Tracing university–industry knowledge transfer through a text mining approach

Woltmann, Sabrina L.; Alkærsig, Lars

Published in:
Scientometrics

Link to article, DOI:
10.1007/s11192-018-2849-9

Publication date:
2018

Document Version
Peer reviewed version

Citation (APA):
Woltmann, S. L., & Alkærsig, L. (2018). Tracing university–industry knowledge transfer through a text mining approach. *Scientometrics*, *117*(1), 449-472. https://doi.org/10.1007/s11192-018-2849-9

1 Introduction

Impact is of increasing importance for universities in addition to traditional tasks of research and teaching. This Third Mission means that many universities expand their efforts beyond the production of knowledge to translate it into socioeconomic relevant contributions(D'Este and Patel, 2007; Etzkowitz et al, 2000). Driven by the need to make their research known in order to secure (public) funding, universities implement various forms of transfer activities, such as adaption of strategic licensing and university patenting, which ensure that their findings are (commercially) utilized (Gulbrandsen and Slipersaeter, 2007). However, the detection of relevant knowledge transfer remains non trivial. Thus, new methodologies are needed to quantitatively assess knowledge transfer of universities to enable a more holistic analysis. This paper examines the transfer of university research to the industry through a novel combination of methods based on established text mining tools. We use additional data sources and metrics, to move beyond the traditional proxy indicators. We aim to identify the potential and limits of contemporary text mining tools for a detection of identical knowledge pieces. Text documents are already used in numerous studies as data sources and are, hence, suitable to answer relevant present-day questions (Zhang et al, 2016). Our approach is unique, since it captures identical knowledge pieces in the university and the industry in its (geographical) proximity. The intention is to capture the transfer without focusing on specific transfer channels, collaboration types or related commercialization mechanisms. The knowledge detection is made through the application of existing text mining methods, namely the latent dirichlet allocation (LDA) and the algebraic indexing method called term-frequency, inverse document frequency (TFIDF). LDA is a known topic model, used to identify underlying structures in entire text collections, while TFIDF indexing can be used to extract keywords for single documents. We use a combination of both methods to identify *identical* knowledge pieces.

2 Literature

Knowledge transfer concerns "(...) the conveyance of knowledge from one place, person or ownership to another. Successful knowledge transfer means that transfer results in successful creation and application of knowledge in organizations "Liyanage et al (2009)[p. 122], including the necessity of utilization of this particular knowledge.[1] Given the particular role of universities within the field of knowledge transfer, a great deal of literature has established a well developed empirical basis for the assessment of university driven knowledge transfer (Agrawal, 2001; Perkmann and Walsh, 2007). The empirical approaches are often derived from integrated models on the institutional

[1] Technology transfer and knowledge transfer are in the literature strongly interrelated concepts and are widely used as interchangeable terms (Grimpe and Hussinger, 2013; Sung and Gibson, 2000).

level, such as the triple helix model (Etzkowitz and Leydesdorff, 2000a), which are regularly reduced to a bilateral university-industry focused concept that investigates collaborations between universities and firms on individual, organizational or national level (Siegel et al, 2003; D'Este and Patel, 2007).

Additionally, the research is also divided into formal and informal knowledge transfer. Formal transfer will eventually "result in a legal instrumentality such as, for example, a patent, license or royalty agreement (. . .)" (Arundel and Marcó, 2008, p. 642), while informal transfer is seen as resulting from informal communication and does not lead to outcomes that fall under intellectual property regulations (Tijssen et al, 2009; Link et al, 2007). Overall, the main attention in the literature on university-industry knowledge transfer has been given to formal knowledge transfer often focusing on the commercial value the knowledge yields (Thursby et al, 2001; Wu et al, 2015; Han, 2017). Due to the abstract nature of knowledge transfer its actual measurement remains challenging and relies heavily on proxy indicators. The indicators for formal (commercialized) knowledge transfer, even though well developed, fail to measure instances where knowledge cannot easily be commercialized, patented or licensed (Cheah, 2016; Cohen et al, 2002; Agrawal and Henderson, 2002). These circumstances have left the research community with a gap in tracing and measuring the university-industry knowledge transfer (Sung and Gibson, 2000) and the need to investigate and assess potential new methods.

2.1 Text mining: empirical applications

One of the contemporary approaches to solve various kinds of measurement or detection challenges is the application of data mining or in particular text mining (Aggarwal and Zhai, 2012). In this regard, computational linguistics, the scientific base of text mining, became increasingly relevant for empirical studies in a number of unrelated academic fields (Yau et al, 2014; Aggarwal and Zhai, 2012; Gaikwad et al, 2014). Previously great insights in disciplines like social sciences, biology, and economics have been achieved through the use of text mining tools (Zhang et al, 2016; Garechana et al, 2017). Text mining applications have also gained traction within research concerning knowledge networks and knowledge flows (Magerman et al, 2010; Leydesdorff, 2004). In studies investigating the influence on knowledge generation and dissemination of universities, the triple helix model (Etzkowitz and Leydesdorff, 2000b) is often used as a foundation to unveil concrete knowledge linkages (Meyer et al, 2003). These studies aim to measure the underlying structures of the (knowledge driven) relationships between governments, academia, and industry and apply regularly text mining based measurements (Khan and Park, 2011). These contemporary text mining applications are often used in combination with other bibliometric tools. An evaluation of university-industry interaction can, for instance be done through and the identification of key words and co-occurrences (Khan and Park, 2011). So today's understanding of the triple helix interaction has been immensely increased by relying on bibliometric, network and text mining approaches (Glänzel and Thijs, 2012; Zhang et al, 2014).

However, approaches on tracing knowledge transfer remain at a relatively rudimentary level. Even though some studies show the successful application of text mining methods (Van Eck and Waltman, 2017; Tussen et al, 2000), the concrete outcomes remain often undetected. The application of these methods also remain challenging (Meyer et al, 2003). The main challenges today include the identification of the actual contributions of university research without limiting analyses to too narrow indicators or being to imprecise. Often only trends are detected, since measures like citations and references do not hold up well in an industry context (Jaffe et al, 2000). Hence, new detailed measurements for knowledge transfer are needed. Our study provides an assessment of the use of text mining methods to extract relevant pieces of knowledge from universities and identify them within companies' public documents. The contribution and innovative approach of this study is to identify the concrete pieces research, such as the results of an experiment or a novel method, from a university publication base and trace them.

3 Methodology

We focus on knowledge transfer overall, which particularly includes the aspects of technology transfer.

Our approach is different to conventional knowledge flow detection in the sense that we aim to identify concrete research outcomes including for instance a concrete technology, method, algorithm, chemical formula etc. and focus less on similar working fields or just coherent topics. This focus makes the actual identification and verification more challenging in a technical sense. We use a combination of two well known techniques the latent dirichlet allocation (LDA) and the term-frequency, inverse document frequency (TFIDF). This combination allows the extraction of relevant keywords per text and also for entire text collections, which allow a keyword comparison.

Generally, text mining can be used to describe the extraction of knowledge from free or unstructured text. This encompasses everything from information retrieval to text classification and clustering (Kao and Poteet, 2007). Current rapid developments in computational linguistics provide improved accuracy and feasibility (Chapman and Hall/CRC, 2010; Collobert et al, 2011). The task of identifying content similarity, however, remains up until today challenging. In particular as the similarity between linguistically highly diverse texts remains widely unsolved.

