Paper to be presented at: DRUID17 
NYU Stern School of Business, New York, June 12-14, 2017

Search for Knowledge: Text Mining for Examination of University-Industry Knowledge Transfer

Sabrina Woltmann
Technical University of Denmark
DTU Management Engineering
swol@dtu.dk

Lars Alkærsg
Technical University of Denmark
DTU Management Engineering
lalk@dtu.dk

Abstract
This paper identifies knowledge transfer between universities and the industry by using novel computational methods. Current research on university-industry knowledge exchange relies mainly on formalized data and indicators such as patents, collaborative publications, and license agreements, to assess the knowledge transfer by universities. We complement the existing research by proposing a novel implementation of a method for comparing the texts from university publications to texts from firm websites. We use a word ranking method from the field of information retrieval called term frequency-inverse document frequency (TFIDF) and topic models identify common key words and thereby identify overlapping contents among the texts. We find that several websites contain content-related and partly traceable back to the university research. Our results show that university research is represented in the websites of industrial partners. We further propose possibilities to enhance the precision of the method. This paper is the first step to enable the identification of common knowledge and knowledge transfer via text mining to increase its measurability.

Jelcodes:032,031
Search for Knowledge: Text Mining for Examination of University-Industry Knowledge Transfer

1 INTRODUCTION

Decades ago a change perception of policy makers and public took place adding a new objective to universities traditional tasks, the third mission. After teaching and research, this mission is an interrelated mission, which aims to translate the teaching and research efforts of universities into economic contributions. Therefore the universities implemented various forms of knowledge transfer activities (Gulbrandsen & Slipersaeter 2007). As with teaching and research, the strategies and activities through which universities pursue the third mission vary from university to university (Ankrah & Al-tabbaa 2015, D’Este & Patel 2007). The knowledge transfer activities are highly depending on exogenous factors, such as legal frameworks (Geuna & Rossi 2011), public research policies, funding incentives (Munari et al. 2016) and the research fields the university is active in (Bekkers & Bodas Freitas 2008).

However, the key objective remains the same: the active dissemination of novel knowledge and technologies to the socioeconomic environment of the university (Zawdie 2010). From a policy perspective, the main incentive is to ensure that knowledge is actually received by the industry or other relevant third parties, which in turn are able to utilize and potentially further develop the novel knowledge (Agrawal & Henderson 2002).

In times where leading edge technologies decide about the economic development of sectors, regions and nations and about their competitive positioning in the global economy, knowledge is the key driver to foster (socio) economic development. University as research institutions have been contributing for centuries to regional knowledge development. Hence, their research plays an important role in terms of technological innovation and knowledge
creation. Research driven innovations lead to economic growth, development and increase competitiveness (Huggins et al. 2008, Vincett 2010).

Most concepts of university research impact assessment are based on the notion of knowledge transfer and is often assessed on a case-by-case basis or via somewhat limiting proxy indicators. This implies challenges to generalize the empirical findings, quantify the knowledge transfer and to draw concrete conclusions about the actual contributions to socioeconomic development. The emphasis of knowledge transfer detection lays on transfers leading to commercially relevant research driven innovations. However, even commercialized knowledge is only detected, if it is protected by patents, declared via scientific co-publications, subject to entrepreneurial activities or sold/licensed to the industry. Mostly it needs to be generating some sort of direct revenue. It is acknowledged that commercialization is only one aspect of knowledge transfer and that several other levels are existing, including the creation, sharing and implementation (Sung & Gibson 2000). Moreover, commercialized inventions are estimated to represent only a small fraction of the actual transferred and used knowledge (Agrawal & Henderson 2002, Drucker & Goldstein 2007). But the transfer of not commercialized knowledge is even harder to trace and measure, which leaves the research community, up until now, with lack of understanding of its occurrences and very limited metrics to capture these. Therefore, the actual knowledge transfer often remains an approximation.

Various attempts to identify and quantify knowledge transfer on diverse levels show that current scholarly literature fails to provide metrics to capture the complete knowledge transferred from universities to the industry (Malerba 2007). Different proxy-indicators and assumptions about knowledge transfers, spillovers and their channels are employed to compensate this lack of direct measurability (Cheah 2016, Lin 2016, Salter & Martin 2001). These indicators are not holistic, but as F. Malerba states for one of them: ‘the use of patent citations in order to examine knowledge flows and networks is a very fruitful research direction, provided that one is aware of their limitations and uses them jointly with other qualitative and quantitative indicators’ (2007, p. 13).

Given these evident research challenges, the three key objectives for this study are i) to identify novel methods that allow direct identification of common knowledge contents between universities and industry using additional data sources, ii) to identify whether these common contents originate from the university and iii) to capture common areas of knowledge in the geographical proximity of the university. The overall goal is to enable a flawless detection of research knowledge contents ensuring generalizable and comparability of findings regardless of the case. We propose a novel method that
identifies and to a certain extent quantifies knowledge transfers from universities to the industry. We aim to capture the transfer without focusing on channels, commercialization or transfer mechanisms.

We are proposing contemporary computational methods from the field of natural language processing (NLP) including text mining and statistical learning tools to adapt a novel metric to measure knowledge transfer. Using text material from corporate websites and academic publications, we aim to identify common and related topics. We use pattern recognition tools and similarity measures to identify overlapping and coherent contents. Point of departure are the scientific texts, assuming that publications are limited to developers of an invention or novel research insights. Hence, we derive content from various scientific texts by organizing them into text corpora and extracting their key concepts. Afterwards we identify similarities to the commercial websites.

