Downloaded from orbit.dtu.dk on: Jun 18, 2021

DTU Library

Tracing Knowledge Transfer from Universities to Industry: A Text Mining Approach

Woltmann, Sabrina; Alkærsig, Lars

Published in:
Academy of Management Proceedings 2017 (AOM)

Link to article, DOI:
10.5465/ambpp.2017.15409abstract

Publication date:
2017

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Woltmann, S., & Alkærsig, L. (2017). Tracing Knowledge Transfer from Universities to Industry: A Text Mining Approach. In *Academy of Management Proceedings 2017 (AOM)* Academy of Management. Academy of Management Proceedings https://doi.org/10.5465/ambpp.2017.15409abstract

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Tracing Knowledge Transfer from Universities to Industry: A Text Mining Approach

ABSTRACT

This paper identifies transferred knowledge between universities and the industry by proposing the use of a computational linguistic method. Current research on university-industry knowledge exchange relies often on formal databases and indicators such as patents, collaborative publications and license agreements, to assess the contribution to the socioeconomic surrounding of universities. We, on the other hand, use the texts from university abstracts to identify university knowledge and compare them with texts from firm webpages. We use these text data to identify common key words and thereby identify overlapping contents among the texts. As method we use a well-established word ranking method from the field of information retrieval term frequency–inverse document frequency (TFIDF) to identify commonalities between texts from university. In examining the outcomes of the TFIDF statistic we find that several websites contain very related and partly even traceable content from the university. The results show that university research is represented in the websites of industrial partners. We propose further improvements to enhance the results and potential areas for future implementation. This paper is the first step to enable the identification of common knowledge and knowledge transfer via text mining to increase its measurability.

Keywords:

Text mining, knowledge transfer, impact assessment, university-industry

INTRODUCTION

Universities, as publicly funded institutions, conducting and disseminating research, are highly valued contributors to the knowledge development for economic growth and development (Feller, 1990; Howells, Ramlogan, & Cheng, 2012). The dissemination of their research outcomes is next to teaching seen as a major part of the impacts universities provide for their environment. In particular, the contribution of university research to economic development by fostering innovation leading to increased competitive advantages for industries and firms is today widely accepted (Cohen, Nelson, & Walsh, 2002; Huggins & Johnston, 2009). Academics and policy makers have in the past decades shown increasing interest in the identification of impact of the dissemination of university research; driven by the desire to ensure optimal allocation of limited public funding (Drucker & Goldstein, 2007; Rothaermel, Agung, & Jiang, 2007). Justification for the utilization of public funds thus became an incentive and are increasing the pressure to provide evidence for the return on public investments, so their societal and economic benefits are increasingly evaluated (O'Shea, Chugh, & Allen, 2008).

The increase of understanding and the evaluation of university research impacts became a political incentive and particularly the aspects of knowledge creation and transfer are in focus of assessments and evaluations (Agrawal, 2001).

Due to the high relevance of the topic, we aim to deepen the understanding of the economic impacts of university research dissemination by contributing with a new indicator and an additional novel measurement. Considering the current framework, this study takes a step back and aims to revive the work on the foundation of university research impact assessments: the notion of knowledge transfer.

The main objectives of our study are (i) to develop an additional measure of knowledge transfer (ii) to evaluate this new method using a case study of the Technical University of Denmark (DTU), chosen due to high accessibility of data, and (iii) to compare two different approaches of creating a relevant sample representing firm knowledge.

We seek to contribute by using computational methods, which are based on data mining processes, to develop our understanding of whether university knowledge is transferred and applied outside of formal collaboration and communication. The main method is derived from the field of natural language processing (NLP) and based on a concurrent text mining technique (Paukkeri & Honkela, 2010).

Text mining enables a trace from university research output, in form of publications, to corporate websites, annual reports or similar texts that give insight into firms' innovations, products and services. The goal is to identify correlations between these two types of texts, using this as an indicator for the transfer of new knowledge from the university to the firm. This paper should be seen as a first step towards identifying and understanding the characteristics of common knowledge between university and the industry. Our study contributes to the research on university-industry knowledge transfer by identifying correlations between university knowledge and firms commercially displayed knowledge via text analysis. We aim to increase insights into areas of common knowledge and mutual interests between universities and companies.

ECONOMIC IMPACT AND KNOWLEDGE TRANSFER

An extensive body of literature is concerned with the economic impact of university research. Since not one domain embraces all relevant aspects of this field of study, it has developed into a body of highly interdisciplinary works (Teixeira, 2014), providing a multitude of perspectives and definitions within the literature. Given the research diversity on publicly funded research impacts, today's understanding is comparatively well developed (Cheah, 2016). However, due to the diversity of scholars within the field, the understanding of 'economic impact' is used in varying contexts encompassing different notions, perspectives and dimensions (Cheah, 2016). Overall findings indicate different levels of economic impact for firms, sectors or regions. The benefit of university-generated knowledge is not spread uniformly across firms and sectors and national contexts (Bodas Freitas, Marques, & Silva, 2013), but examination of literature reviews and most influential empirical works reveals that the significant economic benefits of public-funded research are widely accepted (Agrawal, 2001).

