Inter-technology Relationship Networks: Arranging Technologies through Text Mining

by

Peter Hofmann[a], Robert Keller[b,c], Nils Urbach[a,d]

appears in:
Technological Forecasting and Social Change (2019)

The final publication is available at:
https://www.doi.org/10.1016/j.techfore.2019.02.009

Affiliations:
[a] Project Group Business & Information Systems Engineering of the Fraunhofer FIT, 95444 Bayreuth, Germany
[b] FIM Research Center, University of Augsburg, 86135 Augsburg, Germany
[c] Project Group Business & Information Systems Engineering of the Fraunhofer FIT, 86159 Augsburg, Germany
[d] FIM Research Center, University of Bayreuth, 95444 Bayreuth, Germany

E-mail addresses:
peter.hofmann@fit.fraunhofer.de (P. Hofmann)
robert.keller@fim-rc.de (R. Keller)
nils.urbach@fim-rc.de (N. Urbach)

Inter-technology Relationship Networks: Arranging Technologies through Text Mining

Abstract:

Ongoing advances in digital technologies – which enable new products, services, and business models – have fundamentally affected business and society through several waves of digitalization. When analyzing digital technologies, a dynamic system or an ecosystem model that represents interrelated technologies is beneficial owing to the systemic character of digital technologies. Using an assembly-based process model for situational method engineering, and following the design science research paradigm, we develop an analytical method to generate technology-related network data that retraces elapsed patterns of technological change. We consider the technological distances that characterize technologies' proximities and dependencies. We use established text mining techniques and draw from technology innovation research as justificatory knowledge. The proposed method processes textual data from different information sources into an analyzable and readable inter-technology relationship network. To evaluate the method, we use exemplary digital technologies from the big data analytics domain as an application scenario.

Keywords:

Text Mining; Network; Tech Mining; Patent Mining; Method Construction

Highlights:

- *A text mining-based method arranges technological relationships in a dynamic network*
- *Similarities between technology-related corpora quantifies technologies' relatedness*
- *Separating text processing pipelines allows one to incorporate different textual information sources jointly*
- *The average relatedness of technologies within the big data analytics landscape increased between 2007 and 2017.*

1. Introduction

Evaluating new technologies' impacts on existing technology landscapes is a relevant issue for both researchers and practitioners. Ongoing advances in digital technologies – which enable new products, services, and business models – have fundamentally affected business and society through several waves of digitalization (Bharadwaj et al. 2013; Legner et al. 2017). Emerging technologies now include blockchain, deep reinforcement learning, 5th generation mobile networks, and virtual assistants (Gartner 2017). Although it is hard to forecast technological advances and trends (Adomavicius et al. 2007; Daim et al. 2006), companies must decide which of the many emerging technologies are worth adopting, developing, or examining. A further complicating factor is that historical cases indicate that invention processes do not proceed in a uniform way (Arthur 2007).

Commonly used information sources that provide insights into the evolution of technologies include scientific publications and patent documents (Engelsman and Van Raan 1994) as well as academic proposals, business news, and social media (Zhang et al. 2016). Besides bibliometric data (e.g., authors, citations, keywords), much of the information is only available in (unstructured) text form. The availability of a large corpus of written resources, often readily accessible via the internet, is both a blessing and a curse. While the presence of a large corpus is desirable, manually reading and interpreting a whole corpus is extremely time-consuming as well as complex, and thus often exceeds human information processing capacities (Debortoli et al. 2016; Fan et al. 2006; Hao et al. 2014; Tseng et al. 2007). Nonetheless, it is crucial to utilize an extensive variety of written resources, since it allows one to overcome information silos and thus to develop a – valuable – systematic understanding. Among other approaches, text mining is a promising approach to overcome these limitations. It offers new opportunities for qualitative and quantitative research in the IS domain (Debortoli et al. 2016). Text mining, which originated as knowledge discovery in textual databases (Feldman et al. 1995), generally seeks to automatically extract unknown information from textual data (Fan et al. 2006; Gandomi and Haider 2015; Hearst 1999). Michel et al. (2011), for instance, demonstrate this methodological potential by automatically analyzing word

frequencies in more than five million books so as to observe cultural trends. Among others, they show that the cultural adoption of technology has accelerated since the 19th century. In technology innovation research, different research streams share the approach to extract value-adding information from technology-related documents using text mining (Madani and Weber 2016). Since patent documents are a valuable source of technical knowledge (Gupta and Pangannaya 2000; Lee et al. 2009b), there are many approaches, often subordinated to patent mining or patent analysis (Nakamura et al. 2015). Without limiting itself to patent documents, tech mining or technology mining includes the application of text mining to technology management purposes (Madani 2015; Porter and Cunningham 2005).

However, to understand the underlying patterns of technological change and the emergence of new technologies, we cannot evaluate technologies individually. When analyzing digital technologies, a dynamic system or an ecosystem model that represents interrelated technologies is beneficial owing to the systemic nature of digital technologies (Adomavicius et al. 2008). In this context, an IT ecosystem is "a subset of information technologies in the IT landscape that are related to one another in a specific context of use." (Adomavicius et al. 2008, p. 783). In this ecosystem model, technologies are assemblies of component technologies that can also consist of subordinate component technologies or assemblies (Adomavicius et al. 2008; Arthur 2007). Innovation research prominently considers the recombination of existing technology components or modules as a relevant source of invention or innovation (Fleming and Sorenson 2001; Schoenmakers and Duysters 2010), which emphasizes the importance to incorporate a technology-overlapping perspective. On the one hand, components-based relationships may directly arise as a set of technological components intersect (Aharonson and Schilling 2016). On the other hand, the concept of knowledge-relatedness indicates other ways in which components-based relationships between technologies may arise: knowledge proximity, knowledge commonalities, and knowledge complementarities. While knowledge proximity refers to the results of learning processes (i.e. learning spillovers or local learning), knowledge commonalities may occur as two technologies overlap in their required knowledge. Considering knowledge complementarities, relationships may occur between

technologies that are dissimilar but technically build on one another (Breschi et al. 2003). Thus, one research strategy to retrace elapsed patterns of technological change and to deduce (future) directions of technological development is to consider the technological distances that characterize the proximities and dependencies of technologies or technology fields (Aharonson and Schilling 2016; Schoen et al. 2012). Further, Breschi et al. (2003) describe that the relatedness between technologies influence technological diversification decisions. Building on the ecosystem model, we hypothesize that the proximity of technologies indicates the imminent combination of component technologies into a new innovation.

In this context, there are different methods in the academic literature that arrange technology-related entities in structured representations such as graphs, networks, or maps via text mining techniques (Engelsman and Van Raan 1994; Yoon and Park 2004). Other related research streams without a specific technology focus include knowledge maps (e.g., Hao et al. 2014), science maps (e.g., Klavans and Boyack 2009; Leydesdorff and Rafols 2009), and ontologies (e.g., Navigli et al. 2003; Pesquita et al. 2009). However, the academic literature lacks a text mining method that processes unstructured text to accomplish the following method engineering goals: First, enabling purposeful investigations, we need a text mining method that systematically arranges predeterminable technologies or abstractions of these. Depending on the research task at hand, it may be necessary or useful to conceptually aggregate associated technologies to an abstract technologies set (Arthur 2009). For instance, we can speak concretely about *Apache Spark* or switch to a more granular level and consider *large-scale data processing systems* as the abstraction level (Zaharia et al. 2016). Second, as change occurs over time, we require a dynamic perspective on changing technological distances that allows for in-depth longitudinal analysis. Third, as different technology-related information sources offer different insights, the method needs to be able to incorporate information from different sources. Thus, our research question is:

How can an analytical method using text mining techniques be developed that arranges predefined technologies into a dynamically interpretable inter-technology relationship network?

To answer this question, we develop an analytic method following the design science research (DSR) paradigm in IS research (Gregor and Hevner 2013; Hevner et al. 2004; March and Smith 1995). An assembly-based process model for situational method engineering complements the DSR framework in the construction phase of our method (Brinkkemper 1996; Ralyté et al. 2003). We use established text mining techniques and draw from technology innovation research as justificatory knowledge. We contribute to the literature by providing an initial method assembly that supports technology and innovation management as well as research by systematically arranging technologies in an inter-technology relationship network. We also offer a comparison of two method variants and contrast the results to the assessment of human judgment as well as an alternative count-based approach to verify the results' plausibility and to show the necessity of the method. For this, we used exemplary digital technologies from the big data analytics domain as an application scenario.

The remainder of this paper is structured as follows: In Section 2, we discuss existing approaches to contextualize our method in the existing literature and to provide design knowledge relevant to our study. In Section 3, we describe our research approach. In Section 4, we introduce our proposed method, which processes textual data from different information sources into an analyzable and readable inter-technology relationship network. In Section 5, we review our evaluation activities. We conclude by summarizing our findings and discussing its limitations and future research.

2. Literature Review

Existing approaches that map technology-related entities concentrate on patents and scientific publications by using structured data such as bibliometric features, unstructured data in the form of text, or a combination (Aharonson and Schilling 2016; Yan and Luo 2017). Further, we recognize a research stream that is especially popular in the biomedicine and expert systems domain that uses existing ontologies to calculate semantic similarity measures (Lord et al. 2003; Pesquita et al. 2009; Sánchez et al. 2012). However, to avoid restricting our method's applicability only to domains for which ontologies exist, we don't follow

this approach any further. We will now first introduce bibliometric analysis approaches, to subsequently describe the more recent trends towards text-based measures in Section 2.2.