The clear identification of common knowledge areas of a university and its economic environment, the traceability of shared concepts, knowledge and technologies provides additional tools to enable scholars, policy makers and practitioners to perform more in-depth analysis. The flexibility of the tools and their potential for adaptation make them useful in various contexts.

2 THEORY

The body of literature on knowledge transfer between universities and the industry contains several interconnected topics. These focus on issues including firm and university characteristics when engaging in collaborative activities (Ankrah & Al-tabbaa 2015, Brostroem 2012), the identification and verification of knowledge transfer channels (Agrawal 2001, Grimpe & Hussinger 2013, Schartinger et al. 2002), policy implications, funding and legislative regulations for universities engagement (Munari et al. 2016), the role of academic fields or industry sectors (Bekkers & Badas Freitas 2008, D’Este & Patel 2007), the impact of university research including economic, societal and political dimensions (Drucker & Goldstein 2007, Jong et al. 2014). Key aspects are, among others, the measurable identification of knowledge transfer itself and its subsequent commercialization successes (Thursby & Thursby 2002). This diversity shows that the understanding of university-industry collaboration and subsequent knowledge transfers including its impacts are highly investigated by now. This reflects a well-developed academic research area, which led already to an advanced policy understanding and elaborated empirical studies (Munari et al. 2016).

Given this broad understanding it is evident that knowledge development
and knowledge transfer are highly interrelated and increasingly relevant topics for academic research. Today the area is a multidisciplinary empirical research field including well developed literature on knowledge transfer and university-industry collaboration.

2.1 Knowledge

The majority of studies, however, lack a clear definition of the concept of 'knowledge', as well as of 'knowledge transfer' (Liyanage et al. 2009). The concepts seem to be commonly agreed and the framework seems widely understood; however we see the need to use definite concepts for this paper.

The term itself is highly debated and conceptualized in various philosophical approaches; to limit it to a reasonable scope, we focus on the definition relevant for this particular context. Therefore we use the foundations of knowledge management theories. Knowledge management focuses on two main types of knowledge: explicit and tacit knowledge. “Explicit or codified knowledge involves know-how that is transmittable in formal, systematic language and does not require direct experience of the knowledge that is being acquired (…)” (Howells 2002, p. 872). Tacit knowledge, on the other hand, is described as "non-verbalised, intuitive and unarticulated knowledge" (Polanyi 1962). It represents a form of know-how that is developed by informal acquired behaviours and procedures (Howells 2002, p. 872). For the purposes of this study, we refer to knowledge solely in terms of the concept of explicit knowledge.

Furthermore, this study focuses only on research related and novel knowledge. The scope comprises knowledge and technologies, which are potentially relevant for future innovation processes and are novel to the scientific community. This includes all recent research outcomes by a university, but excludes widely known and commonly accepted knowledge. Therefore, alone novel scientific insights, technological innovations, like leading edge technologies shape the scope of this study.

2.2 Knowledge Transfer

Knowledge transfer and technology transfer are in the body of literature extremely interrelated concepts and thus often used in an exchangeable manner (Agrawal 2001, Grimpe & Hussinger 2013, Sung & Gibson 2000). We, however, focus on knowledge transfer overall, but acknowledge that the term ‘technology transfer’ is, in certain cases, a more accurate description of the issue. A closer look at the literature on knowledge and technology transfer reveals that most studies omit to deliver a clear definition of knowledge
transfer (Liyanage et al. 2009).

We aim to set common ground by explicitly defining knowledge transfer, in accordance with Argote and Ingram (2000, p. 152), who define knowledge transfer broadly as: “the process through which one unit (e.g., individual, group, or division) is affected by the experience of another.” However, this study requires an additional definition, which includes more precisely the mechanisms and outcomes of the transfer. Hence, we expand this notion by including the aspect from the notion of Liyanage et al. (2009, p. 123) describing that “(…) a knowledge transfer process has two main components, i.e. the source or sender that shares the knowledge, and the receiver who acquires the knowledge.” For our purpose it is crucial that the emphasis lays on the fact that the for the transfer to be evident it must be measurable on the receiver’s end. While the variety of definitions in literature, given the above mentioned constrains, this paper will follow closely the final definition again suggested by Liyanage et al.(2009, p. 122) who sees knowledge transfer as “(…) 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.” This definition is particularly appropriate as it includes the necessity of utilization of the transferred knowledge, pointing out the criteria for successful transfer.

2.3 Formal and Informal Knowledge Transfer

The research on university based knowledge transfer to industry is divided into two main categories: “formal” and “informal” knowledge transfer. Some scholars define formal knowledge transfer mechanisms as such, which eventually "result in a legal instrumentality such as, for example, a patent, license or royalty agreement (…)" (Arundel & Bordoy 2008, p. 642), while informal knowledge transfer is seen as a transfer resulting from different forms of informal communication, including consulting or collaborative research (Link et al. 2007). However, we view this definition of ‘informal’ transfer as still defining mainly a formal forms of transfer, as it is still based on formalized agreements pursued under contracts between the two entities. Hence, research joint ventures, and university-based start-ups would be a form of formal knowledge transfer (Link et al. 2007). Therefore we follow a less common understanding of informal knowledge transfer including transfer, which is not based on property rights and the exchange may refer to personal contacts, informal use of data bases, workshops or similar. Here the obligations between the partners are more normative than actually legal (Fernández- esquinas et al. 2015, Grimpe & Hussinger 2013, Link et al. 2007). Overall, it is evident that the main attention in university knowledge transfer has been
given to the formal knowledge transfer including their mechanisms, successful commercialization of inventions, impacts and similar (Link et al. 2007). We aim to consider both types in our study.