In the following, we outline the particular methods and algorithms used to fulfill the study's objectives. We aim to give insights into current developments in the field as well as determine the used methods and specify, parameters and tools used. We aim to trace concrete research outcomes from the university, which requires key word extraction, comparison, efficient pattern recognition and similarity measures.

Identifying similarities within texts is a very particular area of text mining. Similarity measures can be based on probabilistic as well as algebraic models. However, these practices are often used to detect actual paraphrasing and these models are limited to identify word-to-word or phrase-to-phrase similarity (Rus et al, 2013). However, for our purpose these applications are too narrow and focus plainly on the linguistic composition and are only applicable on extremely short text snippets.

We make use some of the same basic tools, but combine them in different manners to identify overlapping content for entire documents.

As we aim at tracing concrete research results from the university, it is necessary to combine the comparison between topics and the TFIDF indexing. Therefore, it is not enough to identify that two corpora (a website and a department) share the same topic, for instance 'wind energy', but that for instance a new assessment model (developed and described in the publication) is used by the company. This insight can only be generated on a document-to-document level, but needs to be supported on a corpus topic level. This is crucial, since there are no other concrete indications for transfer, such as citations or references.

3.1 Pre-processing

To apply text mining procedures, the pre-processing of the data is essential. It entails data cleaning and additionally conversion of unstructured raw text into statistical and computational useful units. The quality of text mining results is highly depending on the thoroughness of the pre-processing. The main objective is to capture relevant characters and erase obsolete items (Paukkeri and Honkela, 2010). We follow the procedures as described by Ponweiser (2012, p 33), i.e.:

- Define word boundaries as white spaces,
- Delete unwanted elements (e.g. special characters, punctuation, ...),
- Convert all characters to lower case,
- Remove stopwords (common words that don't carry content information),
- 'Stemming' words, this reduces words to their morphological word stem (Schmidtler and Amtrup, 2007, 126),
- Remove words that are shorter than three characters.

The pre-processed texts are merged into structured units and, in our case, also thematically classified units, the *text corpora*. To prepare the texts collections into a statistically useful format, the corpora are converted into *document-term matrices*. A document-term matrix is the most common vector space representation of document corpora. Rows correspond to documents and columns to terms. It contains the feature (term) frequencies (number of occurrences) for each document (Richardson et al, 2014; Chapman and Hall/CRC, 2010). These matrices are usually highly dimensional and sparse and accordingly most text mining methods most include dimensionality reduction (Berry and Castellanos, 2007).

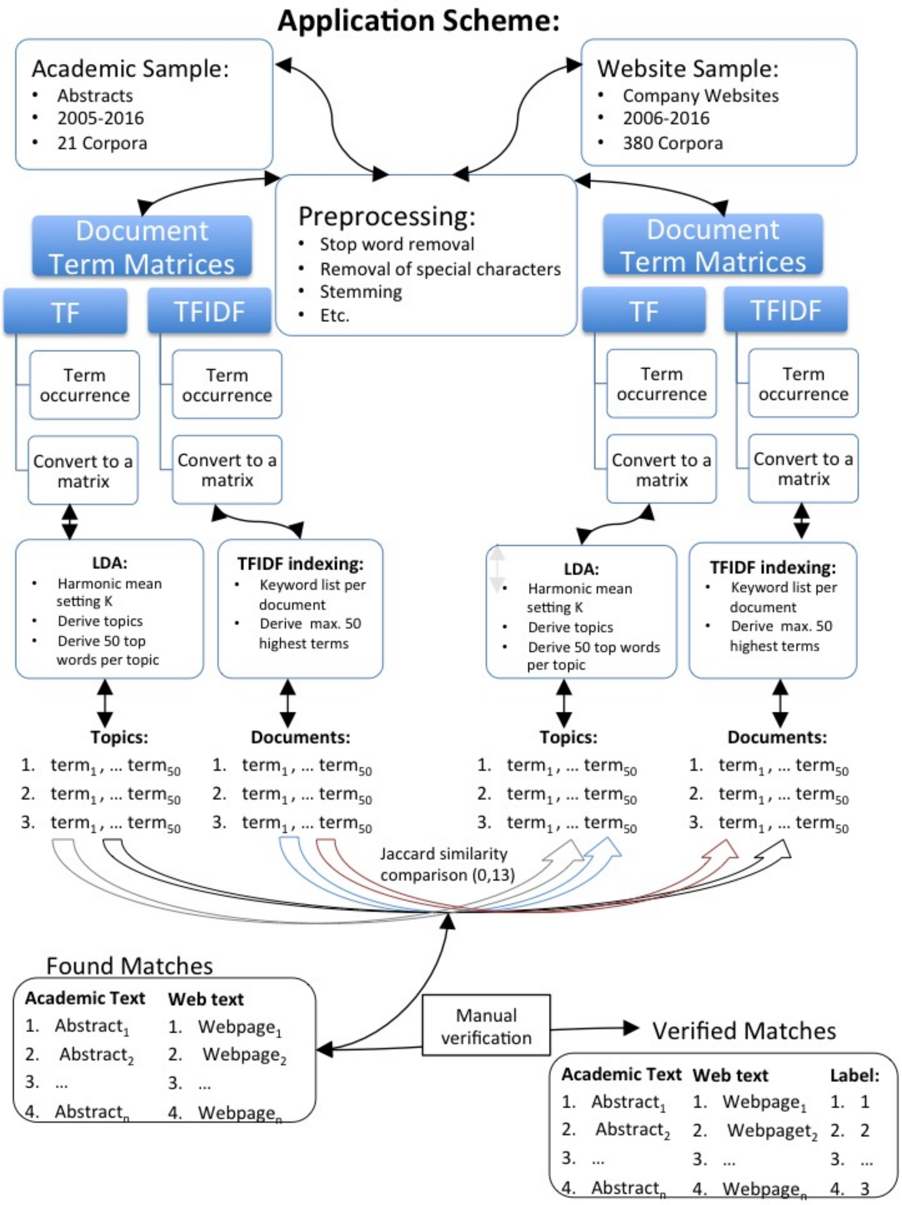

Fig. 1 Steps of statistical method application for the different samples

In a document-term matrix the element at (m,n) is the word count (frequency) of the i'th word (w) in the j'th document (d).

$$Term - Documentmatrix = \begin{array}{cc} & w_n \\ d_m & \begin{pmatrix} x_{1,1} & x_{1,2} & ... & x_{1,n} \\ x_{2,1} & x_{2,2} & ... & x_{2,n} \\ \vdots & & \ddots & \\ x_{m,1} & x_{m,2} & ... & x_{m,n} \end{pmatrix} \end{array} \tag{1}$$

3.2 Term weighting and indexing schemes

Various *term weighting schemes*, determining the value of each entry, are available. The weight for each term can be derived by the application of different measures and is based on the frequencies of term occurrences. Specific text mining models rely on a particular term weighting input (Xia and Chai, 2011).

- Binary weighting takes values 1 or 0 depending on whether or not a term occurs,
- Term-frequency (TF), which is the actual number of occurrences of a term for a given document.
- Term-frequency, inverse document frequency (TFIDF), assigns higher weight to terms that occur in a small number of documents (Xia and Chai, 2011).

The TFIDF is a simple numerical indexing method, which has been applied in various contexts (Franceschini et al, 2016; Zhang et al, 2016) and gives respectable results on its own, but it also serves as basis for various more advanced models, like the Vector Space Model (VSM) or Latent Semantic Analysis (LSA) (Mao and Chu, 2007).