2.1 Bibliometric Analysis

Many early mapping attempts used bibliometric analysis. This limits the evaluation of documents to counts of scientific publications, patent documents, or associated citations to reproduce scientific and technological advances (Narin et al. 1994; Porter and Detampel 1995). In bibliometric analysis, citation-based measures gained much attention, including co-citation analysis, which relies on the number of concurrent citations of two documents in other documents, allowing one to retrieve the extent of a relationship (Small 1973; Small and Griffith 1974). In doing so, networks consisting of scientific articles allow one to follow the development of research fields, since influential articles have high connectivity (Furukawa et al. 2015). Alternatively, citation data can be processed in the form of citing-to-cited relationships between documents (Leydesdorff and Vaughan 2006). Leydesdorff et al. (2014), for instance, arranged patent classes by applying the cosine similarity index on vectors of a citing-to-cited matrix. Kay et al. (2014) used a similar approach on adjusted patent classes. Boyack et al. (2005) compared the results based on co-citation and citing-to-cited relationships. Lastly, bibliographic coupling describes the similarity of two documents by quantifying their shared references (Egghe and Rousseau 2002). Citation-based measures also have major drawbacks: varying citation rates in academic domains diminish inter-disciplinary comparability (Klavans and Boyack 2009). Owing to time lags in the patenting or publishing process, citation-based measures cannot cope with very recent technologies (Yoon and Kim 2011). Further, the necessity of data with citations limits its applicability. Further, citations don't cover the entire relationship structure (Engelsman and Van Raan 1994). In particular, legal and economic rather than knowledge-mapping concerns are prevalent in patent citations (Leydesdorff 2008). For this reason, patent classification-based measures were developed to offer an alternative approach. Patent classification-based measures use information arising from the assignment of patent classes, which are originally used to support information retrieval (Engelsman and Van Raan 1994; Joo and Kim 2010). The issuing patent office refers to a fixed

classification scheme such as the International Patent Classification (IPC) system (Breschi et al. 2003; Leydesdorff 2008). The basic idea behind co-classification is that the more often two patent classes co-occur in patent documents, the smaller the technological distance between them (Breschi et al. 2003; Engelsman and Van Raan 1994; Joo and Kim 2010). However, bibliometric analysis disregards potential insights hidden in the textual data (Engelsman and Van Raan 1994; Lee et al. 2009b), which motivates text-based approaches to measure technological distance (Furukawa et al. 2015; Lee et al. 2009a; Yan and Luo 2017). Swanson (1987) exemplarily showed that bibliographic analysis may not unveil unknown yet logically existing relationships between medical literatures.

2.2 Beyond Bibliometric Analysis

Some of the earlier text-based approaches rely on comparing the occurrence of keywords, following the idea of a vector space model (Yoon et al. 2013). For instance, Ding et al. (2001) applied co-word analysis using scientific publications' keywords to create a bibliometric cartography of the field of information retrieval. The co-word approach assumes that the co-occurrence of keywords represents content-related associations (Callon et al. 1991; Lee et al. 2008). However, the exclusive use of the keywords provided by the authors or editors of a technology-related document has the disadvantage of disregarding a large part of the textual content. Yoon and Park (2004) and Lee et al. (2009a), for instance, used keyword vectors (composed in advance by extracted keywords) that quantify the occurrence of a keyword in a patent document. The relationships between patents is then calculated by the Euclidian distance index between these keyword vectors. In contrast, Lee et al. (2009b) used keyword vectors to apply a principal component analysis. Predefining or extracting meaningful keywords, particularly in the case of emerging technologies, requires much effort and expert knowledge (Yoon et al. 2011). Further, predefining or extracting meaningful keywords directly influences the results and thus decreases a study's objectivity.

To avoid the necessity to predefine keywords, Yoon et al. (2011) used the co-occurrences of properties (i.e. adjectives in patent documents) and functions (i.e. verbs in patent documents) to create a patent network. The parts of speech are also relevant to approaches using the subject-action-object (SAO) patterns to

retrieve content-related similarities between patents. Yoon and Kim (2012), for instance, approached technological distance between patents by calculating sentence similarity between SAO patterns extracted from patent documents. Within SAO patterns, subjects and objects consist of noun phrases, while action refers to verb phrases (Yoon et al. 2013). Thus, sentence similarity is calculated using the *WordNet* semantic dictionary (Miller 1995). This leads to the drawback that the method only calculates similarity based on words covered by the dictionary or a manual extension. By considering the patent application dates, Yoon et al. (2013) extended the SAO approach to create a dynamic patent map. Without concentrating on parts of speech, Nakamura et al. (2015) calculated the similarity between patent clusters by applying the cosine similarity index to weighted word vector representations of patent titles and abstracts.

In contrast, we followed the recent development toward text-based measures, because we see untapped potentials. The research gap consists in the fact that a text-based approach that follows a dynamic, multisource perspective to measure technological distances between predefinable technologies remains unresolved. We apply a broad understanding of *technology*, as per Arthur (2007), defining it as any means that serves a human purpose (e.g., method, process, or device). This allows us to avoid limiting our method's applicability to a specific technological context. Some approaches already compile technologies or technology fields (i.e. abstractions of technologies) using bibliometric features (e.g., Nakamura et al. 2015; Schoen et al. 2012).

3. Research Method

To develop our analytic method, we followed the DSR paradigm. DSR, practiced in technical disciplines such as computer science and software engineering for decades, is also an accepted research approach in the information systems field (Hevner 2007; Iivari 2007; March and Storey 2008). DSR seeks to create artifacts that serve human or organizational purposes instead of understanding reality in the case of natural science (Hevner et al. 2004; March and Smith 1995). Thus, DSR provides a research process that enables rigorously building and evaluating viable artifacts and addresses important and relevant business problems

(Hevner et al. 2004). The DSR methodological approach commonly includes identifying the problem, defining a solution's objectives, designing and developing an artifact, demonstrating how the artifact solves the problem, evaluating the artifact, and communicating the research (Peffers et al. 2007).

Next to constructs, models, and instantiations, artifact types include methods (Hevner et al. 2004; March and Smith 1995); thus, DSR is an appropriate methodological framework for our study. Generally, methods specify how to perform goal-directed activities (March and Smith 1995). We will now introduce the applied approach to construct our method and will outline our evaluation strategy in Section 3.2.

3.1 Assembly-based Method Construction

While using DSR as a methodological frame for our overall design process, we complement principles of situational method engineering (Henderson-Sellers and Ralyté 2010) for rigorously constructing our method. Method engineering uses existing methods to "design, construct and adapt methods, techniques and tools for the development of information systems." (Brinkkemper 1996, p. 276). In this discipline, situational method engineering seeks to create methods that fit specific IS development situations (Henderson-Sellers and Ralyté 2010). Thus, a diverse set of procedure models for situational method engineering exist in the academic literature (Bucher and Winter 2008). We build on the assembly-based method engineering approach by Ralyté et al. (2003) and apply it, as illustrated in Figure 1.

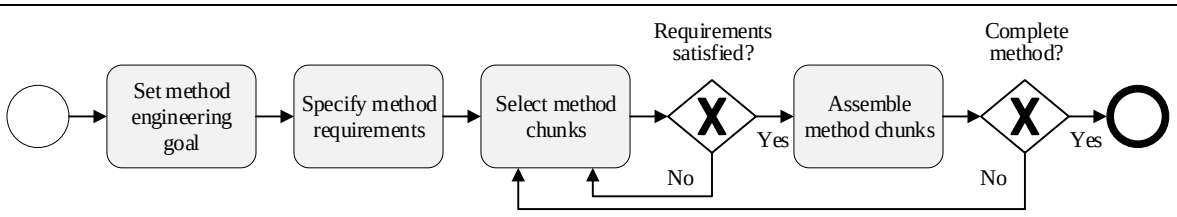

Figure 1. Assembly-based method engineering approach

Step 1 (i.e. setting a method engineering goal) incorporates the definition of what the method should achieve and the decision whether to start from scratch or to take an existing method to build on (Ralyté et al. 2003). We have already delineated our method engineering goals in the introduction section. We constructed our method from scratch, but adhered to existing approaches introduced in the literature review section. Step 2

(i.e. specifying method requirements) involves the identification of requirements that candidate chunks must fulfill (Ralyté et al. 2003). We address these requirements in the artifact description section. Steps 3 and 4 (i.e. selecting and assembling method chunks) iterate until the method engineer accomplishes a complete solution) (Ralyté et al. 2003). Strategies that support the selection of method chunks include decomposing, aggregating, and refining existing methods. Selected method chunks that satisfy the requirements need to be assembled until the composed method meets the completion conditions (CCs) (Ralyté et al. 2003). We set our CCs as follows:

(CC1) Artifact assembly fulfills targeted method engineering goals (i.e. the intended input-output transformation).

(CC2) Each method chunk fulfills requirements.

(CC3) The data processing – including all steps – is transparent and verifiable.

(CC4) The method processes large-scale corpora within a reasonable time.

Although the introduced methodological approach originated from the construction of IS development methods, it is also suitable for constructing our analytic method: First, it allowed us to rigorously combine existing text mining techniques (i.e. method chunks) to develop a novel analytical method that satisfies previously specified requirements and CCs. This construction process iterates in the form of a search process, as proposed in DSR (Hevner et al. 2004; Hevner 2007). We conducted this design process in several iterations of feedback and testing. Second, we could use justificatory knowledge in the design process as recommended by Gregor and Hevner (2013). Thus, we relied on academic literature from the innovation, technology, and text mining research. This procedure allowed for the conceptualization of the natures of technologies and their relationships as well as to extract method chunks from existing (technology-related) text mining approaches. Notably, this paper is about the development of an overall assembly, rather than about configuring specific method chunks in detail.

3.2 Evaluation Strategy

Besides creating an artifact, its thorough evaluation is crucial so as to demonstrate the artifact's inherent utility (Hevner et al. 2004; March and Smith 1995). According to March and Smith (1995), the evaluation criteria for methods comprise ease-of-use, operationality (i.e. feasibility and effectiveness), efficiency, and generality. We implemented a prototype as an instantiation of our method (March and Smith 1995) to apply the method to an exemplary scenario and to discuss the method's utility.