Many studies follow the concept that knowledge transfer from universities to the industry is one of the key aspects of universities impact on the economy (Agrawal, 2001; Perkmann et al., 2013). "*Evidence suggests that even knowledge transferred through the formal university technology transfer channel […], is quite significant."* (Agrawal, 2001, p. 285). The body of academic literature consists of various sorts of impact studies ranging from single case studies, focusing on individual universities, to regional or even national surveys (Drucker & Goldstein, 2007; Huggins & Johnston, 2009; Rosenberg & Nelson, 1993). These diverse studies provide a great variety of methodological approaches aiming to identify university research impact, including qualitative and quantitative approaches.

Qualitative works are often concerned with in-depth understanding of motivation for university-industry collaborations or forms and channels of knowledge exchange, or focus on single universities as case studies (Ankrah, Burgess, Grimshaw, & Shaw, 2013; Perkmann & Walsh, 2009; Rothaermel et al., 2007; Siegel, Waldman, Atwater, & Link, 2004).

Quantitative studies on the other hand often provide particular insights about knowledge generation and knowledge transfer from universities to companies (D'Este & Patel, 2007; Schartinger, Rammer, & Fröhlich, 2002). Indicators used in quantitative studies comprise, among others, number of (co)-publications, number of successful university spin-offs, university income through license agreements, research collaborations and patents (Agrawal, 2001; Crespi, D'Este, Fontana, & Geuna, 2011).

Particularly patents and license agreements are often data of choice for estimating the true economic value of scientific and technical research outcomes (Bodas Freitas et al., 2013; Thursby, Jensen, & Thursby, 2001). Patents and/or licensing agreements are employed to assess the magnitude of knowledge utilized by firms. However, patents, licensing agreements, co-publications and the like do not capture all forms of knowledge exchange by far. They are mainly the most used proxy indicators due to their availability and international comparability (Thursby & Thursby, 2002). However, these indicators face long-standing criticism as they fail to represent a coherent picture of relevant knowledge spillovers (Cohen et al., 2002; Schartinger et al., 2002) and might not represent all specific aspects of successful commercialization as already stated by Agrawal and Henderson (2002). These indicators alone fail to provide a truly comprehensive picture of the knowledge contribution to the economy and yet the literature is dominated by those traditional measurements. Finding more holistic approaches for quantitative impact assessments of knowledge transfer from universities remains a great challenge.

Given these limitations we aim to provide a first step towards a novel measure that is applicable on a single case basis, which provides in-depth understanding like many qualitative studies do and is at the same time an additional quantitative approach, which provides generalizable and comparable results. We propose a computational linguistic approach for this purpose. The goal is to improve the detection of knowledge transfer without focusing on commercialization's, patents or the formal channels of knowledge transfers. The objective is to verify additional data sources and provide potential new indicators for tracing knowledge transfers from universities to the industry or vice versa.

METHODOLOGY

To compare our text samples from the university (DTU) and its partner or related firms, we chose well-established text-mining methods. Using these methods, we aim to identify new patterns of knowledge transfer, which are undetectable by existing indicators. The general assumption is that not all knowledge is necessarily patented or licensed, but it might be displayed in other texts formats. Hence, we use a method that statistically aims to detect word patterns in texts to identify textual pairs that represent the same or similar knowledge.

The applied method is based on the so-called 'bag of word assumption', which presumes that the words' order in a given document is irrelevant for the statistical analysis. Thus, the order of words in a given document is not taken into consideration and is treated as a set of independent features. Obviously, a document with unordered words will surely not express the same message as an ordered one and the features are by no means totally independent, as particular terms tend to occur more often in the particular documents. Furthermore, these methods assume that documents within a corpus are interchangeable and ordering of the documents in a corpus can be disregarded (Blei, Ng, & Jordan, 2003; Hofmann, 2001). However,

these assumptions do not entail any presupposition about for instance the independence or an identical distribution of the variables. The models operate in the space of distributions over words. Typically, documents are represented as feature-vectors, where a feature corresponds to one word (1-gram) or an ordered combination of words (bi- grams, ..., n-grams) (Berry & Castellanos, 2007). In this study, we focus solely on 1-grams, which limits the analysis because bi-grams like 'home made' or 'top ten' are divided in their single components and not identified as contextual unit.

Document-term matrix

The most common vector space representation of a document corpus is a document-term matrix, which contains feature (terms) frequencies associated to each document. Their rows correspond to documents and their columns to terms. The motivation is to achieve a representation of frequencies of semantically and contextual significant terms (Merritt, 2010). These matrices are commonly highly dimensional and sparse matrices (Berry & Castellanos, 2007). There are various schemes for determining the value that each entry in the matrix can take, depending much on the models used (Salton 1988).

