Technological Forecasting & Social Change xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Extending the knowledge base of foresight: The contribution of text mining

Victoria Kayser [a,b,*], Knut Blind [b,c]

[a] Fraunhofer Institute for Systems and Innovation Research, Breslauer Strasse 48, 76131 Karlsruhe, Germany
[b] Technische Universität Berlin, Chair of Innovation Economics, Marchstraße 23, D-10587 Berlin, Germany
[c] Fraunhofer Institute for Open Communication Systems and Rotterdam School of Management, Chair of Standardisation, Erasmus University, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 9 February 2016
Received in revised form 23 September 2016
Accepted 18 October 2016
Available online xxxx

Keywords:
Foresight
Text mining
Data analysis
Roadmapping
Scenario development
Big data

ABSTRACT

While the volume of data from heterogeneous sources grows considerably, foresight and its methods rarely benefit from such available data. This work concentrates on textual data and considers its use in foresight to address new research questions and integrate other stakeholders. This textual data can be accessed and systematically examined through text mining which structures and aggregates data in a largely automated manner. By exploiting new data sources (e.g. Twitter, web mining), more actors and views are integrated, and more emphasis is laid on the analysis of social changes. The objective of this article is to explore the potential of text mining for foresight by considering different data sources, text mining approaches, and foresight methods. After clarifying the potential of combining text mining and foresight, examples are outlined for roadmapping and scenario development. As the results show, text mining facilitates the detection and examination of emerging topics and technologies by extending the knowledge base of foresight. Hence, new foresight applications can be designed. In particular, text mining provides a solid base for reflecting on possible futures.

1. Introduction

The volume of data from heterogeneous sources has considerably grown (Ortner et al., 2014) and the scientific output is constantly increasing (see, e.g., Bornmann and Mutz, 2014). Identifying the relevant data from the huge quantity of available information is challenging and more effort is needed in monitoring thematic fields. Furthermore, (textual) data from websites or social media could be analyzed to address new aspects and research questions (see e.g., Boyd and Crawford, 2012; Kitchin, 2014). For example, the user-generated content on the web may be interesting in the context of foresight, for examining social perspectives and the user's perception of current developments. However, the web data is at present rarely considered for a systematic examination (Yoon, 2012; Cachia et al., 2007; Glassey, 2012). However, new indicators could be established to extend the scheme of present indicators with their focus on science and technology through patent and publication analysis (Kostoff, 2012; Abbas et al., 2014). Today, many relevant information sources are left out, though this data could be used to perceive ongoing changes and make more precise statements about possible future developments and emerging technologies. Therefore, new methods and tools for processing and integrating data for foresight are required.

This article focuses on textual data (e.g., reports, blog entries, or Twitter data) that can be accessed and systematically examined through text mining (Berry, 2004; Feldman and Sanger, 2008). By integrating text mining into foresight, other data sources are accessible to be considered in a comprehensive way, especially unstructured and large datasets. Therefore, the objective of this article is to identify and elaborate the potential of text mining for foresight and its methods. One aspect is to consider data sources such as Twitter or websites and new techniques for data retrieval such as web mining. This article examines the extent to which foresight and its methods can be improved through the results of text mining. Therefore, applications are presented wherein text mining is combined with foresight methods such as roadmapping (e.g., Möhrle et al., 2013; Phaal et al., 2010) or scenario development (e.g., Reibnitz, 1991; van der Heijden, 2005; O'Brien and Meadows, 2013). By additional data and stakeholder views, it is expected to enhance the detection and examination of emerging themes and technologies and to provide a solid base for decision making.

This article begins with the fundamentals of foresight and the basic principles of text mining in Section 2. Section 3 addresses the use of text mining for foresight. Different data sources are described, the state-of-the-art concerning existing implementations is summarized, and further applications are outlined. Finally, the results are discussed in the framework of foresight and a conclusion is drawn in Section 4.

* Corresponding author.
E-mail address: mail@vkayser.de (V. Kayser).

http://dx.doi.org/10.1016/j.techfore.2016.10.017
0040-1625/© 2016 Elsevier Inc. All rights reserved.

2. Foresight and text mining

This section introduces the two main components of this article—foresight and text mining—and gives an overview of the recent debate.

2.1. Foresight

Foresight is a systematic process of looking into the long-term future of science, technology, and innovation (e.g., Martin, 1995; Cuhls, 2003). One definition of foresight is *"opening to the future with every means at our disposal, developing views of future options, and then choosing between them"* (Slaughter, 1995). Foresight thereby considers possible and plausible futures, since there is not only one future. In principal, the future cannot be predicted but is shaped by the decisions and actions of today. Foresight helps to assess the consequences and implications of present actions, early warnings, thinking about desirable futures, and implications of possible future events. Therefore, it is an action-oriented decision support that brings together the relevant stakeholders for an open discourse about possible futures. Foresight builds on a set of different methods (e.g., Popper and Butter, 2008) such as roadmapping (Barker and Smith, 1995; Möhrle et al., 2013) or scenario development (Reibnitz, 1991; van der Heijden, 2005). The set of methods to be selected for application depends on the scope and focus of the foresight exercise and has to be decided from case to case. Foresight, futures studies, and future technology analysis are not further distinguished in the course of this article due to their commonalities.