2.4 Qualitative and Quantitative Approaches

Informal aspects of knowledge transfer are mostly studied in qualitative studies, like case studies. These are often capable of capturing various transfer channels, motives behind collaboration and similar. They provide in-depth insights into potential transfer mechanisms, motivations and rationales of the actors and more (Brostroem 2012, Franco & Haase 2015). Qualitative studies focus often on single cases, including at best national contexts and provide in-depth understanding about potential impacts and benefits of certain activities, projects or particular innovations (Ankrah et al. 2013). Qualitative case studies fail to provide measurable and generalizable results and offer in-depth insights only into very specific scenarios. This limits the comparability of the findings, in particular since knowledge transfer is highly dependent on exogenous influences like policies, legal structures, funding etc. Moreover, many studies fail to actually verify the content of the transferred knowledge, or the extend of it (Rothaermel et al. 2007, Salter & Martin 2001).

Quantitative studies, on the other hand, often deal with the overall contributions of formal knowledge transfer between universities and industry. They aim to capture mainly economic and/or socioeconomic impacts. The level of analysis ranges from firm and university to regional and national comparative studies. However, these studies mainly capture commercialization of products or technologies and revenue generating usage of more or less finalized inventions derived from research knowledge (Cheah 2016). Main proxy indicators for quantitative studies are, among others, licenses and license agreements (Jensen et al. 2003), patents, including patent citations (Arundel & Bordoy 2008, Thursby & Thursby 2002), co-publications by firm and universities (Tijssen et al. 2009), and different kinds of entrepreneurial efforts, like university spin-outs and their generated revenues (Vincett 2010).

However, these indicators face long-standing criticism about their incapability to capture the majority of transferred knowledge. Some of these proxies simplify the transfer to a plain commercially measureable value (e.g. royalties) and others fail to capture the collaborative relationship and focus on potentially never utilized knowledge (e.g. co-publications) (Cheah 2016, Lundberg et al. 2006). Many quantitative studies combine the investigation of formal knowledge transfer, in terms of commercialization with qualitative methods, like expert interviews, to capture a more holistic picture of the knowledge transfer and the collaboration in general (Cohen et al. 2002, Siegel
The indicators for economic impact of quantitative assessments are ‘(…) difficult to obtain and generally suffer from long lag times between public investment and outcomes’ (Arundel & Bordoy 2008, p.6). Besides, it has been pointed out that they fail to provide a holistic picture. Some scholars argue that the measurements are only accounting for very low percentages of actual knowledge transfer (Cheah 2016, Lundberg et al. 2006).

Given the limitations of contemporary empirical work we contribute to the body of academic literature by addressing some of the deficits of the current metrics. We aim to provide a novel measurement method that helps to diminish the limitations. Our method provides in-depth insights about the transferred knowledge. Ideally even traceable to university department, or academic scientist level. It is supposed to provide statistical correlation measures comparable to the ones of patents or licenses analyses. We offer a great extend of independence from the concepts of formal and informal knowledge transfer channels, aiming at an additional more holistic way of capturing knowledge transfer. The approach is less dependent on the examination of external and internal circumstances, as it measures the transferred items and does not require additional assumptions about the potential of knowledge exchange. However, the commercialization is not directly measurable in terms of patents or revenues, but in combination with the traditional measures could provide comparatively precise estimations.

3 METHODS

Texts contain information, and extracting this information has become an increasingly developed part of today’s research fields of machine learning and computational linguistics. Enormous insights in various disciplines were generated via the use of text mining tools over the past decades. Content and sentiment analysis are of increasing relevance in computer science and machine learning during the past decades and the tools advancement is getting more and more promising (Chapman & Hall/CRC 2010, Collobert et al. 2011).

In our case text mining is an appropriate strategy, since text material can be a sufficient data source to detect knowledge transfer. First, academic publications in form of scientific texts, such as journal articles, conference proceedings or books, contain the main outcomes of scientific research. They are seen as output and dissemination channel of university research (Stahl et al. 1988, Toutkoushian et al. 2003). Therefore these publications are texts containing data for all major research findings of a university.

Second, online presences (like websites) are media for companies to dis-
play their novel products, services, R&D strategies and innovations, as web-
sites, blogs, videos or social media entries. These online presences of firms
are mainly in text form and firms place high value on these to ensure their
visibility for potential consumers and investors leading to regular updates
and R&D descriptions (Branstetter 2006, Heinze & Hu 2006).

These two types of texts provide insights into the use and generation
of knowledge. Therefore, the use of statistical tools from NLP is an ideal
approach to identify commonalities in terms of correlations.

3.1 Pre-processing

Text pre-processing converts unstructured raw text into statistical and com-
putational useful units. Pre-processing is part of any text analytic procedure
and might very well be decisive for the later outcome of the analysis. The
quality of the results is highly depending on the thoroughness of the pre-
processing. The main objective is to capture all relevant characters, erase
obsolete items. To identify actual words via the detection of word separation
(tokenization). This enables the application of further text mining methods
(Paukkeri & Honkela 2010).

Pre-processing of text includes:

- To define word boundaries as white space,
- To delete unwanted elements (e.g. special characters, punctuation and
  numbers),
- To convert upper case to lower case characters,
- To remove ‘stopwords’
- To stem the words.