In a term-document matrix, the element at (i,j) is the word count (frequency) of the i'th word (t) in the j'th document (d):

$$Document - Term\ Matrix = \begin{matrix} \\ d_j \end{matrix} \begin{matrix} t_i \\ \begin{pmatrix} x_{1,1} & \cdots & x_{1,i} \\ \vdots & \ddots & \vdots \\ x_{j,1} & \cdots & x_{j,i} \end{pmatrix} \end{matrix}$$

Word count (frequency) is sometimes modified and weighted for a better representation of the relevant feature of each document. Common weighting schemes include:

- Binary weighting, representing whether or not a term occurs in a document;
- Term-frequency weighting (TF), based on the number of occurrences in a document;
- Term-frequency inverse document frequency weight (TFIDF), using TF but assigning higher weight to terms that occur only in a small number of documents.

In our case, we converted all the single text corpora into document-term matrices applying (normalized) TFIDF weighting.

We additionally applied additive filtering of words not relevant to the context of a document by completely removing words that would occur in more than a certain percentage of documents in a corpus. The percentage was arbitrarily adjusted according to the method used, by assessing the outcome of the models and adjusting until obtaining satisfactory results.

TFIDF

This method is a numerical method used in various contexts and applied in text mining to calculate an order of content relevant words for documents. It is applied for text classification, summarization or content identification (Zhang et al., 2016). In order to identify commonalities between two documents, we used the TFIDF indexing to determine most characteristic words per document. These words can be regarded as key words describing the content of a document. The TFIDF indexing increases the value of the most relevant features of each document and devalues the feature occurring in more than a few documents.

TFIDF does not account for any synonymy or similarity and is purely bound to individual words, identifying only limited concepts of texts.

Different weighting calculations are possible for TFIDF indexing, but we opted for the most common weighting scheme, which additionally provides some normalization due to the included log transformation. For $t_i \in d_j$,

$$tf(t_i, d_j) = \sum t_i$$

We further have

$$idf(w, D) = \ln\left(\frac{N}{|\{d \in D : w \in d\}|}\right)$$

With N: Total number of documents and $|\{d \in D : w \in d\}|$: number of documents containing the word w. Finally, the TFIDF is obtained with the following multiplication:

$$tfidf(w, d, D) = tf(wd) \times idf(w, D)$$

We found that the representation of the keywords per document was improved for our comparison purposes, when performing the calculation on two separate corpora coming from two different sources. Both text sources do not have the same writing style. On one hand, websites contain a lot of spoken language and noise around the actual information. On the other hand, abstracts from publication papers are dense literature language. Hence, we chose this unusual approach of having two separate corpora for key word extraction.

Obviously, certain similarity measures could not be applied due to the two instances of word score calculation. We decided to include a maximum of 50 highest scoring terms per document. Reducing the dimensionality of documents to a binary list of maximal 50 terms enabled a comparison of keyword lists with each other. The TFIDF is a comparatively basic method, but is computationally economical and gives proficient results for any further analysis. Especially with short abstracts texts, the TFIDF keyword retrievals often resulted in lists shorter than five words, which needed to be considered for the later comparison.

Jaccard Similarity Coefficient

For the similarity measure between the two sets of identified keywords found thanks to the TFIDF, we used the Jaccard similarity coefficient as the metric. It is a statistic used for

measuring sets similarity. The Jaccard similarity is the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, one indicating most similarity (identical sets) and zero indicating least similar (no common feature in the two sets).

Given the set of keywords from one document of the publication database denoted K_A and the second set of keywords from one page of the websites denoted K_B, the Jaccard similarity denoted $J(K_A,K_B)$ is obtained with:

$$J(K_A, K_B) = \frac{K_A \cap K_B}{K_A \cup K_B} = \frac{K_A \cap K_B}{|K_A| + |K_B| - |K_A \cap K_B|}$$

We chose this similarity measure as it only includes occurrence and leaves order or values aside. The advantage is the low computational expense. This makes it attractive for a basic similarity assessment, which can of course be refined, by applying additional similarity measures to find more accurate matches.

The thresholds for a minimum similarity chosen for further examination were chosen based on brief manual investigation; meaning that we would only consider keyword lists with minimum Jaccard similarity values relevant enough for the manual inspection and potential matching. However, we observed that the Jaccard similarity tends to give better scores to small sets. For example, a 2 words intersection out of two sets of 3 words gives a very high Jaccard similarity (0.5) but is probably not indicating more related content than a 25 words intersection out of 50-words sets (0.33). Hence, we decided to set a common threshold to a minimum of 0.13 and another used indicator threshold consisted in multiplying the Jaccard index with the intersection of the two sets, giving higher weight to sets with a large intersection (higher amount of common words). The number of common words was multiplied with their Jaccard Similarity and needed to exceed 0.15×7, representing approximately 7 words intersection with Jaccard index of 0.15, approximately 7 common words out of 26-words sets. Thus, set pairs with Jaccard

Similarity lower than 0.15 need an higher than 7-word intersection in order to pass the criteria, while set pairs with Jaccard Index higher than 0.15 can have a lower than 7-word intersection in order to pass the matching criteria.