To evaluate effectiveness, we compared the two method variants' methods with each other as well as with human judgment. Comparing automatically calculated similarity measures to the scores of human judgment is a common approach in computational linguistics (e.g., Arts et al. 2017; Lapata 2006). To collect human judgment, we conducted semi-structured interviews according to Myers and Newman (2007), in which we proceeded as follows: First, we introduced ourselves and explained the interview's purpose. Second, we introduced our understanding of technologies and the manifold possibilities of how relationships between technologies may occur. Third, each interviewee completed a symmetric adjacent matrix. Fourth, we processed the adjacent matrix into a network representation, allowing a participant to visually verify its ratings. In case of discrepancies, we returned to step 3. Finally, we discussed the challenges that occurred when completing the adjacent matrix and asked the interviewee how well they assessed their rating performance. Experiences from the pre-tests resulted in the iterative completion of the adjacent matrix. The interviewees showed strong appreciation of this proceeding, since it allowed them to more accurately assess the complex structures of technological relationships. As participants, we selected eight researchers from our research network and two practitioners associated with our research network. Every interviewee had distinct knowledge in the technology landscape in question. The interviews lasted approximately 60 to 90 minutes each. We also discussed face validity against the following assumptions: If a document contains two technologies, there is a relationship between these technologies is. Accordingly, the more documents that satisfy the above condition, the more reliable (not necessarily stronger) the relationship between these technologies. Notably, we did not apply the reverse conclusion (i.e. if there is a relationship between

technologies, they also occur together in documents). An in-depth evaluation regarding generality and efficiency, falls outside the scope of this work. We discuss the fulfillment of the CCs so as to conclude the evaluation.

4. Artifact Description

Based on self-assembled corpora associated with predefined technologies, our method creates an inter-technology relationship network using text similarity measures. In this regard, we assume that the textual similarity between technology-specific corpora indicates the relatedness between technologies. This assumption is in line with text-based approaches introduced in the literature review. To simplify matters, we further assumed that the relatedness between technologies is commutative (i.e. relatedness between technology A and B = relatedness between B and A). To avoid terminological confusion, we will equate technological relatedness and technological proximity and will disregard the opposing notion of technological distance. This is not detrimental to our study, since one can easily convert a similarity measure into a distance measure via inversion or subtraction (Turney and Pantel 2010).

By *inter-technology relationship network*, we mean an ordered sequence of undirected, weighted multigraphs with the edges' weight representing the technological relatedness. Thus, we combined the principles of traditional graph theory (Bollobás 1998; Newman 2003) and time-varying networks (also referred to as evolving, temporal, or dynamic networks). Time-varying networks are exposed to topological changes as a function of time (Casteigts et al. 2011; Holme and Saramäki 2012). In particular, we applied the model analogous to Ferreira (2004), depicting the longitudinal development as a sequence of static multigraphs, each of which represents a time stamp. Here, the use of a multigraph allowed us to distinguish between multiple weighted edges resulting from different technology-related document types (Bollobás 1998), and thus to incorporate for instance patent documents and scientific publications jointly into a single network. Describing the technological space with the help of networks allowed us to maintain as much of

the characteristics of the original high-dimensional technological space and to use an abundance of existing network analysis methods.

To create an inter-technology relationship network, we proposed an individual processing pipeline for each data type. The network sequences resulting from the different data sources were then merged into a sequence of multigraphs. We separated the processing, since each technology-related document type has a textual style. Figure 2 illustrates the overall design of the method, and Figure 3 specifies the structure of the data type-specific processing pipelines, which we explain in some detail in Sections 4.1, 4.2, and 4.3.

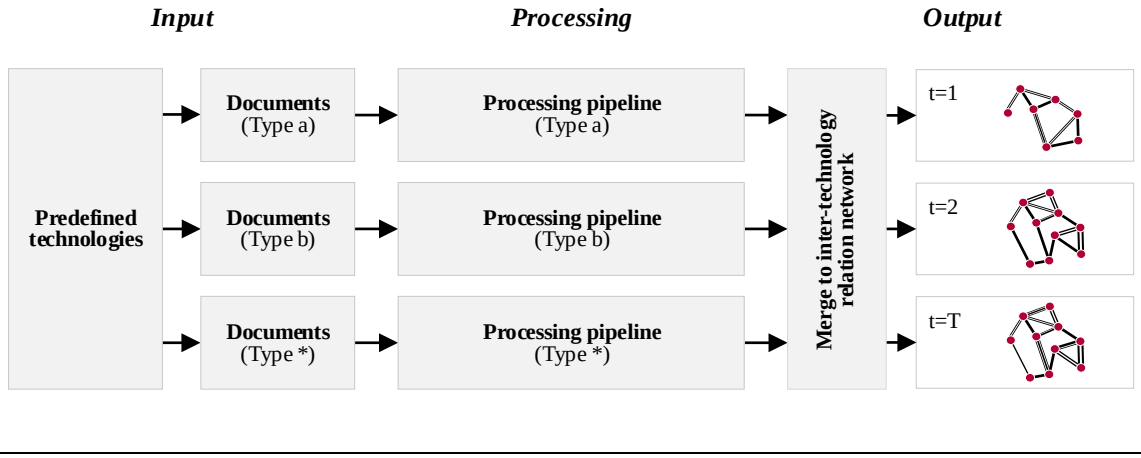

Figure 2. Method overview

The input of our method consists of a list of predefined technologies and technology-related documents. Each technology of interest must be specified in advance with a name and an allocation query (e.g., *'mapreduce' OR 'map reduce' OR 'map-reduce' in the patent title OR abstract*). To compile a list of technologies, one may refer among others to the results of previously conducted technology scouting activities, the academic literature, and existing taxonomies or ontologies. Ontologies for instance may provide precisely defined terms in a specific domain (Maedche and Staab 2002).

Preprocessing then seeks to process technology-related documents such that they provide the necessary data basis for the subsequent processing steps. We refer to the processed documents as assets. To suit our

method, each asset must contain at least a time classification parameter and text. This requirement barely restricts usable data sources. Thus, our method is capable of integrating various data sources such as patents, scientific publications, news articles, or social media posts. We used the processed assets to create technology-specific corpora, each of which represents the development of a predefined technology. This requires techniques to adequately allocate and condense assets to form technology-specific and time-varying corpora. Based on these corpora, we applied text similarity measures to derive time-varying relatedness matrices. This required text similarity measures that describe technological relatedness based on content-related rather than syntactic similarity of the technology-related corpora. Finally, we used the relatedness matrices as adjacency matrices to construct the networks. To avoid loops (i.e. relationships that connect an individual technology to itself), we manually set the diagonal elements to zero. The user may apply a cutoff value, to limit the networks to more pronounced relationships or to eliminate noise.

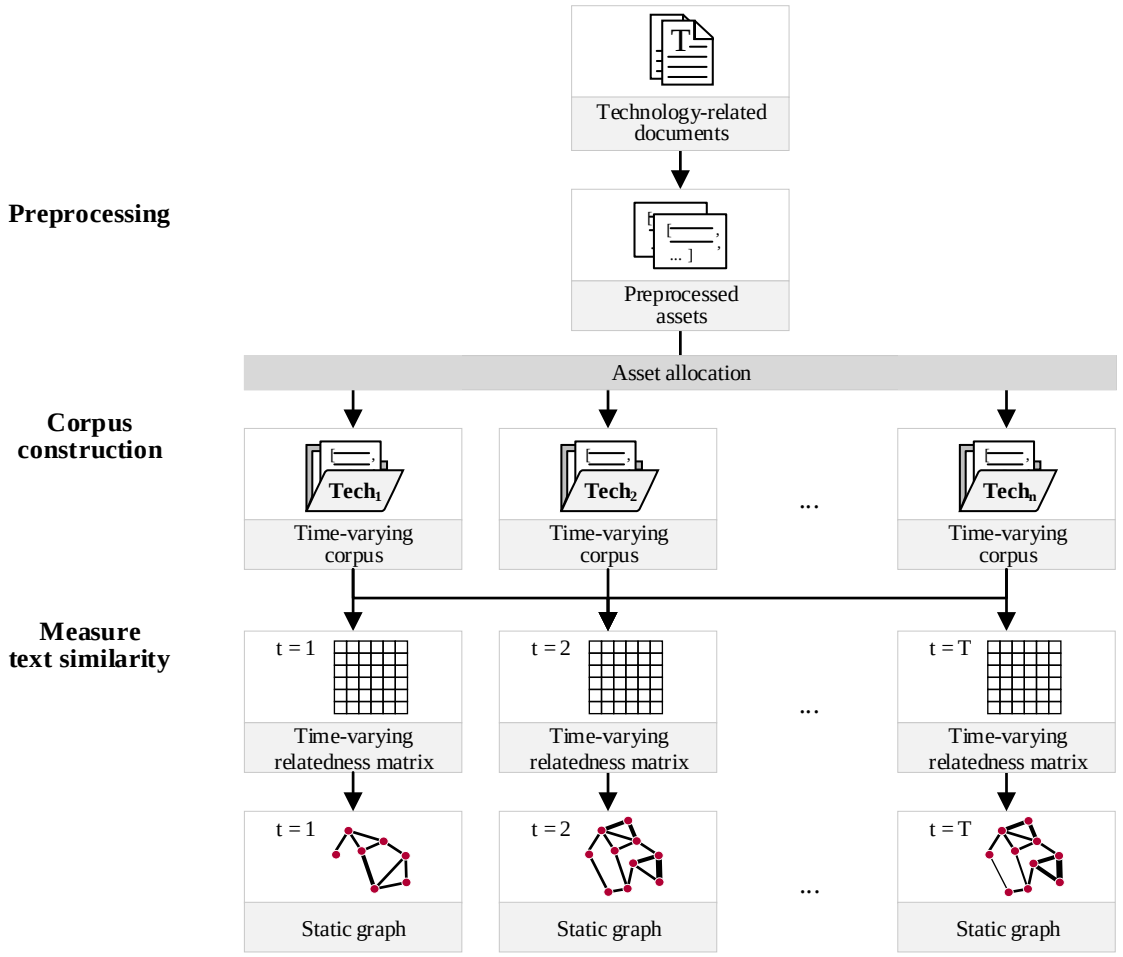

Figure 3. Data type-specific processing pipelines

4.1 Preprocessing

Preprocessing involves providing the right data structure and text normalizing to increase the effectiveness of the subsequently applied text similarity measures (Turney and Pantel 2010). While extracting a time classification parameter from an asset is hardly a problem for any asset type, preprocessing deals primarily with the preparation of the text. To meet the data structure requirements of most text mining algorithms, one must tokenize plain text into single words (Manning et al. 2009). Since the expression that specifies a technology may consist of several words, we recommend recombining single tokens to multiword expressions based on a predefined thesaurus of technology expressions. An example is ‘machine learning.’ We recombined the single tokens ‘machine’ and ‘learning’ into ‘machine_learning’ with an underscore.