Results of our pre-processing revealed some challenges in the case of the
academic abstracts. These abstracts contain, for instance, chemical formu-
las and notations, which rely heavily on numbers and/or special characters.
These are unfortunately lost during the course of the pre-processing. The
only possibility to later identify the same formulas and to use them for sim-
ilarity measures is the assumption that the removal will always result in an

1Stopwords are the most common words in a language, which are not carrying content
relevant information

2Describes the process of reducing words to their word stem or root form. It is a process
for removing the morphological endings from words: connected, connection, connections
become ‘connect’.
identical character string. The result may not be identifiable as the specific formula, but still provide a match \(^3\). However, particularly short strings derived, from formulas or notations, are lost during the pre-processing. Some terms seem to be the result of poor pre-processing, but are in reality just a representation of specific models, formulas or project names shrunk to a unidentifiable string of characters. The pre-processed texts are merged into structured units: the text corpora. All the following methods are based on these corpora, which are an a way of structuring documents as well as organizing them into meaningful content related units.

3.2 Document Term Matrix:

A document-term matrix is the most common vector space representation of a document corpus. It contains the feature (term) frequencies for each document. Rows correspond to documents and columns to terms.

A document-term matrix is usually generated from pre-processed corpora, which results in a representation of semantically and contextual relevant terms (Chapman & Hall/CRC 2010). As document-term matrices are usually highly dimensional and sparse; hence many of the current models aim for sensible dimensionality reduction (Berry & Castellanos 2007). 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 \((d)\).

\[
\text{Document-Term Matrix}(w, d) = \begin{pmatrix}
    a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
    a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{m,1} & a_{m,2} & \cdots & a_{m,n}
\end{pmatrix}
\]

Various schemes determining the value of each entry in the matrix can take have been developed, which are the term weighting schemes. The weight for each term can be derived in various forms from frequencies of the term occurrences. The weighting of terms differ widely depending on the models used. Common weighting schemes include, among others:

- The binary weighting, the entry takes values 1 or 0 depending on whether or not a term occurs,

- Term-frequency (TF), the actual number of times a term occurs,

\(^3\)In some cases HTML tags prevent the identical construction. In this case we did not find a way to identify the matching strings.
• Term-frequency, inverse document frequency (TFIDF), uses TF but assigns higher weight to terms that occur only in a small number of documents.

3.3 Term-frequency, inverse document frequency

For the purpose of this study we chose to use the TFIDF indexing to determine the 50 most characteristic words per document. In case of the academic abstracts we retrieved often less than 50 words. The reduced dimensionality enabled a comparison of keyword lists with each other. We use these lists, generated for each document, to identify common terms between two types of documents, abstracts and website pages.

The TFIDF is a simple numerical indexing method, which has been applied in various contexts (Franceschini et al. 2016, Zhang et al. 2016). It has proven to give respectable results on its own, especially considering its simplicity. However, it is also used in various more advanced models, such as Vector Space Model (VSM) or Latent Semantic Analysis (LSA) (Mao & Chu 2007).

TFIDF can be used to enable a dimensionality reduction providing a small set of content relevant terms, which capture the main content of a document. Further, TFIDF is an indexing scheme that allows identifying the most relevant words by extracting the words most unique to a given text. The principal assumptions are simple: a word that occurs often in a document is relevant for its content (of course after the stopword removal), but words that are additionally used in many documents are less specific for the single document and therefore less relevant. Hence, frequent words that are used in many different texts are seen as carrying less contextual information and obtain a lower score in the TFIDF weighting. The scheme has different proposed calculations, but most commonly the TFIDF weight 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 $dw$.

\[
TF(w, d) = \sum w_i
\]

\[
IDF(w, D) = \log\left(\frac{D}{dw}\right)
\]

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

The TFIDF approach suffers from three main shortcomings:
First, as it calculates term weight based on term frequencies, it might represent only the content of a text fragment, since terms with a high frequency may be only used in a certain part of a document. This is a major drawback especially for long texts.

Second, IDF assumes that terms, which rarely occur over a collection of documents, are more content related, while in reality it just makes it more distinct from the other documents in the collection. So a corpus about, for instance, water issues would probably score the term ‘water’ low, which does not capture the reality of the document content.

Third, empty terms and function terms, like adverbs or modal particles, are often assigned too high scores, which leads to inaccurate weight. Unfortunately, even a thorough stopword removal is not preventing this from happening (Xia & Chai 2011).

3.4 Latent Dirichlet Allocation

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 & 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 (2012, p.15):

1. For each topic: decide what words are likely (term distribution described as $\beta \sim Dirichlet(\delta)$)

2. For each document:
   
   (a) 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. choose a topic ($z_i \sim Multinomial(\theta)$).
      
      B. given this topic, 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 select the optimal number of topics ($K$), we chose to approximate the marginal corpus likelihood (depending on $K$) by taking the harmonic mean of the corpora after applying the LDA. 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. This is an approximation of $p(w|K)$, i.e., the likelihood of the corpus given the number of topics (Ponweiser 2012).

We limited the maximum number of topics for the corpora of websites in each case to 50 and found that only 3 web corpora might have benefited a larger number, we chose this limitation for computational efficiency reasons. For the academic abstracts, the calculation for the optimal number of $K$ was set to a maximum of 200. This was chosen due to the diversity in academic specialized fields. However, results show that the optimal number of topics only rarely exceeded 50.

We chose to set the hyper-parameter ($\alpha$ and $\beta$) so that they allow a more diverse topic distribution over a single document by enforcing more topics per documents with lower probabilities\(^4\). This is appropriate, since we are not trying to classify the documents but working to fine grain the content of the documents to an extent that captures context and topics of text snippets. To improve the performance 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 very specified, especially after the additional pre-processing step.