SAMPLES

The next section outlines steps undertaken for the generation of the text samples. The outline is divided into the generation of the text collections, representing university and industry knowledge and to identify common knowledge.

This study is using the case of the Technical University of Denmark (DTU) as scope of the study. Two main data sources are used in this study.

The first source is the university publication database named Orbit. The data set, provided by Orbit, contains a collection of research publication abstracts. These abstracts present main research outputs by employees of the DTU between 2005 until 2016. The database provides, among other information titles, keywords, author information and in most cases abstracts. Given the challenges to obtain a comprehensive sample of full text publications, abstracts were chosen as proxy of the universities research output, although this will not reflect the complete output.

The second data source, giving information on company knowledge and innovations, was gathered from firm websites. Selection criteria for the companies were (i) an English version of at least part of the website, (ii) a national branch of the company, and (iii) at least one common partner with the university.

Following these criteria the sample was produced using a hyperlink network from the university to its partners and partners of partners.

Publication Database (Orbit)

The selected data set from Orbit included all entries from January 2005 until August 2016, which resulted in a total of 76,627 publication entries. Of these entries, 43,745 included a full abstract, which were then categorized by research area and combined accordingly into separate corpora. This division of fields improves the later statistical analysis by dividing meaningful subsets for the data structure. Furthermore, computation time is reduced if a measure is only applied to smaller subsets of the data. The division resulted in 24 separate fields, which were aligned to department codes, provided by the database. Three of these subfields were irrelevant for the academic output of the university: (i) Publications registered to the university administration, (ii) publications registered to the bachelor program, and (iii) one set that was directly linked to a large company (this might have biased the findings significantly as the firm is directly involved in several hundreds of specially dedicated publications).

The remaining 21 fields are Electrical Engineering, Management Engineering, Physics, Compute, Chemistry, Mechanical Engineering, Environmental Engineering, Energy Conversion and Storage (EngConSto), National Food Institute, Nuclear Technologies, Aquatic Resources, Photonics, National Space Institute, Micro and Nanotechnology, Biochemistry, National Veterinary Institute, Civil Engineering, Wind Energy, Transport, Biosystems and Diverse[1].

These corpora will in the following be referred to as 'academic' corpora or by their individual name in case this is relevant for the interpretation of the results.

Firm Webpages

To identify the relevant firms for the firm based sample, we generated a simple directed network based on the relationships of the university with companies. A first network was

[1] This corpus contains publications, which do not fall under any of the above-mentioned categories.

generated on the basis of hyperlinks between webpages using the university as point of departure, (denoted Sample A). While an additional network was generated using university contracts to identify collaboration partners of the university and their partners (denoted Sample B). All companies connected to the university via hyperlinks and their direct partners were identified and stored, which resulted in a directed un-weighted second-degree network. The identified pages were downloaded and stored as HTML files.

The collected files were subsequently scanned for a Danish firm registration number and added to the text samples only if one was found for each given website. In a following step, the language of the page or the subpages was verified and only the English[2] content was stored. The online text samples were collected during August 2016 and September 2016[3]. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample, to avoid unnecessary pages and unrelated hyperlinks. In Denmark, universities are registered as companies and therefore have a Company registration number (CVR); so they had to be manually excluded.

Sample A

The first network contained 177 nodes, which represent individual company websites. These are connected to the university within a range of a path length of two, meaning that each node is either directly or over a common partner connected to the university page. The hyperlink network shows clear tendency to build clusters and it has some particularly central nodes. The nodes, which are highly interconnected and central for the structure of the network are mainly

[2] Danish firms provide a great amount of their information in English and the academic abstracts are in English, which enables a comparison, based on keywords between Danish firms and Danish university research in English.

[3] The script used to identify and download the pages can be found at https://github.com/nobriot/web_explorer

online service platforms, including transportation and types of yellow pages and firm registries. The texts from this network contain overall around 120,000 unique terms. We assigned each text to its website URL, resulting in 121 single text corpora based on individual websites (with up to 1000 webpages). 56 smaller websites had less than 5 pages after language filtering and were combined to one single corpus, as these would be too small to apply the relevant statistical analysis, as they are mainly composed of brief introduction pages of the home pages, not containing any relevant information.

During the network generation it became apparent that many official partners are not necessarily connected with a hyperlink to the university main pages. We included the web Sample B to account for this.

Sample B

To generate an additional sample another network was created based on Danish companies with a formal connection to the university, namely a collaboration contract. Hence, we commenced building the second network with around 686 first-degree firms, which had a contract with the university between the years 2013 and beginning of 2016. Those new websites were collected and their online partners were also identified. This generated a fully new network including more content related companies. The identified firms operate mainly in technology intensive sectors and are firms with strong R&D divisions.