The principal assumption behind the TFIDF is that words that occur often in a document are relevant for its content, but words that are used in many documents are less content specific for the single document. Frequent words that are used in many texts carry less contextual information and obtain a lower score (Robertson, 2004). TFIDF indexing enables a dimensionality reduction providing a small set of content relevant terms. Most commonly the TFIDF is calculated by multiplying the term frequency TF, the number of times word w appears in document d; and the inverse document frequency IDF, which is the logarithm of the total number of documents D divided by the number of documents that contain the word w denote dw (Aizawa, 2003).

$$TF(w,d) = \sum w_i$$

$$IDF(w,D) = log(\frac{D}{dw})$$

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

The TFIDF approach suffers from some shortcomings. First, it might represent only the content of a particular text fragment, which is a major drawback for long texts. Second, IDF assumes that terms, which rarely occur over a collection of documents, are more content related, while in reality they are just more distinctive. Third, empty terms and function terms are often assigned too high scores (Xia and Chai, 2011). Nevertheless, the TFIDF approach has been proven to provide very robust and high quality results (Robertson, 2004).

For the purpose of this study, we use (among other metrics) the TFIDF indexing to determine the most characteristic words for each document. Hereby we reduce the dimensionality and enable a comparison of keyword of different texts with each other. Hence, the lists, generated for each document are used to identify common terms between two types of documents, abstracts and website pages.

3.3 Latent Dirichlet Allocation (LDA)

LDA is an application of topic modeling and is a fully automated method based on statistical learning, which aims to identify latent (unobservable) topical structure in a text corpus (Blei et al, 2003; Griffiths and Steyvers, 2004). LDA extracts underlying structures of texts and translates them into topics, which are composed of terms that are assigned together with a certain probability to each topic.

LDA works as follows, described by Grün and Hornik (2011, p. 4) and Ponweiser (Ponweiser (2012, p.15)):

1. For each topi,. we decide what words are likely (term distribution described as $\beta \sim Dirichlet(\delta)$
2. For each document,
   (a) we decide what proportions of topics should be in the document, (topic proportions defined by $\theta \sim Dirichlet(\alpha)$.
      i. for each word in the document:
         A. we choose a topic ($z_i \sim Multinomial(\theta)$).
         B. given this topic, we choose a likely word (generated in step 1.) from a multinomial probability distribution conditioned on the topic $z_i : p(w_i|z_i, \beta)$.

To improve the performance of the LDA we added one pre-processing step that excluded terms, which occur in more than 90% of the documents in the document-term matrix. The resulting topics are more specified and do not contain generic terms. The LDA algorithm needs to start with a pre-defined number of topics denoted K. Separate approaches were used for estimating K for the academic corpora and for the companies website corpora. For the academic abstracts, K was estimated using the following approach: we approximate the marginal corpus likelihood (depending on K) by taking the harmonic mean for each corpus after applying LDA for different numbers of K. Hereby we are sampling the best 'fit' for a set of possible K values. The harmonic mean takes one chain of samples as argument to first collect all sample log-likelihoods and subsequently calculates the harmonic mean of these likelihoods. The log-likelihood values are determined by first fitting the model and to do this over a sequence of topic models with different numbers of topics. This is an approximation of $p(w|K)$, i.e., the likelihood of the corpus given the number of topics (Ponweiser, 2012). The upper level for K was set to 200. However, this method is computationally very expensive and is therefore only feasible for the shorter texts in the academic corpora.

For the websites corpora, we set the topic number according to each individual corpus size. We simply use the total number of documents for setting K, assuming that a larger corpus contains more distinct topics:

$$D_m \geq 3000 : K = 200$$

$$D_m \geq 2000 : K = 150$$

$$D_m \geq 1000 : K = 100$$

$$D_m \leq 1000 : K = 50$$

The hyper-parameters for the LDA are in our case aligned to the needs to identify common content rather than to classify a document into a topic. Hence, we use the Gibbs sampling for determining the posterior probability of the latent variables. We use standard α =50/k as parameters of the prior distributions. For more information on determining the posterior probability of the latent variable see B. Grün and K. Hornik Grün and Hornik (2011).

3.4 Jaccard Similarity Coefficient

To measure the similarity between the sets of identified keywords, we use the Jaccard similarity coefficient as metric (Niwattanakul et al, 2013). We chose this similarity measure as it only includes element presence in a given set. It is applicable for the LDA and/or TFIDF generated keywords. This has two major advantages for our purpose: First, Jaccard similarity is not based on the input of scores or probabilities, which would in our case be hard to compare, since they result from different corpora and even usual normalization's are not necessarily good enough. Second, the overlap over terms is comparatively low, due to the high linguistic difference between academic writing and public websites, which is not the case for most other studies, focusing on more similar types of documents (Zhang et al, 2016). Therefore in our case a set comparison is more relevant. The similarity measure yields scores that are highly dependant on pre-processing and data type, and therefore needs specifically adjusted thresholds for our study. However, this said, it is not given that in other circumstances with similar goals other similarity measures, such as the cosine similarity or euclidean distance will not be more appropriate.

The Jaccard similarity is based on the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, 1 indicating most similarity (identical sets) and 0 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:

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

The thresholds for a minimum similarity for further examination were chosen based on preliminary results. In all applications, we only consider it a potential match if keyword lists return a certain minimum Jaccard similarity. However, the Jaccard similarity tends to benefit smaller sets. Hence, we decided to set a common threshold to a minimum of 0.13 and another used indicator threshold consisted in multiplying the Jaccard similarity with the intersection of the two sets, giving higher weight to sets with a higher amount of common words. Two sets with Jaccard similarity lower than 0.15 need more than 7 words in common in order to pass the criteria, while set pairs with Jaccard similarity higher than 0.15 can have smaller intersections.

3.5 Manual classification and verification

To determine whether the findings of the algorithms are actually relevant or valid, we needed human inspection and final verification. This is necessary since we are working with unlabeled text data and would not be able to verify the results without human confirmation. This step verifies the data and enables insights about the performance of the computational tools. Especially, since the data sample does not provide the possibility to have a labeled training data set, meaning that we have no training data, which could offer objective labels for the text matches. However, this is not really a possibility due to the huge amount of text pairs and the high level of complexity of the text documents.

We used 3 independent people from different disciplines to decide about the similarity of the text snippets. They were asked to categorize the text matches into one of 5 categories:

1. Identical topic = University contribution
2. Identical topic = Potential university contribution
3. Common topic = Unlikely directly related
4. Different topic = No match in content
5. Unclear = could not be classified

In the first label we included also findings about identical topics, which are a University contribution, but to a public entity, or media article or news about university research. If needed the people could resort to the actual full text publication, in case the abstract did not provide enough information for a final verification.

The human result classifiers background is as follow: three academics (PhD students) from different fields and one engineer. A fourth person was then making final decisions when disagreement is observed between the three human classifiers.The general idea is to use people that are capable to identify research topics and applications in various context.

4 Test Data Sample

To test our text mining methods we use Technical University of Denmark (DTU) and its economic environment as example case. To establish a first test sample, DTU is an appropriate case for this research for the following reasons: First, focusing on a technical university enabled us to study leading edge technology research with direct connections to industry innovation. Second, DTU provides a well documented case and the number of research institutions in Denmark is rather small, which allows straightforward attributions to a specific university. Third, Denmark has a high level of digitization and data availability, making it a promising setting for applying text mining. The scope is ideal as a first use case especially since DTU has already a comparatively high level of commercially relevant knowledge (`http://www.dtu.dk/english/Collaboration/Industrial_Collaboration`) and industry ties, which supports the assumption that there it is a fruitful case for tracing knowledge transfer. The type of research is very applied and hence highly relevant to the private economy.