We used the obtained 50 words per topic with the highest probability for this particular topic and returned them as list of keywords. We compared to other lists of topic keywords from LDAs from academic corpora and web corpora. The resulting topic pairs show the most similar corpora in terms of their underlying structures.

### 3.5 Jaccard Similarity Coefficient

For the similarity measure between the sets of identified keywords found by applying TFIDF or LDA, we used the Jaccard similarity coefficient as metric. It is a statistic used for measuring the similarity between sets. 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:

\(^4\)We used Gibbs sampling in the LDA model to draw from the posterior distribution. For more information on determining the posterior probability of the latent variable, refer to (Grün & Hornik 2011)
We chose this similarity measure as it only includes element presence in a given set. It is applicable for the LDA and TFIDF generated keywords, as it does not need an input of scores or probabilities, which would not be comparable since they resulted from different corpora. Another advantage is the low computational expense, making it attractive for a basic similarity assessment. However, this could of course be refined by applying additional similarity measures to find more accurate matches. The thresholds for a minimum similarity for further examination were chosen based on previous manual examination; meaning that we would only consider keyword lists with a certain minimum Jaccard similarity as relevant for the manual inspection and potential text matching. 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 index with the intersection of the two sets, giving higher weight to sets with a higher amount of common words. A set of pairs with Jaccard Similarity lower than 0.15 needs more than 7 words in common in order to pass the criteria, while set pairs with Jaccard Index higher than 0.15 can have smaller intersections.

3.6 Sample

The following description of the sample is divided into the generation of the text corpora, representing a) university and b) industry knowledge. Two main data sources were needed for the analysis: academic publications representing university research output and a collection of relevant texts from firms. The methods are applied to data from the Technical University of Denmark (DTU).

The university publication database, Orbit, provided data that include a collection of academic abstracts from university research publications. These abstracts present a summary of the main research outputs by employees of the university between the years 2005 and 2016.

Firm websites are the second data source for this study, providing the company knowledge. Criteria for relevant websites are a) a national (Danish) registered branch of the firm b) at least some English fragments of the firm website, and c) the firm must have been a ‘partner’ of the university between 2013 and 2016.

5The types of relevant partnerships are explained in the next section of the article.
**Publication database**

Given legal challenges to obtain a comprehensive sample of full text publications, we chose available abstracts to serve as proxy of the university’s research output. Approximately 60% of the publications during those years were stored with an abstract. The data availability in the database increases from 2012 onwards here approximately 80% of all abstracts are publications available. The selection criteria required that an entry is a) published between 2005 and 2016, b) has an available abstract and c) is least co-authored by one member of the university staff. These criteria resulted in 43,745 hits while the total number of all publication records is 76,627.

We classified the abstract by their assigned departmental codes, provided by the database, to enable a pre-classification of the texts on the basis of their research field. Both methods, LDA and TFIDF, perform better on more content coherent corpora, as this to enables a better performance of the statistical analysis. In particular in the case of the LDA, one single corpus as input would result in identifying the overall research giving us almost no additional insights. The classification resulted in 24 separate research fields, while three of these were irrelevant for the academic output of the university. The collection of corpora based on academic abstracts is in the following referred to as ‘academic’ corpora or by the individual research field, if it is relevant for the interpretation of the results.

**Firm Websites**

To identify relevant firms for the sample we performed two major steps. First we collected based on Danish companies with a formal connection to the university, namely a collaboration contract. We identified 686 Danish firms, which had a contract with the university between the years 2013 and beginning of 2016. The firms in this sub-set operate mainly in technology intensive sectors and are firms with strong R&D divisions. Therefore it included companies with contents similar to the research performed at the university. Second we generated a network on the basis of hyperlinks between websites using the university as point of origin, identifying the university’s partners linked to the university website. Partners of those partners (second-degree partners) of the university were hereby also identified and added. These websites content were downloaded and stored as HTML files. The list of examined websites contained many online service platforms, including for

---

6We excluded 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 firm (this could have biased the findings significantly as the firm is directly involved in several hundreds of dedicated publications).
example public transportation sites, yellow pages and firm registries. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample. The online text samples were collected between August 2016 and November 2016.

To ensure a connection of the firms to Denmark each page of each website was subsequently scanned for a Danish firm registration number (CVR) and in case one was found the website was added to the sample. Unfortunately, in Denmark, universities, schools for higher education and other public entities are registered via the firm registration number; they had to be manually excluded. Finally the language of the website and/ or the sub-pages was verified and only content with more than 60% English content was stored.

The total sample contains 599 Danish websites, containing English pages with a total of 148939 sub-pages (documents). 464 websites provided more than 5 English pages and were converted into single corpora and used for the TFIDF application. Due to more extensive pre-processing procedures the number of useful websites for the LDA was 404 websites. The number of pages and length of documents varies a great deal between the firm websites. Some provide just an English summary for their main contents, while others, often multinational firms, have their entire website in English, which influences the model performance. One major drawback in our sample collection is the partial absence of PDFs or similar formats stored since these require special treatment for each format.

4 RESULTS and DISCUSSION

The results of this study are divided into the application of the different methods. We aim to provide in-depth details about the performance of each single tool and algorithm. Additionally, we describe interrelated components and the results generated via a combination of those methods.