The second network contained 686 nodes and of which 312 were identified as Danish companies. This sample, resulted in 243 single text corpora, based on individual websites (with up to 1000 webpages) and an additional corpus again containing 69 smaller pages. For the later analysis we will refer to the sample that is solely based on hyperlinks as sample A and the sample including internal contract information as sample B.

Pre-processing

Text pre-processing describes the task of converting unstructured raw text into an order of computationally and statistical useful and linguistically meaningful units. The pre-processing is an essential part of any text analytical procedure, since the characters and words are identified at this stage as the units passed on to further text mining stages (Paukkeri & Honkela, 2010).

Pre-processing of text, which is also known as tokenization includes in our case the following steps:

- Define word boundaries as white spaces.
- Remove unessential elements (e.g. coding tags, punctuation, and numbers).
- Convert all characters to lower case (makes the identification of abbreviations challenging).
- Strip the texts from additional white spaces.
- Remove stopwords, meaning most frequent words, which do not carry content information (in some cases, topic specific stopwords were added).
- Apply stemming which is beneficial to merge the inflected word forms into the corresponding stem.

Results of this pre-processing revealed some challenges especially for the academic abstracts. For instance chemical formulas and similar notations rely on numbers, short abbreviations and punctuation. So after pre-processing the only possibility to identify the concurrent formulas would be the prospect that the removal of numbers and punctuation results in the same string in both types of texts that can be seen as an equivalent to a term representation of the formula. Additionally, some very specific abbreviations are sometimes hard to identify, meaning that the results of the tokenization does not seem to make much sense, but are actually

describing very particular features of some publications (e.g. omniitox, which is name of a European project, or 'modelpbpk', standing for PBPK modeling). Finally, we merged the pre-processed texts into text corpora, which are a large ordered set of documents, to ensure structured sets of texts.

RESULTS

The described TFIDF indexing was used to assess the documents' similarity. We divided the results into the two web data samples for illustration. The results vary greatly due to the high diversity of the text corpora from the firm samples. After all pre-processing steps the sample A encompassed 117 websites containing 30,241 single pages and sample B with 243 websites and 77,421 pages.

We classified the found text pairs or matches into 5 main categories:

- 1st order: Web texts which are related to a university publication
- 2nd order: Web texts which are very likely to be related but miss an actual clear link
- 3rd order: Web texts which clearly come from the same area, but concern a different sub-field of the area
- 4th order: text pairs that contain similar topics but there is no deeper connection
- 5th order: text pairs with no overlap at all.

It has to be remarked that the pairing of the web text files and the abstracts resulted in several recurrent hits, meaning that the overall number of different pairs is significantly lower than the raw found matches, due to the fact that companies often display the same text content on more than one page. However, still one page could have several hits, so we excluded pairs, which represented the same website and the same abstract, but a different page from the website.

We decided to perform the manual investigation on the original texts, without any pre-processing to ensure that the actual content of the documents was understood.

TFIDF results for the academic corpora

The application of the TFIDF indexing on the 21 academic corpora resulted in a given set of key words for each document. Several academic expressions were hereby filtered out and context relevant words were identified. Table 1 shows the 5 most relevant words for each university department.

Insert Table 1 about here

These words represent the content of the departments satisfactorily considering the exclusion of too recurrent words. A manual inspection of the sample confirmed an adequate representation of keywords on corpus (departments) and document level. However, the collected abstracts were relatively short (4-6 sentences), which limited the content and representation of keywords per se. The same comprehensiveness of presentation of keywords accounts for the websites.

Results of the Comparison with Sample A

In the following we compared each keyword set from any website with the keywords of each abstract in every academic corpus. This led overall to 1,306,139,031 comparisons. For the chosen threshold for the Jaccard similarity (see Methodology section), 385 document pairs were considered as matching documents (including all pairs). The matching rate of relevant pairs was 2.9×10^{-5} %. The highest scoring pair reached 0.235 Jaccard Similarity representing in our case 19 common words out of 81 total keywords. As a benchmark, calculating the Jaccard Similarity

between the different abstracts within the academic corpora, given the same threshold, the threshold was exceeded more than 0.009% of the time. The highest Jaccard similarity was in this case close to 1. Showing clearly that the academic corpora documents are found to have more in common among each other than with the sample A.

The average Jaccard similarity for matches in the sample A was 0.125, which is rather low. Only 22 pairs exceeded 0.15 Jaccard similarity. The identified pairs were in the following manually examined. . Highest Jaccard similarity scores were dominated by a word co-occurrence of country names, which is likely to be only of limited contextual relevance. Additionally, some text pairs were identified as similar due to a common foreign language, which was detected in both texts like for instance parts of German or Danish. Indeed many similar pairs, show that the dominating attributes were country names, but that among the top ten pairs were some in which the common words with more content relevance as shown in Table 2.