As we aim to detect knowledge transfer from universities to the industry, we use the research output of the university as baseline since publication texts are the formalized output and dissemination channel of university research and contain all important research findings of a university (Toutkoushian et al, 2003). On the other hand, we use websites, which are companies channels used to ensure their visibility for potential consumers and investors including their most recent R&D successes and collaboration efforts(Branstetter, 2006; Heinze and Hu, 2006). The comparison of these two sources aims to detect knowledge overlap seems feasible.

Furthermore, Denmark, as national context, is ideal as its research is almost exclusively published in English language and most companies also use English as secondary, if not as first corporate language. This is highly relevant for the application of the text pattern recognition and for co-word occurrence measures.

4.1 Publication Database

We focus only on recent research outcomes by a university and exclude widely known and commonly accepted knowledge. Therefore, only novel scientific insights, technological innovations, like leading edge technologies shape the scope of this study.

Table 1 Total publications for the years: 2005-2016

Year	Abstracts	Only Texts	Abstract OR Text	All publications
2005-2010	16,502	2,738	3,854	40,455
2011-2016	28,517	5,137	11,963	38,011
2005-2016	45,019	7,875	15817	78,466

To identify relevant university research, we use the universities publications published by the university between 2005 and March 2016 . In the case of DTU, the data is taken from a database named ORBIT `http://orbit.dtu.dk/en/`. The retrieved entries present main research outputs by at least one employee of the university. However, the registration of research items only became mandatory in the year 2012, so it is important to mention that data coverage is not equal across the all years of the observation. The data provided by the database include a collection of academic abstracts, open-source full-text publications and publication meta-data. The meta-data includes among others: year, author(s), title, journal name, university section id, internal id, and DOI (digital object identifier). The number of all publication records for the time period is 78,466. For more detailed information on the available publication data, see Table 1. We cleaned the abstract data by removing all entries, which had no real text in the abstracts field, which resulted in 55 removed entries.

We classified the texts (abstracts and full-texts) by their database assigned departmental codes, which we converted into collections of research areas. This provides a pre-classification of texts by their fields. The sub-setting resulted in 24 separate research fields (see Section 2) of which three are irrelevant for the academic output of the university. (We excluded approximately 250 articles including 1) publications registered to the university administration, 2) publications registered to the bachelor program, and 3) one set that was directly linked to a large company). The collection of these research area based corpora will in the following be referred to as 'academic' corpora or by their individual name if this is relevant for the interpretation of the results. Most text mining methods perform better on more contextual coherent corpora and hence achieve better performances.

The distribution shows that the coverage and also the research output varies a lot between the research fields. This is crucial to keep in mind when analyzing the amount of observed knowledge transfer according to the fields. Especially, given that fields like Nuclear technology have only 316 abstracts but a high coverage since the entire output is only 422 articles, this might be due to the size of the research group at the university and/ or the groups age (see Section 2)

We chose the abstracts to serve as main research sample. This shortens computational time and enables better investigation of relevant fields and texts. The findings from this preliminary analysis are then used to find most relevant corpora for more in-depth and more extensive exploitation of the methods.

Table 2 Data coverage by research field: 2005-2017

Department	Abstract	Text	Total	% of Abst
Compute/Math	3890	1933	5791	67%
Biochemistry	2343	1038	4338	54%
Chemistry	1420	413	2352	60%
Civil Eng.	2122	1017	3675	58%
Electrical Eng.	3519	1778	4363	81%
Energy Conversion	1244	521	1536	81%
Environmental Eng.	1699	1269	3851	44%
Management Eng.	2569	1886	4521	57%
Mechanical Eng.	2999	1223	4293	70%
Nanotechnology	1935	918	3064	63%
Photonics	4262	2090	5617	76%
Physics	1434	685	1911	75%
Biology	2339	902	3562	66%
Transport	860	470	1686	51%
Wind Energy	1421	1158	1972	72%
Food Sciences	2846	1651	6210	46%
Aquatics	1481	787	4786	31%
Space Research	1432	782	2137	67%
Nuclear Technology	316	200	422	75%
Veterinary Sciences	1520	820	2594	59%
Other	2648	1954	8841	30%

4.2 Companies

The second data source, providing the company knowledge, was gathered from corporate company websites, since knowledge chunks, which are displayed on a website have to be of a certain commercial relevance for a firm.

First, we identified the key criteria for relevant companies, which are defined as: a) having a national (Danish) company registry number (CVR) and b) having had a collaboration contract with the university between 2006 and 2016. This constitutes a direct formal link between the companies and the university, which is the ideal basis to test and verify the new method.

To identify more potentially relevant companies, we generated one network on the basis of hyperlinks between the university and company websites. Hereby we identified additional partners linked to the university website. The list of websites contained many online service platforms. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample.

The websites themselves needed to provide as a minimum a set of 5 English web-pages with in English minimum of and more than 100 English words per page and display the CVR number on the website. We fetched the HTML content of the websites using a self designed web-crawler (`https://github.com/nobriot/web_explorer`) and converted it to usable plain text cleaning it from any remaining code tags. These online text samples were collected between August 2016 and November 2016. Exploring the websites, we visited

908,288 total web-pages (single text documents in total), that had to be filtered by the above mentioned criteria for websites.

The number of total number of companies, which could be identified as collaborators of the university between 2006-2016 was 1225 of which 699 had a CVR number written on their website and 544 were displaying the Anpartsselskab (ApS) abbreviation (which describes limited liability companies in Denmark). Certain companies went out of business, underwent mergers or were just renamed. We tried to identify the new names or entities, however this was not in all cases possible. We were left with a final sample of 445 companies. The firms in this sample operate mainly in technology intensive sectors and are firms with strong R & D divisions. Therefore it included companies with contents related to the research performed at the university.

To provide an overview of the composition of the firm sample we decided to identify the main industry field of each company by using additional text based tools. This is reasonable since the identification of topics and clustering of texts has a long tradition and has successfully been used in various research areas.

Table 3 Page and term numbers per website (descriptive)

	Pages (P.)	P. Mean	P. Median	Terms	T. Mean	T. Median
total	138544	311	69	2185191	4911	2233
lower boundry	5	–	–	38	–	–
1st quantile	22	12	10	905	521	523
2nd quantile	69	42	40	2233	1476	1408
3rd quantile	257	142	130	6018	3819	3675
4th quantile	10106	155	591	67351	13866	10466

We applied the LDA for the clustering of companies (for more details see Section 3 to identify the main categories for the firms, showing the overall distribution of firms that work within one topic or field. We used our knowledge of the sample to set the optimal number of topics ($K = 45$). To avoid too generic topic clusters we erased all words that were used in more than 80% of the websites, which removes website specific terminologies, such as the contact information, impressums and similar. For a better understanding we summarized the single topics with their most relevant keywords for each topic (see Table 4). The clustering cannot be assumed as reliable as the labels from the scientific fields, however they show clear focus in some fields (see Table 5).