4.1 LDA

As described earlier, the LDA is a representation of the hidden structures of the content of a given text corpus determined through a set of topics. The main words per topic show an adequate representation of the overall topics of the corpora. It means that for example the themes of the abstracts in the corpus of Chemistry are represented in 37 topics.

This extract shows that the words are representing overall topics of the academic corpus quite satisfactorily.
Table 1: The top 5 words for academic topics

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>enzym</td>
<td>bind</td>
<td>dynam</td>
<td>forc</td>
<td>oligosaccharid</td>
<td>indic</td>
</tr>
<tr>
<td>domain</td>
<td>site</td>
<td>vibrat</td>
<td>particl</td>
<td>branch</td>
<td>chain</td>
</tr>
<tr>
<td>amino</td>
<td>conform</td>
<td>motion</td>
<td>hydrophob</td>
<td>carbohydr</td>
<td>complet</td>
</tr>
<tr>
<td>residu</td>
<td>enzym</td>
<td>coupl</td>
<td>friction</td>
<td>donor</td>
<td>size</td>
</tr>
<tr>
<td>express</td>
<td>residu</td>
<td>excit</td>
<td>layer</td>
<td>polysaccharid</td>
<td>mean</td>
</tr>
</tbody>
</table>

Table 2: The top 5 words for website topics

<table>
<thead>
<tr>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
<th>Topic 11</th>
<th>Topic 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>speci</td>
<td>chemistri</td>
<td>optim</td>
<td>raman</td>
<td>treatment</td>
<td>situ</td>
</tr>
<tr>
<td>ester</td>
<td>analyt</td>
<td>factor</td>
<td>spectra</td>
<td>strontium</td>
<td>redox</td>
</tr>
<tr>
<td>previous</td>
<td>research</td>
<td>occur</td>
<td>band</td>
<td>bone</td>
<td>electrochem</td>
</tr>
<tr>
<td>iridoid</td>
<td>uniqu</td>
<td>design</td>
<td>laser</td>
<td>reduc</td>
<td>stm</td>
</tr>
<tr>
<td>isol</td>
<td>european</td>
<td>reveal</td>
<td>mixtur</td>
<td>treat</td>
<td>microscopi</td>
</tr>
</tbody>
</table>

However, in the case of the websites we observe a more diverse outcome with less diversification among the topics. One could say that the single topics within the web corpora are less coherent and provided more heterogeneous themes than the academic ones. The keywords of the topics of the web corpora seem to be more generic. This is attributed to the length of texts (abstracts are shorter than websites) and the content diversity (abstracts contain mainly one single theme).

In order to capture relevant pairs of academic abstracts and website texts we decided to combine three different approaches to compare the keywords between the LDAs.
Academic topics vs. Websites topics

The first approach is to compare the similarity of each topic from the web corpora with each topic of the academic corpora. Each topic identified by the LDA is converted into a simple word list. The Jaccard similarity measure was used to compare and match word lists. The number of list comparisons made to assess the topic similarity was 17,417,025.

Given the chosen Jaccard threshold (as explained in the method section), only $1.3 \times 10^{-4}$% (23 pairs) exceeded the required similarity for the topic per topic comparison. 9 of these matches were Danish language fragments embedded in the topics. They were removed and the remaining 14 matches were manually inspected. The common words are comparatively generic, but show relatedness. The common words indicate a clear overlap in the topics, but are slightly unspecific, as they lack any reference to relevant models, notations, formulas or relevant proper nouns. This might very well be a result of the fine tuning of the LDA and the LDA’s potential to identify fields rather than specific content.

<table>
<thead>
<tr>
<th>12 common words between 2 topics word list</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 function</td>
</tr>
<tr>
<td>2 gene</td>
</tr>
<tr>
<td>3 dna</td>
</tr>
<tr>
<td>4 express</td>
</tr>
<tr>
<td>5 isol</td>
</tr>
<tr>
<td>6 microorgan</td>
</tr>
<tr>
<td>7 cell</td>
</tr>
<tr>
<td>8 strain</td>
</tr>
<tr>
<td>9 bacteri</td>
</tr>
<tr>
<td>10 bacteria</td>
</tr>
<tr>
<td>11 communiti</td>
</tr>
<tr>
<td>12 popul</td>
</tr>
</tbody>
</table>

Table 3: The top 10 most common words academic topics vs website topics

Interestingly most common words within the remaining 14 pairs were based on 8 distinct corpora. One corpus alone accounted for almost a third the total matches. The corpus that accounted for that many pairs was the corpus, which contains ‘diverse’ research areas that could not meaningfully be fitted in any other department. This was surprising, but given the corpus diversity it would represent the most mixed research topics which are present in several websites.
This corpus lacks the coherence and specificity of the other academic corpora and is therefore highly

**Academic topics vs. Websites**

The second approach is to combine all topics of one website corpus into a word list representing the entire corpus and compare this web corpus list with each individual topic of the academic corpora. The academic topics are more defined or specific and are therefore the adequate choice. The goal is to find new relevant abstracts and website matches.

Given a new Jaccard threshold \((0.45 \times 10\text{ words})\), due to the increase in number of potential common words we obtained \(4.6 \times 10^{-3}\%\) (17 pairs) matching pairs for 370.575 comparisons. No purely language based pairing occurred in this instance.

The most common words were comparatively generic and the number of their occurrence was rather low, with a maximum of nine co-occurrences, showing that the pairing was based on comparatively diverse keywords.

<table>
<thead>
<tr>
<th>10 common words</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

**Table 4**: The top 10 most common words topics vs websites

**Departments vs. Websites**

The third approach is to create one combined word list for each of the web corpora and a second combined word list for each of the academic corpora. These comparatively long keyword lists are subsequently compared with each other.