Insert Table 2 about here

With a manual inspection of the found pairs we found a limited number of common contents and DTU related research content. We found the following classifications:

- 1st order matches: 4
- 2nd order matches: 2
- 3rd order matches: 10
- 4th order matches: 4
- 5th order matches: 5

There were no 1st order or 2nd order pairs identified below the Jaccard Similarity threshold of 0.130. It should be mentioned that this sample contained a considerable number of

1st order pairs (16), which were representing websites of public entities, which in some cases were even part of the university itself. Hence, these pairs were subtracted from the overall 1st order pairs. However, these were correctly identified pairs. The overall correct identification would therefore be 20 correct identified pairs. Eliminating the country pairs and hits under 0.130 we have 41 relevant pairs left and the 1st and 2nd order pairs were 53.65% from the overall findings. The common contents were mainly related to system inventions, or presentations given by DTU employees and mentioned on the respective websites.

Comparison of Sample B

For each page per website of sample B, we calculated the keywords via the TFIDF indexing and compared with the academic keyword sets. In the case of sample B this accounted for 3,343,890,411 compared pairs and 974 of them passed the chosen threshold. This is again a percentage of 2.9×10^{-5} % found pairs, which is identical to sample A's matching rate. This resulted in 25 text pairs scoring a Jaccard Similarity over 0.15 but none over 0.18, which is lower than Sample A's result. The average Jaccard similarity was 0.121 for found matches, which was lower than the one from sample A.

Most common words were more diverse than the ones of sample A. The resulting matches of keywords consisted of words that have more content relevance, however the highest pairs are still consisting country related words (refer to Table 3).

Insert Table 3 about here

The manual verification of the text pairs revealed that the matches scoring under 0.130 Jaccard similarity are definitely less relevant and contain mainly 4th order pairs than the pairs that

exceed this threshold. After removing all pairs under 0.130 Jaccard Similarity and excluding the country pairs we were left with 89 relevant pairs. We identified the following numbers for the classes of the text pairs:

- 1st order matches: 13
- 2nd order matches: 10
- 3rd order matches: 22
- 4th order matches: 23
- 5th order matches: 16

This means that 27.38 % of the matching pairs were clear references to the university knowledge or were highly likely related. We had 5 pairs (5.95%), which we could not clearly classify as the information provided by the abstract was too limited, or the content too specific and would require an expert opinion of the specific field. Only 19.05% were pairs that have no overlap and were wrongly identified.

DISCUSSION

Generally is evident that the results from sample A and B vary in their quality (text content) and quantity. The most relevant matches 1st and 2nd order describe clearly the use of common, partly by the university invented methods and their direct application. Three of the websites state the university as source of these methods or tools. Some of the matches are towards the same website but identify different contents, so one site is responsible for 4 of the 1st order matches. Within the 1st order we found one match where the company that does display the content refers to another company with which the university has the topic related contracts and the content matched extremely well. In other cases, parts of the actual abstract are directly quoted, but without a clear reference to the university.

The 2nd order pairs show often a strong overlap in scope content and used methods, but lack a clear verification or linkage to the university, which the 1st order pairs contain.

Sample A's 1st order pairs were mainly a clear display of research results either on the pages of other public entities, conference summaries or similar. Resulting in the identification of clear related, but in terms of commercial use and knowledge transfer maybe not very relevant. Sample B's 1st order pairs are dominated by the use of university developed tools and models and are therefore extremely relevant in terms of our research objectives.

Given that sample A is a sample containing mainly websites that are not related to any of the university's research this is a positive outcome, as it verifies that the method finds communalities where there are some present. Generally, the performance of this simple measure is comparatively successful as it succeeds in identifying knowledge overlaps.

A further confirmation is the significantly higher number of commonalities among the academic keywords than between websites and academic corpora, even though they refer often to different topics, especially since a technical university as such has a great overlap among the research fields. In sample A, many pairs were correctly identified but the identification of purely private enterprises was not impeccable. The comparatively small number of 1st and 2nd order pairs show that there would be additional identification mechanisms suitable to obtain more results. However, it shows that the pairing can identify the use of university related knowledge and even the use of university created knowledge.

The high number of 3rd and 4th degree order in sample B represents companies that use the common contents like particular models, instruments, or metrics in the same or closely related fields, but are rather unlikely connected to the university's research.

The performance of the TFIDF indexing, especially given the benchmark comparison matching between the different academic corpora, shows that it identified 186.414 pairs that reach the threshold even though the abstracts are significantly shorter than the webpages, which means the quantity of text for matching is reduced and findings should be less. Some more trails to find optimal thresholds need improvement and additional randomized testing is necessary, but the results are promising.