Normalizing text by converting superficially different terms into the same form increases the effectiveness of the subsequently applied text similarity measures (Turney and Pantel 2010). Thus, terms of morphological variants match, which would otherwise not coincide (Hull 1996). We applied lemmatization to canonicalize tokens (Manning et al. 2009). For lemmatizing, we used the morphy function of the lexical database *WordNet* (Miller 1995). Further, we suggest removing stop words (e.g., of, the, that), which often occur but contribute little to the content (Manning et al. 2009; Turney and Pantel 2010), and using all lowercase letters. While the introduced preprocessing steps worked well with formal language appearing for instance in patent documents and scientific publications, adapted preprocessing methods are necessary for information sources such as news articles or social media. For this reason, we applied a processing pipeline for each data type individually. This example illustrates the preprocessing steps.

Raw string: 'We applied Machine Learning algorithms in the cloud'
Case folding: 'we applied machine learning algorithms in the cloud'
Tokenizing: 'we', 'applied', 'machine_learning', 'algorithms', 'in', 'the', 'cloud'
Lemmatizing: 'we', 'apply', 'machine_learning', 'algorithm', 'in', 'the', 'cloud'
Filtering stop words: 'apply', 'machine_learning', 'algorithm', 'cloud'

4.2 Corpus Construction

By allocating assets to individual technologies, we obtained a selection of assets for each technology. There are multiple approaches to achieving this, since the problem shows parallels to information retrieval. For instance, Schoen et al. (2012) and Nakamura et al. (2015) used patent classes to identify technologies. For the sake of simplicity, we stuck with a simple exact term matching approach by assigning a document to a technology when a predefined, technology-specific query was true for an asset. Thus, we consider terminology as the surface appearance of technologies (Navigli et al. 2003). Table 1 describes an example of the asset allocation in the case of patent documents. Avoiding false positive allocations, we applied allocation queries for patent documents only to patent titles and abstracts.

Technology name	Patent allocation query	Asset 1	Asset 2	Asset 3	…
MapReduce	'mapreduce' **OR** 'map reduce' **OR** 'map-reduce' **in** the patent title **OR** abstract	True	False	False	
Sentiment Analysis	'sentiment analysis' **OR** 'sentiment-analysis' **OR** 'opinion mining' **OR** 'opinion-mining' **in** the patent title **OR** abstract	False	False	False	…
…	…	…	…	…	…

Table 1. Example of asset allocation for patent documents

We used the assigned assets to create technology-specific and time-varying corpora. To accomplish this, we suggest to accumulate the assets' terms into a sequence of terms for every technology on an annual basis in the case of patents and scientific publications. Thus, a technology-specific corpus A_t at time t equals the sequence of terms from A_{t-1} appended by the terms of the assets at time t. This proceeding is based on the assumption that both patents and scientific publications refer to all existing knowledge. Nonetheless, it is necessary to verify whether this assumption is also valid for other data types (e.g., Twitter).

4.3 Measuring Relatedness

The academic literature offers a wide range of approaches to determining similarities between texts. The count-based bag-of-words approach (BOW) is one of the most common approaches to measuring the text-based similarity between documents (Le and Mikolov 2014). The underlying hypothesis of the BOW approach is that the more two texts resemble each other in their word frequencies, the more similar they are (Li et al. 2006; Salton et al. 1975; Turney and Pantel 2010). Following the distributional hypothesis (Harris 1954), the Vector Space Model describes a document as a vector of terms (Salton and Buckley 1988). Given the technologies $tech_1, \dots, tech_n$ and the terms contained in the entire corpus $w_1, \dots, w_m$, we defined a term-technology matrix as $A_t \in \mathbb{R}^{m \times n}$ where a_{ij} represents the occurrence of the term w_i in the corpus of technology $tech_j$ at time t. Based on the term-technology matrix, the distance or angle between the column vectors allows one to quantify the similarities between the technology-related corpora (Salton et al. 1975; Salton and Buckley 1988). In this context, the academic literature commonly refers to cosine similarity (Lowe 2001; Turney and Pantel 2010). This yields to a symmetric technology-to-technology relatedness matrix $R_t = (rel_{n,m})$ at time t.

With the objective to improve the basic BOW approach's performance, advanced techniques commonly transform the count-based word vectors. This includes primarily reweighting document vectors and smoothing them (Baroni et al. 2014). Reweighting seeks to improve document vectors' distinctness by increasing the influence of surprising events as opposed to expected events (Turney and Pantel 2010). We selected the commonly used term frequency inverse document frequency (tf-idf) weighting (Manning et al. 2009; Salton et al. 1983), referring to it as BOW-tf-idf. Thus, a term in a document has a higher weight if it appears more often in this document and the less often other documents contain the term. This yields to a weighted term-technology matrix A_t. Visualizing large corpora, existing approaches use matrix decomposition and factorization methods such as singular value decomposition (SVD) and principal component analysis (Gretarsson et al. 2012). SVD, as a part of latent semantic analysis (Landauer and Dumais 1997) or latent semantic indexing (Dumais et al. 1988), reduces the dimensionality of a (weighted) term-document matrix (Landauer et al. 1998) or, in our case, the term-technology matrix. Dimensionality reduction tries to address the effect of synonymy (i.e. different terms describing the same object) and polysemy (i.e. the same term describing different objects) (Deerwester et al. 1990). As in the basic BOW approach, the cosine is applied to the documents' vector representations (Bradford 2008). Thus, the definition of the dimensionality reduction parameter is complex, which directly influences the results (Bradford 2008).

As a disadvantage, the BOW approach ignores word order, so that two documents may have the same word vectors, although they express different content (Le and Mikolov 2014). However, for large corpora, the abundance of co-occurring words weakens the effects of neglecting syntactic information (Li et al. 2006). Recently, prediction-based measures, such as the Word2Vec model (Mikolov et al. 2013), became the subject of discussion (Baroni et al. 2014; Zhu and Iglesias 2017). The Word2Vec model creates distributed word vectors by learning to predict words from its context (i.e. surrounding words) (Mikolov et al. 2013). While the Word2Vec model enables vector representations of words, it does not natively support vector representation of multiple words or whole documents. Thus, Le and Mikolov (2014) proposed an

unsupervised framework (often referred to as Doc2Vec) that produces document vectors or, in general, paragraph vectors in which text length can range from sentences to whole documents (Le and Mikolov 2014). Doc2Vec provides two approaches to learn document vectors, namely the distributed memory model (used in this study) and the distributed BOW version (Le and Mikolov 2014). The distributed memory model trains both word vectors (analogous to Word2Vec) and document vectors (Le and Mikolov 2014). We then applied the cosine to the document vectors to retrieve their similarity. We will now compare the results of BOW-tf-idf and Doc2Vec.

5. Evaluation

For our illustrative scenario, we consider the case of big data analytics, which significantly affects academics and business (Agarwal and Dhar 2014). Being relevant to various domains, big data challenges all actors to cope with new technologies for storing, processing, and analyzing a massive amount of data (Buhl et al. 2013; Hu et al. 2014) and therefore changes the corporate landscape (Chen et al. 2016). Big data's ambitions involve the volume, variety, velocity, and veracity of data, as well as the value from data (Chen et al. 2015). According to Hu et al.'s (2014) layered architecture model, the big data system affects the infrastructure layer (i.e. the pool of computing, networking, and storage resources, including cloud infrastructure), the computing layer (i.e. middleware, including data tools such as data integration, data management, and a programming model), and the application layer (i.e. the interfaces provided by the programming models for various data analysis functions). The data used for big data analytics ranges from structured, to semi-structured, unstructured, and real-time data (Kambatla et al. 2014). We selected this technology landscape as our illustrative scenario, for the following reasons: First, since data analytics has been discussed for several years, a profound data basis exists. Second, new questions and opportunities concerning the availability of data and machine intelligence are changing the technology landscape (Agarwal and Dhar 2014). To apply our proposed method to this scenario, we extracted a set of 32 technologies by relying on the academic literature. We consider this technology set as a sample excerpt and do not claim that it is exhaustive.

To prevent the interviews from exceeding a reasonable extent and from becoming too complex, we limited the number of technologies. Thus, we randomly selected 11 technologies from the technology set and added Bluetooth to control against false positive results, since technologies in the big data analytics ecosystem are somewhat interrelated. We contained ratings from interviewees using a rating scale ranging from 0 (i.e. technologies bear no relation to one another) to 100 (i.e. technologies are identical or always co-occur together).