The average keywords list per academic corpus contained 1900 keywords, of course depending on number of topics. The web corpora had an average 2300 keywords per corpus list.
We compared the total keyword lists of both and find the next pairs. For the corpus based keyword comparison, we had only 8505 comparisons and due to the high number of words per corpus list we had to adjust the Jaccard similarity thresholds to a minimum Jaccard similarity score of 0.2 and a minimum word intersection matching of $0.18 \times 500$ words. We obtained 41 positive matches, which contained non based on foreign language fragments. The main corpus matches were based on several Departments. The combination of websites was, again, diverse without any clear patterns or specific corpus domination.

We compared in the outcome of the LDA comparisons with the manual investigation conducted after the application of the TFIDF.

<table>
<thead>
<tr>
<th>Academic Topic vs. Web-pages Topic</th>
<th>Academic Topic vs. Websites</th>
<th>Departments vs. Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Diverse</td>
<td>Energy Conversion</td>
<td>Management Engineering</td>
</tr>
<tr>
<td>2 Mechanical Engineering</td>
<td>Electrical Engineering</td>
<td>Diverse</td>
</tr>
<tr>
<td>3</td>
<td>Diverse</td>
<td>Civil Engineering</td>
</tr>
<tr>
<td>4</td>
<td>Civil Engineering</td>
<td>Compute Math</td>
</tr>
</tbody>
</table>

Table 5: Results for the LDA

### 4.2 TFIDF

The TFIDF indexing resulted into a set of keywords for each single document, for both academic and website corpora. This resulted in 3,343,890,411 comparisons. Every match (a comparison, which exceeds the threshold) was stored as text pair for later manual assessment. We excluded multiple matches between the same academic abstract and the same website (but different web-page within the site). We kept only the match with the highest score Jaccard similarity score, because some companies display the same texts on more than one page. However, we left matches that referred to the same university department and the same website, but to a different abstract in the sample. Since abstracts of the same department are less likely to be identical than text snippets on the same website.

We found exactly 100 pairs that exceeded the chosen Jaccard similarity threshold. However, after some manual investigation of the outcomes we found that we had to exclude matches that were based on country names. Additionally, some matches were based on other language fragments entailed in the abstracts and the websites. These were mainly displaying German,

---

10A full exclusion of country names for future applications is considered, but seemed not necessary for the current sample.
French and Danish content. This was solved by the application of a simple language filter which identified all Danish, German and French key words in the sets and removed the match in case it had more than 5 hits. After this removal, the total matching pairs was decreased to 88.

The left matching keyword matches were manually assessed and we found that the dominant words were highly diverse and many entailed not real words. The character strings derived from trademarks, proper nouns, models, software names and formulas, were most present and helpful to identify relevant matches. Terms like ‘novirhabdovirus’, or ‘mxgs’ (Gamma ray Sensor module (MXGS)) account for a number of hits.

After the quality assessment of the keywords per document we paired the relevant texts and manually checked their similarity. We then classified the pairs into 7 different categories.

1. Identical topic = University contribution
2. Identical topic = Potential university contribution
3. Identical topic = Unlikely university contribution
4. Identical topic = Newspaper article about university
5. Different topic = No match in content
6. Identical topic = University contribution to a public entity
7. Unclear = could not be classified

The manual classification was undertaken taking into consideration the full text publication, since in many cases the abstract would not provide sufficient information to establish whether contents are actually related.

Additionally, we had to make qualitative distinctions between the entities, which display the university research, since all of them fulfill our requirements, but not all of them are actually private firms using the research. We found newspaper articles presenting university research; we found several public entities (with CVR numbers), which use and promote the university research. These can be seen as correct pairing from the TFIDF (true positives), but show that the differentiation between public and private entities needs to be improved. To exclude newspaper articles and there like might be rather challenging, but with a news registry this might be achievable.

The results present a relatively high number of content related, but unlikely truly related matches (category 3) this shows that the TFIDF finds
related content, but an additional measure to minimize these hits would be beneficial. Even more so for the pairs off classification label 5, which provided not even the same content and result in a high number of false positives.

Figure 1: Results classification

![Histogram of occurrences of matches orders](image1)

Figure 2: Results classification with classes 1, 4 and 7. combined

![Histogram of occurrences of matches orders](image2)
The academic departments that occurred in the pairs of documents were diverse, but certain departments dominated the pairing. The main academic departments found to be most often in the pairs were as indicated in Table 6. The websites on the contrary were very diverse and no particular websites were matched more frequent than the others. We had a maximum of five hits to the same web-page.

<table>
<thead>
<tr>
<th>Category 1</th>
<th>Category: 1 and 2</th>
<th>Category: 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1    Bio Systems</td>
<td>Bio Systems</td>
<td>Management Engineering</td>
</tr>
<tr>
<td>2    Electrical Engineering</td>
<td>Electrical Engineering</td>
<td>Mechanical Engineering</td>
</tr>
<tr>
<td>3    National Space Institute</td>
<td>National Space Institute</td>
<td>National Food Institute12</td>
</tr>
<tr>
<td>4    Diverse13</td>
<td>Diverse</td>
<td></td>
</tr>
<tr>
<td>5    National Veterinary Institute</td>
<td>National Food Institute14</td>
<td></td>
</tr>
<tr>
<td>6    National Food Institute</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Departments present in the results

