety of statistical tools to analyze the texts and to achieve our objectives. The analysis includes statistical measures that identify document relatedness, correlations or different types of regression parameters. Hereby, we quantify the extent of correlations between documents of the two corpora and the corpora themselves.

To detect the similarities between texts from the two corpora we use specially adapted machine learning algorithms, which extract key features from the reference corpus and compare them with the company corpus. We aim to include semantically correlated and content related approaches, to ensure the methods capture not only obvious semantic, but also content correlations. Accordingly, this approach allows us to detect shared contents among documents and enables the tracing of knowledge, which provides evidence-based insights in the 'relatedness' between the corpora.

We use statistical models, which include, but are not limited to, methods for dimensionality reduction like latent semantic analysis (LSA) (Landauer, Foltz & Laham 1998) and, for uncovering the underlying structures of the documents, probabilistic topic models for instance latent dirichlet allocation (LDA) and correlated topic models (CTM) (Blei, Ng & Jordan 2003). However, as NLP is a comparatively young field its methods undergo continuous development, therefore specific adjustments to its models are inevitable.

[1] http://orbit.dtu.dk/en/about.html

Evaluation of the method
Given the identification of the extent of knowledge transfer by tracing linguistic and semantic content, we seek to extrapolate the research areas, which spread most knowledge and the companies, which make most use of university research within their proximity. To evaluate the relevance of our findings and to conclude whether our findings truly increase the understanding and measurement of (indirect) impacts we will compare our results to those of conventional measures.

POLICY RELEVANCE AND POTENTIAL

The study provides a supplementary perspective for the detection of research dissemination and impact of university innovations. Our intention is to contribute to the understanding of university performance by enhancing the detection of impacts of publicly funded research. Current computational methods provide novel possibilities for measurements allowing additional benchmarking as foundation for decision-making processes.

The goal is to provide policy makers with additional insights on the applications of university knowledge, allowing them to evaluate the benefits of government funding of research in a more holistic manner by including so far undetected, but essential impacts. This study can shed new light on the contributions universities make to economy and society.

Advantages of this novel approach are firstly, the availability of data, contrary to conventional assessments, which rely highly on university databases, which vary in quality and accessibility. Secondly, the potential to apply these measures in different regional, societal and economic contexts. Thirdly, the instantaneous nature of the measurement could capture the outcomes and the status quo almost in real-time.

After an in-depth evaluation of our approach against existing measures, we will be able to verify the extent of additional information that can be drawn from this new approach. Ideally, it will provide a greater overview about (informal) knowledge exchange from universities to companies, providing a more detailed picture for future oriented decision-making.

REFERENCES

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol: O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Broström, A. (2012). Firms' rationales for interaction with research universities and the principles for public co-funding. *The Journal of Technology Transfer, 37*(3), 313-329.

Drucker, J., & Goldstein, H. (2007). Assessing the regional economic development impacts of universities: a review of current approaches. *International regional science review, 30*(1), 20-46.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Huggins, R., & Johnston, A. (2009). The economic and innovation contribution of universities: a regional perspective. *Environment and Planning C: Government and Policy*, *27*(6), 1088-1106.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284.

Palmer, D. D. (2010). *Handbook of natural language processing (second edition)*. Boca Raton: CRC Press.

Jongbloed, B., Enders, J., & Salerno, C. (2008). Higher education and its communities: Interconnections, interdependencies and a research agenda. *Higher education*, *56*(3), 303-324.

Ramos-Vielba, I., Fernández-Esquinas, M., & Espinosa-de-los-Monteros, E. (2009). Measuring university–industry collaboration in a regional innovation system. *Scientometrics*, *84*(3), 649-667.

Vincett, P. S. (2010). The economic impacts of academic spin-off companies, and their implications for public policy. *Research Policy*, *39*(6), 736-747.