We acquired patent grant full-text data issued weekly from January 2007 to December 2017 from the bulk data storage system of the United States Patent and Trademark Office (2018) on February 11, 2018. While parsing the provided XML files, we filtered the patents so that only technical patents (i.e. utility patents) that fulfilled classification requirements remained. Our patent data comprises patents classified by the International Patent Classification (IPC) as G (i.e. Physics) and H (i.e. Electricity) or by the Cooperative Patent Classification (CPC) as G (i.e. Physics), H (i.e. Electricity), and Y (i.e., among others, new technological developments and cross-sectional technologies). We initially filtered patent data moderately on the basis of their classification in order to reduce the influences of term ambiguities, but also to avoid excluding relevant patents. As part of the subsequent asset allocation, technology-specific queries then filtered irrelevant patent documents more strictly. We used the application year as the time classification parameter. Regarding patenting activities, it seemed sufficient to rely only on U.S. patent data, since they are a reliable representation of patenting activities (Leydesdorff et al. 2014) and we are not addressing legal issues. However, one may also incorporate patent data from other patent offices, such as the European Patent Office or the Japan Patent Office. We retrieved titles, abstracts, and publication years of scientific publications for 2007 to 2017 from Web of Science (Clarivate Analytics 2018) on April 15, 2018. We applied the following search specification:

"(TS=("Cloud Computing" OR "Cloud Service" OR "Cloud-based" OR "Infrastructure as a service" OR "Software as a Service" OR " Platform as a Service") OR TS=(Big AND Data*) OR TS=("Data Science" OR "Data Analytics" OR "Data Analysis" OR "Data Mining" OR "Text Mining") OR TS=("Machine Learning" OR "Neural Network") OR TS=("Bluetooth") OR TS=("Speech recognition") OR TS=("Stream Processing" OR "Stream Processor") OR TS=("Image Analysis") OR TS=("In-memory Computing" OR "In-memory database") OR TS=("NOSQL") OR TS=("Internet of things")) *AND* **LANGUAGE:** (English) *AND* **DOCUMENT TYPES:** (Article) **Refined by: WEB OF SCIENCE CATEGORIES:** (ENGINEERING ELECTRICAL ELECTRONIC OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR COMPUTER SCIENCE INFORMATION SYSTEMS OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE SOFTWARE ENGINEERING OR COMPUTER SCIENCE CYBERNETICS OR INFORMATION SCIENCE LIBRARY SCIENCE OR COMPUTER SCIENCE HARDWARE ARCHITECTURE) **Timespan:** 2007 to 2017. **Indices:** SCI-EXPANDED."

In sum, patent data acquisition yielded 14,859 patents that matched both patent classification requirements and our patent allocation queries. The abovementioned search specification resulted in 59,400 scientific publications. We will now evaluate our method by discussing its effectiveness (Section 5.1), feasibility, ease-of-use (Section 5.2), and fulfillment of the CCs (Section 5.3).

5.1 The Method's Effectiveness

For each technology pair, we derived a set of relatedness scores, including the results of the BOW-tf-idf and the Doc2Vec approaches, as well as human judgment. Owing to the limited dimensionality reduction potential resulting from the relatively small number of nodes in our exemplary technology set, we did not use the SVD approach. To compare the scores, we used Kendall's τ (Kendall 1938) to apply a non-parametric rank correlation test with the null hypothesis of no differences between the scores. By using a rank correlation coefficient, we addressed the feedback of some interviewees that, despite the scale used, they were unable to provide parametric scores. Table 2 contains the results of the comparison, using Kendall's τ. Each cell in the correlation matrix represents the correlation between the results of the method variants in the case of patent data and scientific publications. Thus, the method's results are based on the smaller technology subset so as to be comparable to human judgments.

	Patent data			Scientific publications		
	BOW-tf-idf	**Doc2Vec**	**Averaged human judgments**	**BOW-tf-idf**	**Doc2Vec**	**Averaged human judgments**
BOW-tf-idf	1	0.329*	0.123	1	0.254*	0.173*
Doc2Vec	0.329*	1	0.171*	0.254*	1	0.108
Averaged human judgments	0.123	0.171*	1	0.173*	0.108	1

Note: * two-sided p-value < 0.05.

Table 2. Correlation between the results

Table 2 indicates significant correlation between the method variants and the averaged human judgment only in certain instances. However, while the human judges agreed for some technology pairs (e.g., machine learning and neural network), there were significant differences for other technology pairs (e.g., speech recognition and image analysis). This disagreement between human judges manifested in an average standard deviation of 21.2 and a Kendall's coefficient of concordance (Kendall and Smith 1939) of 0.58. This discrepancy occurred, although the interviewees at least considered their assessments to be good; however, they substantiated their relatedness scores with different rationales. Thus, we don't regard the averaged human judgments as a conventional gold standard. Nonetheless, the results underline the importance of a transparent and verifiable method that complements human judgments. Besides this, Table 2 depicts that the method variants have weak positive, significant correlations between one another. Although the method variants shared some agreement, it is apparent that they weigh the relationships differently.

Directly comparing the results of patent data and scientific publications using the same method variants based on the smaller technology subset with the null hypothesis of no differences between the scores yielded a Kendall's τ of 0.535 (two-sided p-value < 0.05) for BOW-tf-idf and -0.08 (two-sided p-value > 0.05) for Doc2Vec. This comparison highlights that, especially for Doc2Vec, the data source that is used influences the results and thus provides a different perspective on the technology landscape. To discuss face validity,

we followed a visual approach, providing three charts (Figure 4) that illustrate the number of documents in which a technology pair co-occurs ordered by the results of the individual method variants descends from left to right. To calculate the values of the method variants, we applied the prototype implementation to patent data.

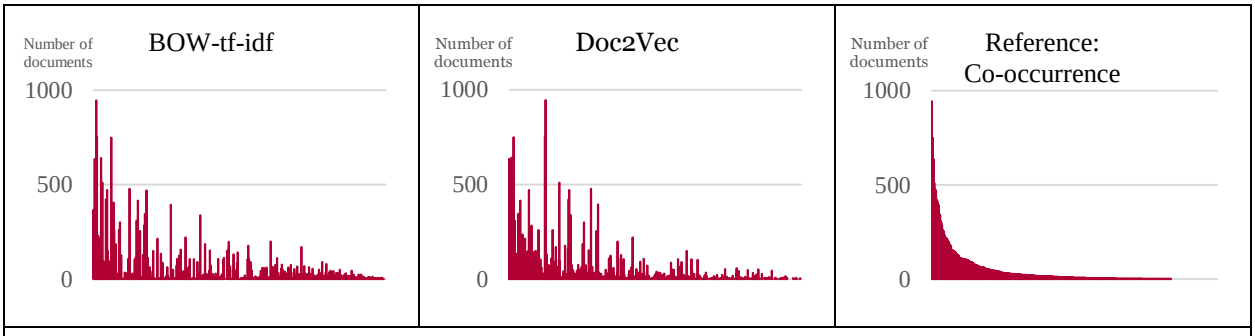

Figure 4. Face validity against co-occurrence of technologies in patents

At first glance, it becomes apparent that both algorithms tend to value technology pairs, which often co-occur in documents, with a high value. The accumulation of high bars on the left demonstrates this phenomenon. However, Figure 4 also illustrates that the algorithms come to different rankings and even give high weights to edges of technologies with a small number of co-occurrences. In this context, a low or non-existent number of co-occurrences does not mean that there is barely or even no relationship between the technologies. For instance, key-value databases and column-oriented databases are apparently related, but didn't appear together in the abovementioned patent collection. Although these technologies were not mentioned together in a single document, the method variants measured at least a moderate relatedness. To get a better impression of the resulting networks, we provide an exemplary network slice (Figure 5), using the BOW-tf-idf approach based on patent data as at 2017. The average edge degree determines the size of the nodes, and the edge thickness displays the relatedness between two technologies resulting from measuring the text similarity of the technology-specific corpora. By using cosine similarity, the edge weights have values between 0 and 1; thus, no further normalization is necessary. To ensure clarity, we filtered the edges based on their weights, so that they had to exceed a minimum value.

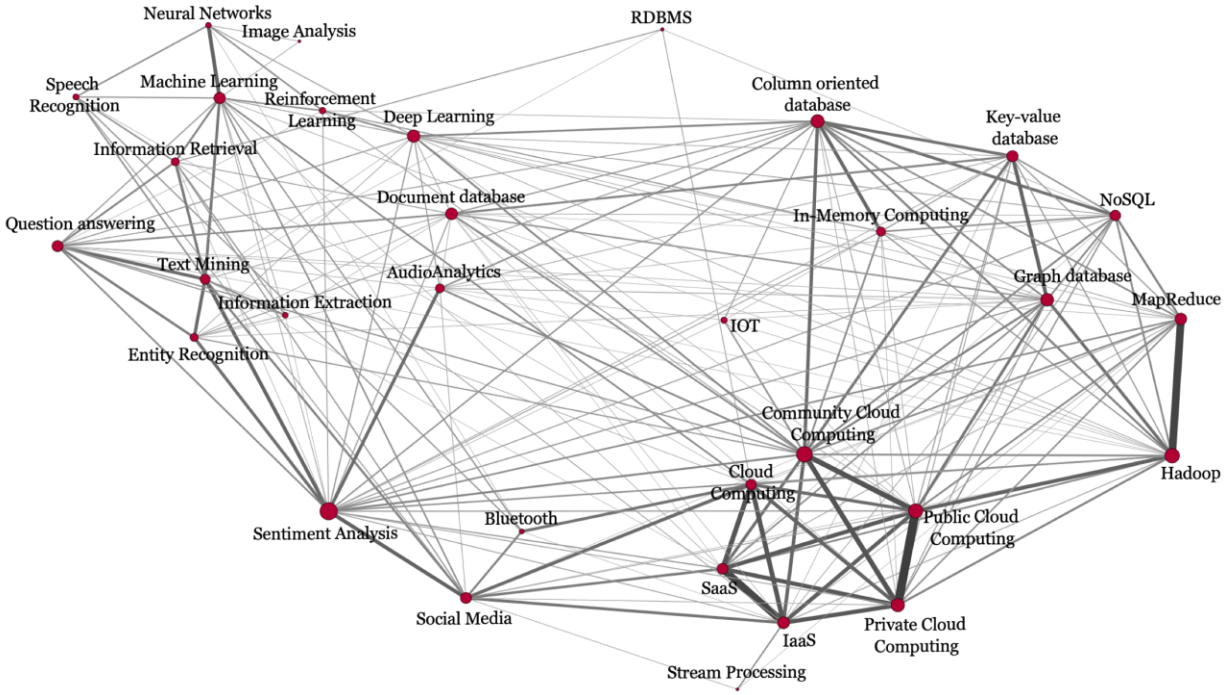

Figure 5. Exemplary network slice based on patent data as at 2017

Based on the network representation, it turns out that a basic structure is given through the highly weighted edges. For instance, the network representation exposes a group of technologies directly associated with cloud computing. Besides the thick edges, there are many thin edges. Notably, Bluetooth, which we initially intended for a false positive test, has among others relationships to cloud computing, social media, and speech recognition. The number of documents in which these technology pairs co-occur justified these relationships, since 749 patents mentioned both cloud computing and Bluetooth, 260 social media and Bluetooth, and 470 speech recognition and Bluetooth. For instance, the input device for speech recognition may be a Bluetooth-enabled device. The technologies Bluetooth and social media, as another example, are linked via patents in the context of smartphones.