CONCLUSION

This study provides a first attempt to develop an additional measure of knowledge transfer by using texts as main data sources. Our test case shows that the identification of university knowledge in firms' websites is clearly possible by applying the given statistical measures. We examined two different samples of websites and our results suggest that our approach does work for formal as well as for informal or second-degree partners of the university. The overall outcome identifies common grounds between companies and the university.

We can identify texts that show on the one hand either a clear relation to university knowledge and furthermore identify the companies that deal with very related topics. This can be used to identify the universities knowledge transfer and additionally most common areas of interests from universities and companies. We see this as a great step towards the actual detection of knowledge spillovers and transfer, even though it is certainly just an addition to current metrics.

Limitations

The text samples of firm websites for the study are not exhaustive as especially PDF formats and similar were not yet included in the sample. Additionally an additional identification

of Danish firms would be beneficial. Regarding the representation via abstracts of publications must be said that the availability of full text would have been beneficial especially since the content of academic abstracts is per se very limited.

Finally, the TFIDF indexing is a rather simple method, which is incapable to capture contexts, meaning that in case different words are used to describe the same subject this method would fail to identify a connection.

Future research

Next steps for the improvement of this approach are to increase the quality and quantity of the text data, by gaining access to full text publications and potentially annual reports from relevant firms. For future research we also aim to provide automated classifications into the 5 classes, which will only have to be verified by humans to decrease the amount of manual labor. We aim to combine our approach it with additional statistical approaches to increase the performance. Concurrent machine learning approaches will come in handy and enable us to enhance the current results. Ideally we will be able to test our next results against the outcome of traditional metric.

REFERENCES*

Agrawal, A., & Henderson, R. (2002). Putting Patents in Context: Exploring Knowledge Transfer from MIT. ***Management Science***, 48(1): 44–60.

Agrawal, A. K. (2001). University-to-industry knowledge transfer: literature review and unanswered questions. ***International Journal of Management Reviews***, 3(4): 285–302.

Ankrah, S. N., Burgess, T. F., Grimshaw, P., & Shaw, N. E. (2013). Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. ***Technovation***, 33(2–3): 50–65.

Berry, M. W., & Castellanos, M. (2007). ***Survey of Text Mining II: Clustering, Classification, and Retrieval***. London: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. ***Journal of Machine Learning Research***, 3(4–5): 993–1022.

Bodas Freitas, I. M., Marques, R. A., & Silva, E. M. D. P. E. (2013). University-industry collaboration and innovation in emergent and mature industries in new industrialized countries. ***Research Policy***, 42(2): 443–453.

Cheah, S. (2016). Framework for measuring research and innovation impact. ***Innovation***, 18(2): 212–232.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and Impacts : The Influence of Public Research on Industrial R & D. ***Management science***, 48(1): 1–23.

Crespi, G., D'Este, P., Fontana, R., & Geuna, A. (2011). The impact of academic patenting on university research and its transfer. ***Research Policy***, 40(1): 55–68.

D'Este, P., & Patel, P. (2007). University-industry linkages in the UK: What are the factors underlying the variety of interactions with industry? ***Research Policy***, 36(9): 1295–1313.

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.

Feller, I. (1990). Universities as engines of R& D-based economic growth: They think they can. ***Research Policy***, 19(4): 335–348.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. ***Machine Learning***, 42: 177–196.

Howells, J., Ramlogan, R., & Cheng, S. L. (2012). Innovation and university collaboration: Paradox and complexity within the knowledge economy. ***Cambridge Journal of Economics***, 36(3): 703–721.

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.

Merritt, D. (2010). ***Adventure in Prolog***. New York: Springer

O'Shea, R. P., Chugh, H., & Allen, T. J. (2008). Determinants and consequences of university spinoff activity: A conceptual framework. ***Journal of Technology Transfer***, 33(6): 653–666.

Paukkeri, M., & Honkela, T. (2010). Likey : Unsupervised Language-independent Keyphrase Extraction. ***Proceedings of the 5th International Workshop on Semantic Evaluation***: 162–165.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., et al. (2013). Academic engagement and commercialisation: A review of the literature on university-industry relations. ***Research Policy***, 42(2): 423–442.

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: Impacts of university-industry relations on public research. ***Industrial and Corporate Change***, 18(6): 1033–1065.

Rosenberg, N., & Nelson, R. R. (1993). American universities and technical advance in industry. ***Research Policy***, 23: 323–348.

Rothaermel, F. T., Agung, S. D., & Jiang, L. (2007). University entrepreneurship: A taxonomy of the literature. ***Industrial and Corporate Change***, 16(4): 691–791.

Schartinger, D., Rammer, C., & Fröhlich, J. (2002). Knowledge interactions between universities and industry in Austria: Sectoral patterns and determinants. ***Innovation, Networks, and Knowledge Spillovers: Selected Essays***, 31: 135–166.

Siegel, D. S., Waldman, D. A., Atwater, L. E., & Link, A. N. (2004). Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: Qualitative evidence from the commercialization of university technologies. ***Journal of Engineering and Technology Management***, 21(1–2): 115–142.