The number and length of pages varies a great deal between company websites (see Table 3). Some have an English summary for their main contents, while others, often multinationals have their entire website in English. This difference in length clearly influences the performance of the statistical models, since long text documents generally influence these models more than short ones. In this sample collection, we also ensured to capture the content of PDFs or similar formats stored on the websites. These required special treatment

Table 4 Example topics for the company websites with their top terms

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
design	gas	hear	health	product	share
product	oil	loss	sustain	food	report
partner	develop	implant	board	process	annual
custom	report	support	report	sugar	cash
read	million	sound	news	farm	market
Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
water	drink	lab	network	oil	health
power	milk	cell	support	vessel	journal
plant	cream	order	data	gas	research
system	process	center	center	ship	clinic
pump	fill	support	switch	power	medic
Topic 13	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**
custom	drill	wind	light	cancer	plan
data	reservoir	project	electron	influenza	consult
platform	seismic	system	power	prevent	project
network	fluid	public	wire	flu	design
cloud	data	product	tool	control	environment

Table 5 Topic distribution of the websites

Topic nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	26	3	4	2	37	6	1	9	12	7	1	7	21	25	1
Topic nr.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	34	1	16	23	13	9	2	2	3	2	7	2	10	2	7
Topic nr.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	1	61	2	17	5	4	2	11	1	3	21	3	4	13	2

and are treated as pages of the websites. Each website is stored as its own corpus. Even though this might seem drastic, it is a sufficient way to ensure a comparable pre-classification like research fields and fosters the performance of the statistical methods.

5 Results

We divide this section according to the results of each applied method to give an explicit insight into the performances and future potential of the single applications. It is crucial to keep in mind that this study is a first step to verify effectiveness, limitations and eventually identify applicable thresholds and suggest future improvements. Finally, we set the results into context and evaluate the outcomes based on the studies objectives. In each subsection we clearly describe which data samples are used and why. This is crucial because of the varying demands of the different methods. The different methods generated different outcomes in terms of keyword lists, due to their different levels of application (document or corpus level)(see Table 6).

Our pre-processing revealed some specific challenges, in particular in the case of the academic abstracts. The abstracts contain, for instance, chemical formulas and notations, which rely heavily on numbers and/or special characters. These are removed during the course of the pre-processing and therefore lost in the subsequent application. The only possibility to later identify same formulas to use them for similarity measures is the assumption that the removal of those characters will always result in an identical end character string, but it might not always be the case. Often the result may not be identifiable as the particular formula, but still provides a match. In some rare cases HTML, or other code tags prevented the identical deconstruction and in such cases, we did not find a way to identify the matching strings. However, some terms may seem like the result of poor pre-processing, but are in reality just a representation of specific models, formulas or project names shrunk to an unidentifiable string of characters. The websites on the other hand are challenging in a different way: they contain different language snippets, which are embedded in every site forcing language detection on lower levels. Therefore we decided to only integrate web-pages that have a minimum of 80% English terms. Additionally we found that the linguistic composition of websites is comparatively repetitive within a website, meaning that the words companies use to describe products or services are not very diverse, which leads to high number counts for single terms. Publications, on the other hand, have a much richer vocabulary and therefore suffer less from this skewed word distribution. To account for this different composition of the two text types we normalized or removed the words in question when needed.

5.1 Text Comparisons

To identify potential text documents with identical knowledge pieces we first compare the keywords from publications and websites with the computational methods.Hereby, we identify text pairs that potentially contain identical knowledge content. However, in the final step these potential matches have to be manually verified.

The keywords are derived through TFIDF indexing or extracted from the topics of LDA. The LDA on the academic corpora resulted in 915 distinctive topics for all 21 academic corpora. The LDA for the websites resulted in 8250 distinct topics (see Table 6).

To verify the performance of the LDA application, we manually inspected the derived topics for several corpora to ensure the performance, including the decision regarding topic numbers and prior settings.

The manual inspection revealed a much clearer picture with the academic texts than with the websites. The topics for the single scientific areas seem very distinct and reasonable (see Table 7).

Table 6 Keyword lists

Method	Total number of topics/keyword lists	Number of corpora
TFIDF web	138552	380
TFIDF orbit	44294	21
LDA topics orbit	915	21
LDA topics web	8250	340

Table 7 Topic example for one academic corpus

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
enzym	bind	dynam	forc	oligosaccharid
domain	site	vibrat	particl	branch
amino	conform	motion	hydrophob	carbohydr
residu	enzym	coupl	friction	donor
express	residu	excit	layer	polysaccharid

While the LDA applied to websites still gave some good indication about their main area, the topics seem less distinct. However, the manual inspection suggests that LDA is capable to represent the main content of a website, but due to the previously mentioned word repetition adjustments in terms of too frequent words need to be made. Accordingly, we removed all terms occurring in more than 90% of documents in a website.

Since the LDA can only be applied on texts that have a certain length, as the algorithm depends on the amount of text data input, we had to exclude 40 smaller website corpora. These corpora are suited for the TFIDF application, but not for LDA. Therefore the sample size of LDA is slightly smaller than for TFIDF (see Table 6). The LDA provides a certain number of topics for each corpus (these vary according to the website length (see Section 3.3). Each of these topics are composed of specific words, which we extract and combine to a keyword list. For the LDA comparison we selected the 50 most relevant (probable) words for each topic (LDA allows term re-occurrence in different topics with different probabilities). We compared each topic from one of the 21 academic corpora and 340 website with each other. Each time a keyword list is compared to another and the Jaccard similarity is computed for each comparison. More than 7,548,750 individual comparisons were performed.

Examining the Jaccard scores revealed that none of the comparisons scored higher than the set threshold of 0.13 (the first set threshold). The first matches between topics were around the threshold of 0.08. This is a really low similarity score and shows that the academic and web corpora are very diverse in the main areas. 12 document pairs could be identified exceeding a Jaccard threshold of 0.08. Comparing the academic topics from different departments with each other reached scores up to 0.82 Jaccard similarity, which shows how much closer academic corpora are related.

The TFIDF provides keywords for each document, hence the number of lists equals the number of documents in each corpus. We extracted up to 50 highest indexed terms for each document (see Table 6). We compared each TFIDF keyword list from the academic documents with all keyword lists from the websites.The maximum length of the keyword lists was set to 50 extracting the words with the highest TFIDF scores (see Section 3). However, some texts, mostly the academic abstracts, were too short to generate a list of 50 words, hence we decided to set the length of the list of words to all words remaining after cleaning and pre-processing. We additionally excluded around 10 websites, as they were too short for the application of the TFIDF. However, comparisons for shorter texts are object to the adjusted Jaccard threshold (see Section 3) to ensure that the short keyword lists are not dominating the final match sample with less relevant matches. We retrieved 44,294 lists for the academic abstracts. and 138,552 keyword lists for the websites resulting in 6,137,022,288 comparisons.

Compared to the LDA application some keyword list pairs scored comparatively high. 124 pairs with 0.13 Jaccard similarity threshold were identified. After a preliminary manual inspection we decided to apply another cleaning step for the TFIDF matches, since some particular matches share no contents, but only certain distinct words that are irrelevant for the content, such as foreign language fragments or country names (see Table 8). We excluded there fore all the pairs that were matched on those kind of keywords. 91 final text pairs that were after the cleaning procedures which represents a very low $1.48 \times 10^{-6}\%$. For the purpose of comparison we tested two different academic corpora from 'Mechanical Engineering' and 'Computer Science and Mathematics' and compared their TFIDF keyword lists. The assumption is that the contents are more related and the linguistic composition closer. This test resulted in 487,961,509 comparisons. By applying the same thresholds a total of 1377 matches was identified which is $2.8 \times 10^{-4}\%$ matches, way higher than in the websites against academic documents . This comparison shows that the single match between academic and website documents is more relevant, since these are not commonly coincidental. It also confirms the high diversity between the two sets of documents.