Figure 6 illustrates the dynamic perspective of the inter-technology relationship network based on patent data from 2007 to 2017 by plotting the normalized average edge weights. The line graph shows that the average edge weight in our network increased over time, indicating the tighter integration of technologies

over the years. It is also striking that our method variants Doc2Vec and BOW-tf-idf both recognized the high interlocking of technologies even before the co-occurrence of technologies in patents increased.

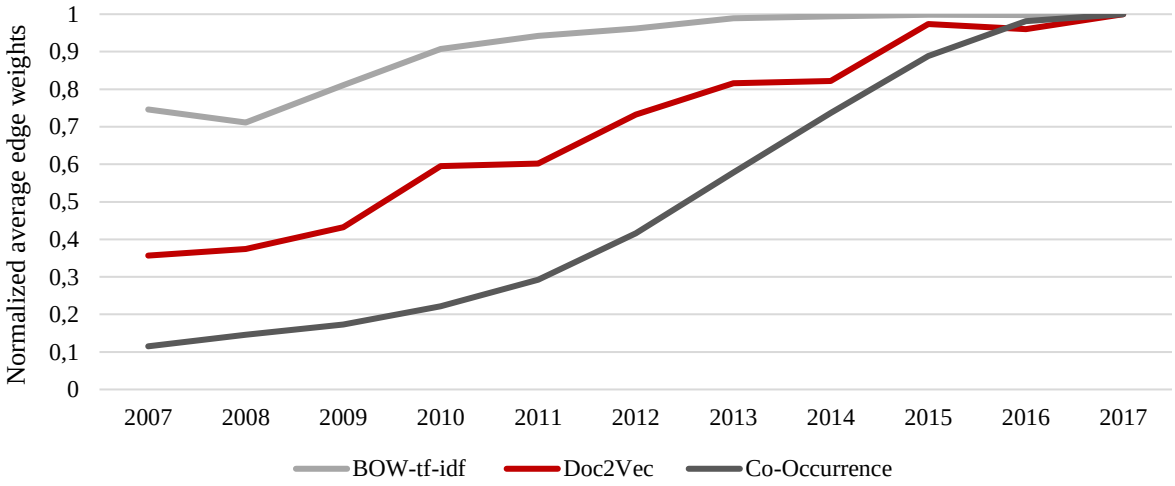

Figure 6. Exemplary network dynamics based on patent data

5.2 Feasibility and Ease-of-use

To discuss our method's feasibility, we implemented a prototype in *Python* as an instantiation of the proposed method (March and Smith 1995). This prototype also generated the paper's included figures. The possibility of swiftly assembling existing text mining techniques with the help of available, open-source libraries in environments such as *Python* fosters the proposed method's ease-of-use. We refer to *NLTK* (Bird et al. 2009) for natural language processing, *Scikit-learn* (Pedregosa et al. 2011) for BOW models, *genism* (Rehurek and Sojka 2010) for the Doc2Vec model, *NetworkX* (Hagberg et al. 2008) for the creation of our networks, and *Gephi* (Bastian et al. 2009) for visualization purposes. The availability of required data directly influences the method's applicability. While the United States Patent and Trademark Office provides patent full-texts in a processable way and free of charge, licensing issues concerning scientific full-texts complicate the method's application. Concerning required computing resources, experiences from the use of the prototype demonstrate that even a low-end workstation (6 CPU cores, 16GB of main memory) can run the processing pipelines of small networks within a reasonable amount of time. More extensive networks (i.e. more nodes or more data) raise the demands on computer resources. Besides developing

software to execute the proposed method, its application demands both expertise in text mining as well as in the technological domain under assessment. While using the simple BOW approach or its weighted extension (i.e. tf-idf) do not require any definition of parameters, the other variants require an understanding of parameter selection. The user of the proposed method requires in-depth knowledge of the technology landscape and the terminologies for defining a technologies set and appropriately configuring the asset allocator. The outlined imperative of method-based and domain-based knowledge restrict the user set.

5.3 Implications for the Completion Conditions

The application of the proposed method reveals several insights: First, it validates that the proposed method is suitable for measuring the relatedness between technologies in a real-world scenario. In this regard, the method supplies both obvious and unexpected relationships. Since each method variant highlights relationships differently, it remains open to derive the properties of the results of each method variant. Second, from the comparison of the results based on patents and scientific publications, we conclude that the inclusion of different asset types is worthwhile, in order to extend the explanatory power of an inter-technology relationship network. Third, the evaluation demonstrates the difficulty of getting a benchmark set against which method variants can be evaluated or trained. For one thing, the method variants' results only scarcely matched human judgments, although the interviewees approached their relatedness scores from different perspectives, resulting in limited agreement between the judges. For another thing, the method variants' results performed well against face validity.

Against the imposed completion conditions, we draw the following conclusion: The method meets the completion conditions CC1 (i.e. intended input-output transformation), CC3 (i.e. transparency and verifiability of the data processing steps), and CC4 (i.e. processing within a reasonable amount of time). We don't consider CC2 (i.e. each method chunk fulfills its requirements) to be fully assessable, owing to contradictions in the evaluation. This partial fulfillment motivates further research to compare the influences of different method chunks on the results. Notably, the overall assembly of the method works as a complement to decision-making rather than as a single point of truth.

6. Discussion and Conclusion

Ongoing advances in digital technologies – which are enabling new products, services, and business models – have fundamentally affected business and society through several waves of digitalization. Companies must decide which of the many emerging technologies are worth adopting or developing. Using an assembly-based process model for situational method engineering, and following the design science research paradigm, we have developed an analytical method to generate technology-related network data that retraces elapsed patterns of technological change. Thus, we consider the technological distances that characterize technologies' proximities and dependencies. We used established text mining techniques and drew from technology innovation research as justificatory knowledge. The method processes textual data from different information sources into an analyzable and readable inter-technology relationship network that is intended to be the input for further analyses. For instance, it may be the basis for the construction of domain-specific ontologies.

Although we have followed a rigorous research approach, this study has limitations. While we compared different method variants concerning text similarity measures, we only considered a selection of existing approaches and did not develop new algorithms or tailor existing ones. Further, we did not evaluate the influences of different preprocessing techniques, such as stemming instead of lemmatizing. We also simplified the problem by only allowing documents in English. Besides this, asset allocation follows a fairly simple approach so as to avoid false positive annotations, neglecting the potential information hidden in unallocated documents. While the proposed method quantifies relationships between technologies based on different data sources, it remains open to clarifying their contextual differences. There is always a limitation in the informative value of the data used, considering the use of data sources. In the case of patent data, different factors may result in biases and inconsistences across technology fields (Choi and Park 2009). These factors include for instance the strategic decision to keep an invention secret instead of patenting it (Kultti et al. 2007). Besides this, we derived historical developments without considering asset type-specific time lags (e.g., time lags induced by academic journals' peer-review processes). In addition to the

limitations in the method construction, the method's evaluation provided partial evidence and demands supplementary evaluation efforts. Evaluation efforts may include increasing the number of interviewees as well as having the method's results evaluated by the interviewees. It is also worth verifying whether the average of the interviewees' scores is an effective means. Further research may also evaluate the method's utility in real-world scenarios or usage cases and may include a comparison with other state-of-the-art approaches.

Further, we advocate research into appropriate methods to analyze our inter-technology relationship network. This endeavor is in line with related research (e.g., Choi et al. 2011; Yoon et al. 2011; Yoon and Kim 2011). In this context, the following research directions motivate us to use inter-technology relationship networks as an intermediate result for further analyses:

I. Regarding specific technologies, it is desirable to understand how an individual technology develops within a (specially) assembled network or in a static consideration of what relative importance it has (for a company) at a certain point in time. Besides gaining an overview of a technology landscape, our method allows practitioners to evaluate technologies individually in the technology context of their company, avoiding non-targeted assessments. For instance, it may support technology roadmapping and may provide insights into the technological development paths (Yan and Luo 2017).

II. Besides considering single technologies, companies may use a technologies set to evaluate the similarities between technology portfolios within and across companies. Thus, we are confident that methods based on inter-technology relationship networks support competitor analysis and decision-making in the context of mergers and acquisitions.

III. Nonetheless, our method and its results are not closed to theory-building. Among others, this could cover an inductive reasoning process that compares the evolution of previous innovations or, in more general technologies, in their lifecycles. Related to the network topology, we see potential in examining and comparing network structures both statically and dynamically.