Teixeira, A. A. C. (2014). Evolution, roots and influence of the literature on national systems of innovation: A bibliometric account. ***Cambridge Journal of Economics***, 38(1): 181–214.

Thursby, J. G. J. J. G., Jensen, R. a., & Thursby, M. C. M. (2001). Objectives, characteristics and outcomes of university licensing: A survey of major US universities. ***The Journal of Technology Transfer***, 26(1): 59–72.

Thursby, J. G., & Thursby, M. C. (2002). Who Is Selling the Ivory Tower? Sources of Growth in University Licensing. ***Management Science***, 48(1): 90–104.

Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. ***Technological Forecasting and Social Change***, 105: 179–191.

TABLE 1*

Most relevant words for each DTU department

Department (corpus)	Most relevant words				
Compute/Math	attack	secur	graph	network	code
ChemBiochem	enzym	polym	membran	oil	catalyst
Chemestry	hydrogen	zeolit	liquid	membran	hydrogen
CivilEng	solar	crack	collector	moistur	stress
ElectEng	antenna	convert	fault	robot	flow
EngConSto	magnet	membran	carbon	anod	field
EnviEng	landfil	sludg	methan	bioga	climat
MAN	servic	network	materi	configur	risk
MechEng	weld	stress	steel	wind	bear
MicroNano	magnet	graphen	cantilev	laser	reson
PhotoEng	quantum	thz	dispers	data	convers
Physics	nanoparticl	pbri	water	mode	plasma
BioSys	biofilm	peptid	resist	dna	aeruginosa
Transport	brbr	til	der	ship	capac
Wind	ref	composit	instal	fibr	accord
Food	efsaq	claim	substanc	salmonella	vitamin
Aqua	egg	prey	migrat	codend	genet
Space	burst	graviti	mcrab	cluster	nustar
Nuc	msupsup	neutron	iodin	supsupi	risø
Vet	resist	serotyp	intestin	fmdv	genotyp
Diverse	magnet	film	grain	turbin	electrod

TABLE 2*

Word co-occurrence for Sample A

Words and their co-occurrences (100 top-words)							
latvia	103	research	23	dtu	10	phospholipid	8
hungari	103	slovakia	23	factor	10	enzym	8
cyprus	102	support	18	phospholipas	10	die	7
bulgaria	102	sweden	17	ischem	10	experiment	7
lithuania	100	renew	16	european	10	plant	7
estonia	99	electr	16	fibril	10	digest	7
finland	91	technolog	15	atrial	10	nation	7
greec	88	risk	14	obes	10	qsar	7
slovenia	88	energi	14	industri	10	procedur	7
czech	88	student	14	stratif	10	sustain	7
republ	88	grid	13	cost	9	ist	7
romania	81	ion	13	physic	9	microscopi	7
technic	49	consumpt	13	fuel	9	knowledg	7
univers	47	fast	13	young	8	databas	7
denmark	41	scatter	12	hydrolysi	8	dynam	7
engin	39	thomson	12	austria	8	und	7
electron	38	collect	12	den	8	countri	7
list	37	power	12	emiss	8	interest	7
sourc	36	der	11	earth	8	properti	7
issu	33	wind	11	liposom	8	ein	7
publish	33	suppli	11	comment	8	von	7
depart	32	gas	11	member	8	programm	7
note	32	learn	10	secretori	8	pretreat	7
luxembourg	26	coronari	10	netherland	8	specif	7
ireland	24	myocardi	10	bioga	8	storag	7

TABLE 3*

Word co-occurrence for Sample B

Words and their co-occurrences (100 top-words)							
electr	91	heat	39	properti	28	ist	24
cycl	84	caus	39	market	28	amplifi	24
fuel	80	sourc	37	spot	28	med	23
der	78	stress	37	damag	28	month	23
environment	70	den	37	document	28	des	23
impact	64	von	36	failur	28	sustain	23
solar	61	mit	35	coal	28	layer	23
die	61	advanc	34	das	28	countri	22
renew	60	werden	34	nois	28	som	22
life	60	depend	33	turbin	27	cost	22
assess	60	tension	33	review	27	consum	22
und	59	deform	32	econom	27	conduct	21
gas	55	analys	32	characteris	27	har	21
fossil	54	mass	32	fibr	27	produc	21
wind	53	emiss	32	obes	27	storag	21
lca	52	auf	31	gain	27	decis	21
temperatur	52	biomass	31	creat	26	electrochem	21
greenhous	49	calcul	30	figur	25	manag	21
power	49	degrad	30	til	25	equat	21
grid	48	energi	30	global	25	growth	21
für	47	mechan	29	resourc	25	sector	20
ein	47	consumpt	29	suppli	25	smart	20
demand	47	determin	29	technolog	25	index	20
plant	41	weld	29	denmark	25	ion	20
climat	40	futur	29	transport	25	averag	19