Table 8 Example of word combinations which had to be excluded from potential matches

Countries		German		Danish	
"kingdom"	"franc"	"wird"	"auf"	"eller	"flere"
"germani"	"european"	"der"	"ein"	"som"	"til"
"poland"	"finland"	"die"	"bis"	"til"	"det"

We have also compared the retrieved keywords from the TFIDF of the websites with the keywords found with the LDA topics computed on the academic corpora. We again set an upper threshold to 50 words per topic. This comparison yielded to a total of 33 matches and after the second clean-up, only 13 potential matching pairs.

To identify the actual documents belonging to an actual topic generated by the LDA is not straightforward, since only a probability distribution over documents is given. Hence, we used for each topic the two documents with the highest probability. This resulted in each TFIDF text having two potential matches for academic abstracts.

5.2 Human verification of the text pairs

The results generated with the TFIDF to TFIDF comparison and the LDA to LDA comparisons show a significant theme overlap between the text documents. Comparing the text pairs retrieved from these both applications resulted in 10 common matches, meaning that the TFIDF and LDA returned 10 times the same text pairs. Interestingly, in the manual verification these documents are websites that achieved many hits via both applications, but refer only to overall similar content, but did not share identical research content. This means in the classical application of topics models to detect knowledge flows these pairs would have been a valid match. In our case,however, we are tracing some more specific content and these pairs do not provide clearly the same concepts, models or other knowledge. This is crucial, since these are the matches that would have been a positive identification of knowledge flows according to traditional measures using only LDA. Certain research areas revealed to be particularly dominating the text pairs, as well as in the true positives and in the entire matched sample. The overall comparison suggests a clear dominance of certain university departments in the matches. Some Departments are most represented in the matched pairs.

The combination of LDA and TFIDF reveals common interests of firms and the university and also shows which departments most are represented within the pages. Especially given that some of the comparatively small academic corpora (see Table 2) are most relevant according to the Jaccard similarity and the actual matches. In Figure 2 we can clearly see that some departments are much more dominant when it comes to the pair-wise comparison. This means that the methods are most successful for those corpora, not determining whether it is only a content relatedness or fully identical contents. However, the other corpora seems not suited for our approach.

This is an example for a true positive, so a real text pair, which has common content and refer to the same knowledge would be the following two texts:

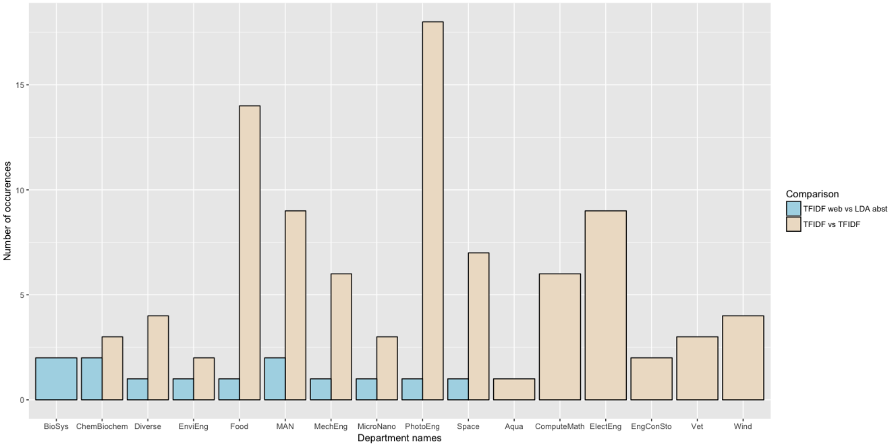

Fig. 2 TFIDF and LDA showing the most dominant research areas leading to matches that exceed the threshold

Academic abstract

"Swarm is the fifth Earth Explorer mission in ESAs Living Planet Programme to be launched in 2009. The objective of the Swarm mission is to provide the best ever survey of the geomagnetic field and its temporal evolution. The innovative constellation concept and a unique set of dedicated instruments will provide the necessary observations that are required to separate and model the various sources of the geomagnetic field (...)"

Website document

"Absolute Scalar Magnetometers from CNES and CEA/LETI which were selected by the ESA for the Swarm mission. (...) The Swarm mission ; a constellation of three identical satellites in three different polar orbits between 400 and 550 km altitude to measure the Earths magnetic field (...)"

These texts show that the company is actually displaying the 'Swarm', which is the topic of the academic publication. In particularly hard cases or very limited information from the abstract the validators could fall back on the full texts of the publication.

Academic abstract	Website document
"Higher-Order ambisonics (HOA) ; and a matrix inversion method. HOA optimizes the reproduced sound at a sweet spot in the center of the array with radius determined by a spherical microphone array ; which is used to derive the spherical harmonics decomposition of the reference sound. The four-loudspeaker-based method equalizes the magnitude response at the ears of a head and torso simulator (HATS) for sound reproduction (...)"	*"Higher-order ambisonics ; matrix inversion method ; ETSI TS 103 224 and matrix inversion method optimized for a specific device. For each method ; the quality of the reproduced sound was evaluated both objectively and subjectively ; at microphones close to a device under test and at the ears of a Head And Torso Simulator (HATS) (...)"*

The second example is according to the human verification only thematic related and does not qualify as a full match. Hence, we have to declare it a false positive. In this particular case they are very closely linked thematically, but the publication is based on the four loudspeaker method, which is not the case in the website. Therefore, these pages are labeled under category 3.

Given this examples it is obvious that the actual task is not simple and is it might appear in the first place. Therefore, we needed to ensure the quality of the assessment and ensured that several persons from different backgrounds were performing the assessment. The manual verification was performed by three persons, two researchers (PhD candidates) and one engineer, and a fourth person to handle possible mismatches in the assessment. All three are scientists and hence familiar with research and the interpretation of research results. The topic to topic comparison, with an adjusted threshold of 0.08 Jaccard similarity resulted in no positive evaluated match between texts, this confirms the assumption that the threshold has to be carefully chosen, in particular in regard to semantically very diverse texts.

Table 9 Overlap in manual decisions

	Academic 1 & 2	Academic 1 & Eng.	Academic 2 & Engineer	All
Total	67%	61%	58%	48%
Category 1	80%	65%	60%	60%
Category 1 & 2	74%	56%	50%	44%
Category 2	21%	29%	14%	21%
Category 3	61%	54%	48%	38%
Category 4	51%	49%	43%	30%

Given the assessment it is clear that the engineer has a much harder time to verify identical contents that are not within his area of expertise. To see the confidence levels of each verification they made qualitative comments to their decisions, which enabled a more accurate final assessment. In Table 9 certain inconsistencies become evident. The overlap within the relevant categories 1 and 2 the low consistency was solely caused by their different understanding of the definition and was finally solved and decided based on their qualitative comments. They also commented on pairs that seemed unclear or difficult to classify to them, or in which they claimed to have specific expertise,the final labeling could be made very accurate. In particular most of the academics assessments revealed an insecurity between two labels while the other was certain about a particular label. However, the overlap for the engineer was much lower and the comments showed only a few certain classifications within his area of expertise. Revealing that mainly trained academics, used to reading academic texts, are capable to mange this tasks with sufficient confidence levels.

The fourth person (academic) had evaluate the qualitative statements, read the texts and make a final decision in alignment to the previous assessments.This strategy ensured the the quality of the results. Given the distribution of decisions (see Figure 3) one of the main inconsistencies in the overall distribution was also the low usage of number label number 5 by the second academic, this label however should be inconsistent since it is the label for to remove the text pairs where the validator was really insecure about the labels.