In sum, we have contributed to the literature by introducing a complementary method for technology and innovation management as well as technology innovation research. The method allows one to create a network picture of an arbitrary technology landscape that is worth a million words. Thus, we have closed the addressed research gap by following a dynamic, multisource strategy to retrace technological distances between predefined technologies. The method's results open new possibilities for research and practice, advancing the discourse on the development of technology landscapes and occurring phenomena as well as the development of decision support systems such as technology forecasting tools. However, this network picture is not an end in itself, but an intermediate step in supporting data-driven decision-making. We strongly encourage future researchers to build on our method and to use its results as input for their analyses and methods.

References

Adomavicius, G., Bockstedt, J. C., Gupta, A., and Kauffman, R. J. 2007. "Technology roles and paths of influence in an ecosystem model of technology evolution," Information Technology and Management (8:2), pp. 185–202.

Adomavicius, G., Bockstedt, J. C., Gupta, A., and Kauffman, R. J. 2008. "Making Sense of Technology Trends in the Information Technology Landscape," *MIS Quarterly* (32:4), pp. 779–809.

Agarwal, R., and Dhar, V. 2014. "Editorial —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," *Information Systems Research* (25:3), pp. 443–448.

Aharonson, B. S., and Schilling, M. A. 2016. "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution," *Research Policy* (45:1), pp. 81–96.

Arthur, W. B. 2007. "The structure of invention," *Research Policy* (36:2), pp. 274–287.

Arthur, W. B. 2009. *The nature of technology*: *What it is and how it evolves*, New York, NY: Free Press.

Arts, S., Cassiman, B., and Gomez, J. C. 2017. "Text matching to measure patent similarity," *Strategic Management Journal* (67:1), pp. 62–84.

Baroni, M., Dinu, G., and Kruszewski, G. 2014. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Toutanova and H. Wu (eds.), Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 238–247.

Bastian, M., Heymann, S., and Jacomy, M. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks," in *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-2009)*, E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov and B. Tseng (eds.), Menlo Park, CA, USA: The AAAI Press, pp. 361–362.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. V. 2013. "Digital Business Strategy: Toward a Next Generation of Insights," *MIS Quarterly* (37:2), pp. 471–482.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*: *Analyzing Text with the Natural Language Toolkit*, Sebastopol: O'Reilly Media Inc.

Bollobás, B. 1998. *Modern Graph Theory*, New York, NY, USA: Springer New York.

Boyack, K. W., Klavans, R., and Börner, K. 2005. "Mapping the backbone of science," *Scientometrics* (64:3), pp. 351–374.

Bradford, R. B. 2008. "An empirical study of required dimensionality for large-scale latent semantic indexing applications," in *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, J. G. Shanahan (ed.), New York, NY, USA: ACM, pp. 153–162.

Breschi, S., Lissoni, F., and Malerba, F. 2003. "Knowledge-relatedness in firm technological diversification," *Research Policy* (32:1), pp. 69–87.

Brinkkemper, S. 1996. "Method engineering: Engineering of information systems development methods and tools," *Information and Software Technology* (38:4), pp. 275–280.

Bucher, T., and Winter, R. 2008. "Dissemination and Importance of the "Method" Artifact in the Context of Design Research for Information Systems," in *Proceedings of the Third International Conference on Design Science Research in Information Systems and Technology (DESRIST 2008)*, V. Vaishanvi and R. Baskerville (eds.), Atlanta, GA, USA: Georgia State University, pp. 39–59.

Buhl, H. U., Röglinger, M., Moser, F., and Heidemann, J. 2013. "Big Data: A Fashionable Topic with(out) Sustainable Relevance for Research and Practice?" *Business & Information Systems Engineering* (5:2), pp. 65–69.

Callon, M., Courtial, J. P., and Laville, F. 1991. "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry," *Scientometrics* (22:1), pp. 155–205.

Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. 2011. "Time-Varying Graphs and Dynamic Networks," in *Ad-hoc, mobile and wireless networks: 10th international conference, ADHOC-NOW 2011*, H. Frey, X. Li and S. Ruehrup (eds.), Berlin, Germany: Springer, pp. 346–359.
Chen, H.-M., Kazman, R., Haziyev, S., and Hrytsay, O. 2015. "Big Data System Development: An Embedded Case Study with a Global Outsourcing Firm," in *First International Workshop on Big Data Software Engineering - BIGDSE 2015*, Piscataway, NJ, USA: IEEE Press, pp. 44–50.
Chen, H.-M., Schutz, R., Kazman, R., and Matthes, F. 2016. "Amazon in the Air: Innovating with Big Data at Lufthansa," in *Proceedings of the 49th Annual Hawaii International Conference on System Sciences*, T. X. Bui and R. H. Sprague (eds.), Piscataway, NJ, USA: IEEE Press, pp. 5096–5105.
Choi, C., and Park, Y. 2009. "Monitoring the organic structure of technology based on the patent development paths," *Technological Forecasting and Social Change* (76:6), pp. 754–768.
Choi, S., Yoon, J., Kim, K., Lee, J. Y., and Kim, C.-H. 2011. "SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells," *Scientometrics* (88:3), pp. 863–883.
Clarivate Analytics 2018. *Web of Science*. http://apps.webofknowledge.com/. Accessed 12 April 2018.
Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. 2006. "Forecasting emerging technologies: Use of bibliometrics and patent analysis," *Technological Forecasting and Social Change* (73:8), pp. 981–1012.
Debortoli, S., Junglas, I., Müller, O., and vom Brocke, J. 2016. "Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems* (39:1), pp. 110–135.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. "Indexing by latent semantic analysis," *Journal of the American Society for Information Science* (41:6), pp. 391–407.
Ding, Y., Chowdhury, G. G., and Foo, S. 2001. "Bibliometric cartography of information retrieval research by using co-word analysis," *Information Processing & Management* (37:6), pp. 817–842.
Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. 1988. "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, J. J. O'Hare (ed.), New York, NY, USA: ACM, pp. 281–285.
Egghe, L., and Rousseau, R. 2002. "Co-citation, bibliographic coupling and a characterization of lattice citation networks," *Scientometrics* (55:3), pp. 349–361.
Engelsman, E. C., and Van Raan, A.F.J. 1994. "A patent-based cartography of technology," *Research Policy* (23:1), pp. 1–26.
Fan, W., Wallace, L., Rich, S., and Zhang, Z. 2006. "Tapping the power of text mining," *Communications of the ACM* (49:9), pp. 76–82.
Feldman, Ronen, Dagan, and Ido 1995. "Knowledge Discovery in Textual Databases (KDT)," *KDD* (95), pp. 112–117.
Ferreira, A. 2004. "Building a reference combinatorial model for MANETs," *IEEE Network* (18:5), pp. 24–29.
Fleming, L., and Sorenson, O. 2001. "Technology as a complex adaptive system: Evidence from patent data," *Research Policy* (30:7), pp. 1019–1039.
Furukawa, T., Mori, K., Arino, K., Hayashi, K., and Shirakawa, N. 2015. "Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions," *Technological Forecasting and Social Change* (91), pp. 280–294.
Gandomi, A., and Haider, M. 2015. "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management* (35:2), pp. 137–144.
Gartner (Gartner) 2017. *Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017*. https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/. Accessed 5 March 2018.

Gregor, S., and Hevner, A. R. 2013. "Positioning and presenting design science research for maximum impact," *MIS Quarterly* (37:2), pp. 337–355.
Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., and Smyth, P. 2012. "TopicNets," *ACM Transactions on Intelligent Systems and Technology* (3:2), pp. 1–26.
Gupta, V. K., and Pangannaya, N. B. 2000. "Carbon nanotubes: Bibliometric analysis of patents," *World Patent Information* (22:3), pp. 185–189.
Hagberg, A., Swart, P., and Chult, D. S. 2008. "Exploring Network Structure, Dynamics, and Function using NetworkX," in *7th Python in Science Conference (SciPy2008)*, G. Varoquaux, T. Vaught and J. Millman (eds.), Pasadena, CA, USA. August 21, pp. 11–15.
Hao, J., Yan, Y., Gong, L., Wang, G., and Lin, J. 2014. "Knowledge map-based method for domain knowledge browsing," *Decision Support Systems* (61), pp. 106–114.
Harris, Z. S. 1954. "Distributional Structure," *WORD* (10:2-3), pp. 146–162.
Hearst, M. A. 1999. "Untangling text data mining," in *Proceedings of the conference, 37th Annual Meeting of the Association for Computational Linguistics*, M. A. Hearst (ed.), San Francisco, CA, USA: Morgan Kaufmann, pp. 3–10.
Henderson-Sellers, B., and Ralyté, J. 2010. "Situational Method Engineering: State-of-the-Art Review," *Journal of Universal Computer Science* (16:3), pp. 424–478.
Hevner, A. R. 2007. "A Three Cycle View of Design Science Research," *Scandinavian Journal of Information Systems* (19:2), pp. 87–92.
Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.
Holme, P., and Saramäki, J. 2012. "Temporal Networks," *Physics Reports* (519:3), pp. 97–125.
Hu, H., Wen, Y., Chua, T.-S., and Li, X. 2014. "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access* (2), pp. 652–687.
Hull, D. A. 1996. "Stemming algorithms: A case study for detailed evaluation," *Journal of the American Society for Information Science* (47:1), pp. 70–84.
Iivari, J. 2007. "A Paradigmatic Analysis of Information Systems As a Design Science," *Scandinavian Journal of Information Systems* (19:2), pp. 39–64.
Joo, S. H., and Kim, Y. 2010. "Measuring relatedness between technological fields," *Scientometrics* (83:2), pp. 435–454.
Kambatla, K., Kollias, G., Kumar, V., and Grama, A. 2014. "Trends in big data analytics," *Journal of Parallel and Distributed Computing* (74:7), pp. 2561–2573.
Kay, L., Newman, N., Youtie, J., Porter, A. L., and Rafols, I. 2014. "Patent overlay mapping: Visualizing technological distance," *Journal of the Association for Information Science and Technology* (65:12), pp. 2432–2443.
Kendall, M. G. 1938. "A New Measure of Rank Correlation," *Biometrika* (30:1/2), pp. 81–93.
Kendall, M. G., and Smith, B. B. 1939. "The Problem of m Rankings," *The Annals of Mathematical Statistics* (10:3), pp. 275–287.
Klavans, R., and Boyack, K. W. 2009. "Toward a consensus map of science," *Journal of the American Society for Information Science and Technology* (60:3), pp. 455–476.
Kultti, K., Takalo, T., and Toikka, J. 2007. "Secrecy versus patenting," *The RAND Journal of Economics* (38:1), pp. 22–42.
Landauer, T. K., and Dumais, S. T. 1997. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review* (104:2), pp. 211–240.
Landauer, T. K., Foltz, P. W., and Laham, D. 1998. "An introduction to latent semantic analysis," *Discourse Processes* (25:2-3), pp. 259–284.
Lapata, M. 2006. "Automatic Evaluation of Information Ordering: Kendall's Tau," *Computational Linguistics* (32:4), pp. 471–484.
Le, Q., and Mikolov, T. 2014. "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on Machine Learning, PMLR* (32:2), pp. 1188–1196.