The overall identification of the TFIDF went surprisingly well, given that the majority of academic abstracts contain only between 100 and 200 words. In many cases the human verification needed additional information to the abstract (like the full text of a publication) to ensure the match was an actual true positive. This is very promising; especially considering the improvements that would be possible with a full text sample from the academic publications. For example the following abstract text provides enough textual information for a class 1 pair:

'A method for reproduction of sound, based on crosstalk cancellation using inverse filters, was implemented in the context of testing telecommunications devices. The effect of the regularization parameter, number of loudspeakers, type of background noise, and a technique to attenuate audible artifacts, were investigated. The quality of the reproduced sound was evaluated both objectively and subjectively with respect to the reference sounds, at points where telecommunications devices would be potentially placed around the head. The highest regularization value gave the best results, the performance was equally good when using eight or four loudspeakers, and the reproduction method was shown to be robust for different program materials. The proposed technique to reduce audible artifacts increased the perceived similarity.' (Gil Corrales et al. 2015)

11Category entails category 7, since it is a performance indicator of the text mining application and not a performance measure for our firm/ non-firm classification.
4.3 Combining Results: LDA and TFIDF

Given the findings from the TFIDF and LDA potential improvement of the outcomes is possible if the results are combined. Hence, we used the different outcomes of the LDA to identify academic corpora that are more prone to generate false positives. The TFIDF true positives and false positives were reduced to their departments and these were compared to the LDA results.

Corpora suggested as relevant in the LDAs are mainly accounting for general content matches by the TFIDF, meaning that these corpora are at least having related contents with the matching websites.

The topic to topic comparison does not capture the most relevant corpora for the later found matches. However, this could suggest that there are more matches within the data, which are not yet uncovered. We assume that additional measures would be adequate to improve this outcome. However, we found that if a department was more than once identified in the academic topic to website topic comparisons the department was likely a true positive pair.

In the case of the comparison of academic topics to website corpora lists, we found that a department, found in this combination, is also likely to be a true positive. None of the identified departments was only in the false positive collection, hence a match here also suggests a content relation between texts.

In the final case, one keyword list for an entire website corpus was compared to one keyword list for an entire academic corpus, we found that one department was only present in this pairing and here over-represented. It also accounted for 25% of all false positives in the TFIDF application. The academic corpus is the one of the department of Management Engineering and is noticeable due to its generic keywords in both LDAd and TFIDF. This could suggest that this last LDA comparison focuses on similar patterns as the TFIDF when identifying the false positives.

If a department was not in any of the LDA comparisons, it was also likely to be a false positive. However, it was not possible to detect the false positives that share a department with a true positive. For this matter an investigation of the websites could be considered.

5 CONCLUSION

What are the implications of these results for the use of text mining as metrics in the studies of university-industry knowledge transfer?

First, our results show that there are many occurrences of commercially used knowledge transfers, which are not necessarily only identifiable via
patent, license agreements or similar. It clearly shows that university research is used and displayed on firm websites and that these instances are computational traceable. Hence, we are confident that we can confirm Agrawal and Henderson (2002) findings who stated that patents present only a small fragment of the knowledge transfer between universities and the industry. Even though our sample size is not large enough to estimate the extent of additional knowledge transfer that can be identified via our method, we can say for certain that we captured additional knowledge transfer.

Second, we see that our findings are in agreement with the notion that certain academic fields are more prone to knowledge transfers than others. This confirms the notion that the transfer of applied sciences is more frequent than the one of basic research. However, since our sample is currently limited to one case it does not yet provide generalizable results.

6 LIMITATIONS

The limitations of our study are numerous and are technical as well as conceptual. First, the data on academic research outcomes is limited, since abstracts hardly display the true output of the research. The use of abstracts was necessary due to availability issues and copyright issues for full-text publications. However, in future we aim to complement the data with full-text publications. Second, the manual classification is not ideal as it is time intensive, especially since the text pairs are often hard to understand and therefore difficult to classify. It often requires expert knowledge from the specific research field. We hope to address this shortcoming in future by building a computational classifier that would at least propose a first potential classification, which would only have to be verified by human inspection. Third, technically we could have used further text mining methods to improve the results. For this purpose we suggest to include other machine learning approaches in the future; in particular word2vec vector and correlated topic modeling (CTM). Fourth, we aim to perform a more traditional analysis with traditional metrics, including patents and license agreements, to verify the actual additional component of our approach and compare the results.

Finally, we need to implement a metric that aims to measure the actual impact of the knowledge presented by the company. Currently we only aim at the binary measure whether knowledge is transferred or not. It would be relevant to assess how important this specific knowledge is for the firm. This could enable a clear measurement of knowledge transfer contribution.
References


Innovation 18(2), 212–232.

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, 

are the factors underlying the variety of interactions with industry?’, Research Policy 36(9), 1295–1313.


nological Forecasting & Social Change Tracing the flows of knowledge transfer : Latent dimensions and determinants of university – industry in-
teractions in peripheral innovation systems’, Technological Forecasting & Social Change 113, 266–279.

communities about sustainability and innovation. A bibliometric journey 
around sustainable terms’, Journal of Cleaner Production 127, 72–83.

Franco, M. & Haase, H. (2015), ‘University-industry cooperation: Re-

Geuna, A. & Rossi, F. (2011), ‘Changes to university IPR regulations in Eu-
rope and the impact on academic patenting’, Research Policy 40(8), 1068–1076.

Gil Corrales, J., Song, W. & MacDonald, E. (2015), Reproduction of Re-
alistic Background Noise for Testing Telecommunications Devices, Audio Engineering Society.