The results of the verification show great overlap in content and a number of certain positive matches. As previously described the LDA comparison retrieved 12 potential matches, TFIDF 91 and the combination of both approaches 13 (see Table 10). These 12 document pairs reveled close topic connection (meaning they contained related content), but did not refer to any concrete common knowledge piece . This is not surprising since LDA is applied on a corpus level and does not in detail represent documents. Nevertheless, the common topics and themes helped to set a base for the TFIDF application. In the final verification process it was revealed that out of the 91 potential matches 27 could be verified by humans, which gives a 30% successful detection of identical knowledge pieces .

After this first comparison the performance of the TFIDF showed more success in identifying potential common contents (see Table 10). Remarkable is also that the only clearly non technical field

Table 10 Number of identified potential matches

Methods	Comparisons	Matches	Matches verified
TFIDF web vs TFIDF orbit	6,137,022,288	91	20
TFIDF web vs LDA topics orbit	126,775,080	13	2
LDA topics web vs LDA topics orbit	75,487,50	12	0

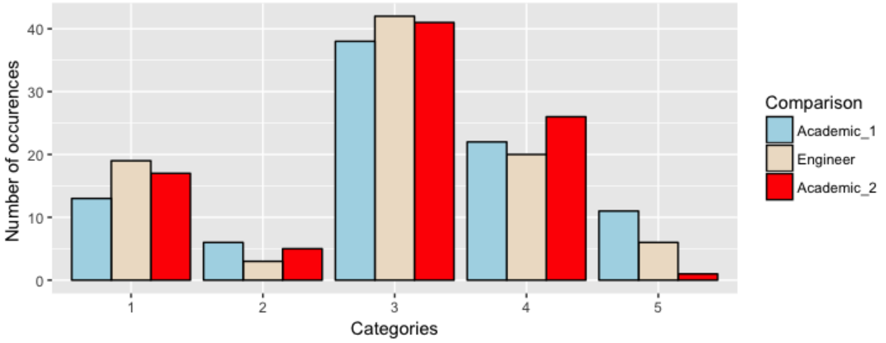

Fig. 3 Decision distribution of the manual assessment

5.3 Technical Considerations

Given the progress made in the past decade the text similarity measures might become sophisticated enough to compare full texts, but for the time being we will have to apply additional strategies. For further refinement and extension, it could be considered to adopt another method for associating the documents to the LDA topics. For example, we could pick the documents connected to the highest ranking words in a given LDA topic instead of taking the highest topic probability in a document. This might be another option for future research. However, due to the size of the original sample and the complexity of the actual labeling, for now it is not possible to estimate the error on how much of the actual knowledge transfer, or the true positives are not identified.

Our findings suggest that our first estimated thresholds proved to be not accurate enough. 0.134 Jaccard distance would have been the ideal threshold for finding all text pairs for the TFIDF with a list union size close to 100 words. The best threshold would have been 0.144, here we have the best trade of between false positives and missing findings. In Figure 2, we show the potential changes in categories (label assignment) with improved thresholds of the Jaccard measure. We lost only one match and reduced the error rates by more than 50%. The amount of first and second order matches gradually decreases with lower Jaccard similarity, as well as the content relatedness. Therefore, we suggest to evaluate the hits in future sequential, meaning to rank the hits by their Jaccard similarity and assess the first hits and stop when the amount of positive hits decreased significantly.

6 Conclusion

The purpose of this study was to offer new insights into both, formal and informal modes of knowledge transfer. The outcome is the development of novel detection and measurement approach for knowledge transfer, capturing instances of knowledge transfer, which are largely overlooked by current methods (Agrawal, 2001). Hence, this study enables new perspectives and further in-depth understanding for reshaping existing notions on what constitutes successful university-industry collaboration in particular for policy makers and other stakeholders. It also provides generalizable and comparable findings and identifies and verifies the transfer of concrete pieces of knowledge, enabling the detection of common knowledge.

The tools we applied to detect university research as being used and displayed by private firms were indeed able to identify those instances. This study detects the use of publicly produced knowledge and moves beyond the traditional proxy indicators. Our results are not bound to the usual formal indicators and capture formal and informal knowledge transfer, as long as it is displayed from the company side. The high level of detail enables the study to show, which knowledge pieces are relevant enough for the industry to display. The trace of knowledge transfer can be directly linked to specific studies or research areas. More than 5 % of the firms actually displayed some concrete knowledge driven from the university on their websites. Additionally, we still traced highly related working topics and working areas proven to be simple among the university collaborators, which adds a value to the method allowing universities to capture the most related topics with the firms in their environment (see Figure 3 all matches contained in the third label (category 3)). In summary, the method provides insights about the transferred knowledge and is a novel quantitative assessment. It provides statistical correlation measures, which could be used supplementary to already existing methods from the Triple Helix concept.

Even though the findings are still on a comparatively small scale, this outcome indicates that the method can successfully detect knowledge transfer. It found several instances where models, methods and clinical studies of the university were used but not directly cited. This is only a first step, but shows clearly the potential of the methods. And even though our approach reveals nothing about the underlying processes and the how of knowledge transfer from university to industry we broaden the measurement spectrum for the instances where knowledge transfer happened,regardless of the channels or mechanisms. Furthermore, the applied methods show that it is actually possible to identify concrete pieces of research knowledge in linguistically very diverse documents. This study is a first step towards an novel supplementary identification of concrete university-industry knowledge transfer.

This insight increases the understanding of the principal value of university research independently from its direct commercial success, highlighting the dissemination potential and the absorption of relevant research. Based on these findings, we might broaden the definition of 'valuable' research beyond what normally is considered valuable through patenting and licensing contracts. This would include changes within the focus on commercial value of public research, lending further support to the potential of new streams of research not identified through more traditional measurements. This could improve the funding situation for relevant but not easily commercialized research in the future, since it would enable decision makers in the university and externally to take into consideration what knowledge is actually used in the industry later. Obviously, our methods still require adjustments, but it is certainly a step to improve the understanding about public research relevance, and a strong indication that the current measures are insufficient in capturing all commercially valuable research outputs.

7 Future Outlook and limitations

From an application perspective, several dimensions must be evaluated before the method can be widely adapted. For instance, it is crucial to benchmark the new method against the traditional indicators to assess the actual knowledge gain. Additionally, this method could be applied in different empirical settings to better understand the overall performance and application possibilities.

From a conceptual point of view, it has to be determined what this knowledge actually represents for companies and research dissemination. This estimation might not be as straightforward as it is in the case of patents or licenses, but must represent a commercial value to a company. Patents and licenses typically carry a certain commercial value, whereas the value of information on corporate websites is less understood.

From a performance perspective of the method our work can be viewed as a first step, using comparatively established methods. Technically, however, there are several improvements and bench-marking options possible. Hence, we suggest to refine the statistical methods and add more advanced statistical learning methods to improve the error rates. Focusing on the best performing research areas (see Figure 2) would also be an option to improve the performance by strategically adjusting it to the given field.

Given these results, simpler classification might be necessary in future. Additionally, in the contrary to our expectations, the rightful classification seems to be difficult for non academics particularly when the content does not match the area of expertise. This speaks for the high quality performance of the method: if human cannot easily distinguish false and true positives means that the method is performing well, since humans are usually performing better when it comes to this kind of tasks.

Despite the current limitations, we see clear future potential as the flexibility of the tools including potential for adaptation make them useful in various contexts.

Acknowledgements We thank the people performing the human validation of our results and the helpful comments we received on several conferences.