Lee, S., Lee, S., Seol, H., and Park, Y. 2008. "Using patent information for designing new product and technology: Keyword based technology roadmapping," *R&D Management* (38:2), pp. 169–188.
Lee, S., Yoon, B., Lee, C., and Park, J. 2009a. "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping," *Technological Forecasting and Social Change* (76:6), pp. 769–786.
Lee, S., Yoon, B., and Park, Y. 2009b. "An approach to discovering new technology opportunities: Keyword-based patent map approach," *Technovation* (29:6-7), pp. 481–497.
Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Mädche, A., Urbach, N., and Ahlemann, F. 2017. "Digitalization: Opportunity and Challenge for the Business and Information Systems Engineering Community," *Business & Information Systems Engineering* (59:4), pp. 301–308.
Leydesdorff, L. 2008. "Patent classifications as indicators of intellectual organization," *Journal of the American Society for Information Science and Technology* (59:10), pp. 1582–1597.
Leydesdorff, L., Kushnir, D., and Rafols, I. 2014. "Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC)," *Scientometrics* (98:3), pp. 1583–1599.
Leydesdorff, L., and Rafols, I. 2009. "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology* (60:2), pp. 348–362.
Leydesdorff, L., and Vaughan, L. 2006. "Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment," *Journal of the American Society for Information Science and Technology* (57:12), pp. 1616–1628.
Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. 2006. "Sentence similarity based on semantic nets and corpus statistics," *IEEE Transactions on Knowledge and Data Engineering* (18:8), pp. 1138–1150.
Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003. "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics* (19:10), pp. 1275–1283.
Lowe, W. 2001. "Towards a Theory of Semantic Space," in *Proceedings of the twenty-third annual conference of the Cognitive Science Society*, J. D. Moore and K. Stenning (eds.), Mahwah, NJ, USA: Lawrence Erlbaum.
Madani, F. 2015. "'Technology Mining' bibliometrics analysis: Applying network analysis and cluster analysis," *Scientometrics* (105:1), pp. 323–335.
Madani, F., and Weber, C. 2016. "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis," *World Patent Information* (46), pp. 32–48.
Maedche, A., and Staab, S. 2002. "Measuring Similarity between Ontologies," in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, A. Gómez-Pérez and V. R. Benjamins (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, 251-163.
Manning, C. D., Raghavan, P., and Schütze, H. 2009. *Introduction to information retrieval*, Cambridge: Cambridge Univ. Press.
March, S. T., and Smith, G. F. 1995. "Design and natural science research on information technology," *Decision Support Systems* (15:4), pp. 251–266.
March, S. T., and Storey, V. C. 2008. "Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research," *MIS Quarterly* (32:4), pp. 725–730.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. 2011. "Quantitative analysis of culture using millions of digitized books," *Science* (331:6014), pp. 176–182.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. "Efficient Estimation of Word Representations in Vector Space," in *International Conference on Learning Representations 2013*, Scottsdale, Arizona, USA. May 2-4, pp. 1–12.
Miller, G. A. 1995. "WordNet: A lexical database for English," *Communications of the ACM* (38:11), pp. 39–41.

Myers, M. D., and Newman, M. 2007. "The qualitative interview in IS research: Examining the craft," *Information and Organization* (17:1), pp. 2–26.
Nakamura, H., Suzuki, S., Sakata, I., and Kajikawa, Y. 2015. "Knowledge combination modeling: The measurement of knowledge similarity between different technological domains," *Technological Forecasting and Social Change* (94), pp. 187–201.
Narin, F., Olivastro, D., and Stevens, K. A. 1994. "Bibliometrics/Theory, Practice and Problems," *Evaluation review* (18:1), pp. 65–76.
Navigli, R., Velardi, P., and Gangemi, A. 2003. "Ontology learning and its application to automated terminology translation," *IEEE Intelligent Systems* (18:1), pp. 22–31.
Newman, M. E. J. 2003. "The Structure and Function of Complex Networks," *SIAM Review* (45:2), pp. 167–256.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* (12:Oct), pp. 2825–2830.
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 44–77.
Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. 2009. "Semantic similarity in biomedical ontologies," *PLoS computational biology* (5:7), e1000443.
Porter, A. L., and Cunningham, S. W. 2005. *Tech mining*: *Exploiting new technologies for competitive advantage*, Hoboken, New Jersey, USA: Wiley-Interscience.
Porter, A. L., and Detampel, M. J. 1995. "Technology Opportunities Analysis," *Technological Forecasting and Social Change* (49:3), pp. 237–255.
Ralyté, J., Deneckère, R., and Rolland, C. 2003. "Towards a Generic Model for Situational Method Engineering," in *Advanced Information Systems Engineering: 15th International Conference, CAiSE 2003*, J. Eder and M. Missikoff (eds.), Berlin and Heidelberg, Germany: Springer, pp. 95–110.
Rehurek, R., and Sojka, P. 2010. "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* Valletta, Malta. May 22, Valletta, Malta: ELRA, pp. 45–50.
Salton, G., and Buckley, C. 1988. "Term-weighting approaches in automatic text retrieval," *Information Processing & Management* (24:5), pp. 513–523.
Salton, G., Fox, E. A., and Wu, H. 1983. "Extended Boolean information retrieval," *Communications of the ACM* (26:11), pp. 1022–1036.
Salton, G., Wong, A., and Yang, C. S. 1975. "A vector space model for automatic indexing," *Communications of the ACM* (18:11), pp. 613–620.
Sánchez, D., Batet, M., Isern, D., and Valls, A. 2012. "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications* (39:9), pp. 7718–7728.
Schoen, A., Villard, L., Laurens, P., Cointet, J.-P., Heimeriks, G., and Alkemade, F. 2012. "The Network Structure of Technological Developments; Technological Distance as a Walk on the Technology Map," in *Proceedings of STI 2012 Montréal: 17th International Conference on Science and Technology Indicators*, É. Archambault and O. s. e. Des technologies (eds.), Montreal: Science-Metrix, pp. 733–742.
Schoenmakers, W., and Duysters, G. 2010. "The technological origins of radical inventions," *Research Policy* (39:8), pp. 1051–1059.
Small, H. 1973. "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science and Technology* (24:4), pp. 265–269.
Small, H., and Griffith, B. C. 1974. "The Structure of Scientific Literatures I: Identifying and Graphing Specialties," *Science Studies* (4:1), pp. 17–40.

Swanson, D. R. 1987. “Two medical literatures that are logically but not bibliographically connected,” *Journal of the American Society for Information Science* (38:4), pp. 228–233.
The United States Patent and Trademark Office (USPTO) 2018. *United States Patent and Trademark Office*. https://bulkdata.uspto.gov/. Accessed 8 March 2018.
Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. 2007. “Text mining techniques for patent analysis,” *Information Processing & Management* (43:5), pp. 1216–1247.
Turney, P. D., and Pantel, P. 2010. “From Frequency to Meaning: Vector Space Models of Semantics,” *Journal of Artificial Intelligence Research* (37), pp. 141–188.
Yan, B., and Luo, J. 2017. “Measuring technological distance for patent mapping,” *Journal of the Association for Information Science and Technology* (68:2), pp. 423–437.
Yoon, B., and Park, Y. 2004. “A text-mining-based patent network: Analytical tool for high-technology trend,” *The Journal of High Technology Management Research* (15:1), pp. 37–50.
Yoon, J., Choi, S., and Kim, K. 2011. “Invention property-function network analysis of patents: A case of silicon-based thin film solar cells,” *Scientometrics* (86:3), pp. 687–703.
Yoon, J., and Kim, K. 2011. “Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks,” *Scientometrics* (88:1), pp. 213–228.
Yoon, J., and Kim, K. 2012. “Detecting signals of new technological opportunities using semantic patent analysis and outlier detection,” *Scientometrics* (90:2), pp. 445–461.
Yoon, J., Park, H., and Kim, K. 2013. “Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis,” *Scientometrics* (94:1), pp. 313–331.
Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., and Venkataraman, S. 2016. “Apache Spark: a unified engine for big data processing,” *Communications of the ACM* (59:11), pp. 56–65.
Zhang, Y., Chen, H., and Zhu, D. 2016. “Semi-automatic Technology Roadmapping Composing Method für Multiple Science, Technology, and Innovation Data Incorporation,” in *Anticipating Future Innovation Pathways Through Large Data Analysis*, T. U. Daim, D. Chiavetta, A. L. Porter and O. Saritas (eds.), Cham: Springer International Publishing, pp. 211–227.
Zhu, G., and Iglesias, C. A. 2017. “Computing Semantic Similarity of Concepts in Knowledge Graphs,” *IEEE Transactions on Knowledge and Data Engineering* (29:1), pp. 72–85.