References

Aggarwal CC, Zhai C (2012) Mining text data. Springer Science & Business Media

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

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

Aizawa A (2003) An information-theoretic perspective of tf–idf measures. Information Processing & Management 39(1):45–65

Arundel A, Marcó CB (2008) Developing internationally comparable indicators for the commercialization of publicly-funded research

Berry MW, Castellanos M (2007) Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition p 241

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993–1022

Branstetter L (2006) Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States. Journal of International Economics 68(2):325–344, DOI 10.1016/j.jinteco.2005.06.006

Chapman, Hall/CRC (2010) Handbook of Natural Language Processing, Second Edition. DOI 10.1007/978-1-4612-3426-5_15

Cheah S (2016) Framework for measuring research and innovation impact. Innovation 18(2):212–232, DOI 10.1080/14479338.2016.1219230

Cohen WM, Nelson RR, Walsh JP (2002) Links and impacts: the influence of public research on industrial r&d. Management science 48(1):1–23

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural Language Processing (almost) from Scratch. The Journal of Machine Learning 12:2493–2537

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

Etzkowitz H, Leydesdorff L (2000a) The dynamics of innovation: from national systems and mode 2 to a triple helix of university–industry–government relations. Research policy 29(2):109–123

Etzkowitz H, Leydesdorff L (2000b) The dynamics of innovation: from National Systems and 'Mode 2' to a Triple Helix of university... Research Policy 29(2):109

Etzkowitz H, Webster A, Gebhardt C, Terra BRC (2000) The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm. Research policy 29(2):313–330

Franceschini S, Faria LGD, Jurowetzki R (2016) Unveiling scientific communities about sustainability and innovation. A bibliometric journey around sustainable terms. Journal of Cleaner Production 127:72–83, DOI 10.1016/j.jclepro.2016.03.142

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

Garechana G, Río-Belver R, Bildosola I, Salvador MR (2017) Effects of innovation management system standardization on firms: evidence from text mining annual reports. Scientometrics 111(3):1987–1999

Glänzel W, Thijs B (2012) Using core documents for detecting and labelling new emerging topics. Scientometrics 91(2):399–416

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America pp 5228–35

Grimpe C, Hussinger K (2013) Formal and informal knowledge and technology transfer from academia to industry: Complementarity effects and innovation performance. Industry and innovation 20(8):683–700

Grün B, Hornik K (2011) topicmodels : An R Package for Fitting Topic Models. Journal of Statistical Software 40(13):1–30

Gulbrandsen M, Slipersaeter S (2007) The third mission and the entrepreneurial university model. In: Universities and Strategic Knowledge Creation: Specialization and Performance in Europe, chap 4, pp 112–143

Han J (2017) Technology commercialization through sustainable knowledge sharing from university-industry collaborations, with a focus on patent propensity. Sustainability 9(10):1808

Heinze N, Hu Q (2006) The evolution of corporate web presence: A longitudinal study of large American companies. International Journal of Information Management 26(4):313–325, DOI 10.1016/j.ijinfomgt.2006.03.008

Jaffe AB, Trajtenberg M, Fogarty MS (2000) Knowledge spillovers and patent citations: Evidence from a survey of inventors. American Economic Review 90(2):215–218

Kao A, Poteet SR (2007) Natural language processing and text mining. Springer Science & Business Media

Khan GF, Park HW (2011) Measuring the triple helix on the web: Longitudinal trends in the university-industry-government relationship in korea. Journal of the Association for Information Science and Technology 62(12):2443–2455

Leydesdorff L (2004) The university–industry knowledge relationship: Analyzing patents and the science base of technologies. Journal of the Association for Information Science and Technology 55(11):991–1001

Link AN, Siegel DS, Bozeman B (2007) An empirical analysis of the propensity of academics to engage in informal university technology transfer. Industrial and Corporate Change 16(4):641–655

Liyanage C, Ballal T, Elhag T, Li Q (2009) Knowledge communication and translation - a knowledge transfer model. Journal of Knowledge Management 13(3):118–131

Magerman T, Van Looy B, Song X (2010) Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics 82(2):289–306

Mao W, Chu WW (2007) The phrase-based vector space model for automatic retrieval of free-text medical documents. Data and Knowledge Engineering 61(1):76–92, DOI 10.1016/j.datak.2006.02.008

Meyer M, Siniläinen T, Utecht JT (2003) Towards hybrid triple helix indicators: A study of university-related patents and a survey of academic inventors. Scientometrics 58(2):321–350

Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S (2013) Using of jaccard coefficient for keywords similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, vol 1

Paukkeri Ms, Honkela T (2010) Likey : Unsupervised Language-independent Keyphrase Extraction (July):162–165

Perkmann M, Walsh K (2007) University–industry relationships and open innovation: Towards a research agenda. International Journal of Management Reviews 9(4):259–280

Ponweiser M (2012) Latent Dirichlet Allocation in R. PhD thesis

Richardson GM, Bowers J, Woodill aJ, Barr JR, Gawron JM, Levine Ra (2014) Topic Models: A Tutorial with R. International Journal of Semantic Computing 08(01):85–98

Robertson S (2004) Understanding inverse document frequency: On theoretical arguments for idf. Journal of Documentation 60:2004

Rus V, Niraula N, Banjade R (2013) Similarity Measures Based on Latent Dirichlet Allocation, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 459–470

Schmidtler MA, Amtrup JW (2007) Automatic document separation: A combination of probabilistic classification and finite-state sequence modeling. In: Natural Language Processing and Text Mining, Springer, pp 123–144

Siegel DS, Waldman DA, Atwater LE, Link AN (2003) Commercial knowledge transfers from universities to firms: improving the effectiveness of university-industry collaboration. The Journal of High Technology Management Research 14(1):111–133

Sung TK, Gibson DV (2000) Knowledge and Technology Transfer : Levels and Key Factors. Proceeding of the 4th International Conference on Technology Policy and Innovation

Thursby JGJJG, Jensen Ra, Thursby MCM (2001) Objectives, characteristics and outcomes of university licensing: A survey of major US universities. The Journal of Technology Transfer 26(1):59–72

Tijssen RJ, Van Leeuwen TN, Van Wijk E (2009) Benchmarking university-industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities. Research Evaluation 18(1):13–24

Toutkoushian RK, Porter SR, Danielson C, Hollis PR (2003) Using publications counts to measure an institution's research productivity. Research in Higher Education 44(2):121–148

Tussen R, Buter R, Van Leeuwen TN (2000) Technological relevance of science: An assessment of citation linkages between patents and research papers. Scientometrics 47(2):389–412

Van Eck NJ, Waltman L (2017) Citation-based clustering of publications using citnetexplorer and vosviewer. Scientometrics 111(2):1053–1070

Wu Y, Welch EW, Huang WL (2015) Commercialization of university inventions: Individual and institutional factors affecting licensing of university patents. Technovation 36:12–25

Xia T, Chai Y (2011) An improvement to TF-IDF: Term distribution based term weight algorithm. Journal of Software 6(3):413–420

Yau CK, Porter A, Newman N, Suominen A (2014) Clustering scientific documents with topic modeling. Scientometrics 100(3):767–786

Zhang Y, Zhou X, Porter AL, Gomila JMV, Yan A (2014) Triple helix innovation in chinas dye-sensitized solar cell industry: hybrid methods with semantic triz and technology roadmapping. Scientometrics 99(1):55–75

Zhang Y, Zhang G, Chen H, Porter AL, 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