Foresight is modular and a sequence of steps. Depending on the objectives and application level, different methods and tasks are combined. Building on previous studies (see, e.g., Martin, 1995; Horton, 1999; Voros, 2003; Da Costa et al., 2008; de Miranda Santo et al., 2006), foresight exercises might be categorized into three phases as illustrated in Fig. 2-1—input, process, and output.

2.1.1. Input

Besides some overall objectives, a process scope is defined, a time horizon is set, and information about recent trends and developments is gathered with regard to the considered field. At the beginning of almost every process, the state-of-the-art has to be summarized. Therefore, this first step relates to collecting and summarizing the available information to get an overview of the present situation (Horton, 1999).

2.1.2. Process

Future technology analysis might be seen as a process of knowledge creation (Eerola and Miles, 2011). Specific foresight methods are applied according to the scope and process objectives. By this, important information about the future and possible future developments is gathered and knowledge is generated, which later serves as decision support.

2.1.3. Output

The results are assessed, priorities are set, and strategies are formulated (de Miranda Santo et al., 2006). This phase is about taking action (Horton, 1999). Diverse interests or expectations related to foresight outcomes exist. One intention of foresight is to support the design of future-oriented strategies. Furthermore, political or governmental actors expect recommendations for planning or setting priorities for research programs (Havas et al., 2010; Könnölä et al., 2011; de Smedt, 2013).

2.2. Text mining

Due to the increasing volume of data from heterogeneous sources, the effort needed to study thematic fields and developments and to read the published studies and literature has increased (Ortner et al., 2014). In times of big data, techniques are required that can automatically process this data and be adapted to varying requirements and data sets. Text mining is suitable for this purpose as it offers methods to access and analyze these textual data sources (Weiss, 2010; Feldman and Sanger, 2008) as introduced in the following.

2.2.1. The process of text mining

Text mining processes unstructured textual data into a structured format for further analysis. Text mining can be summarized in three steps, as indicated in Fig. 2-2. First, a data source is selected. Then this data is preprocessed (Step 2) and analyzed (Step 3). Finally, the results are interpreted.

2.2.1.1. Text selection. The selected data source should be able to answer the research question (e.g. social media, patents, standards). For searching data, at least some principal knowledge of the subject or technology under consideration is necessary. While some data is retrieved from databases (e.g. patent, standards, scientific publications), other data has to be manually gathered (e.g. reports).

2.2.1.2. Text preprocessing. Text has to be structured and transformed into a machine- readable format for further processing. Therefore, the text is divided into its individual elements as words (tokenization). To extract the relevant terms, two different approaches are distinguished—working with stopwords and working with grammatical instances. When using the grammatical instances, speech tags are assigned to each word, such as verb, article, or noun. From this, relevant phrases or chains of words are extracted. Alternatively, stopwords are used to remove irrelevant terms and function words (articles, conjunctions, pronouns, etc.). Further techniques like stemming (which reduces each word to its basic form) or

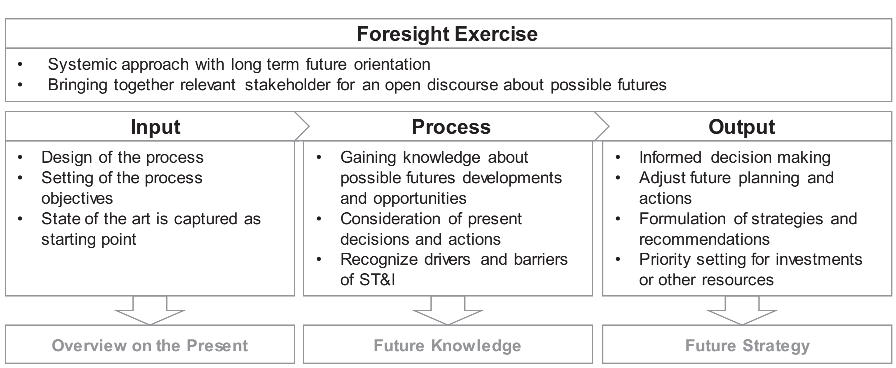

Fig. 2-1. The process of foresight.

Please cite this article as: Kayser, V., Blind, K., Extending the knowledge base of foresight: The contribution of text mining, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.017

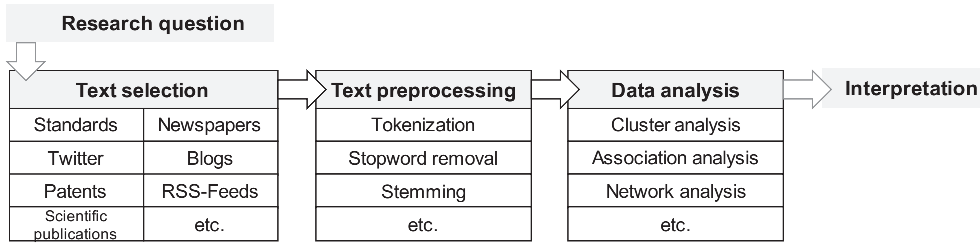

Fig. 2-2. Text mining process.

lemmatization (which reduces word to root form based on a dictionary) are then applied. Finally, the frequency of the terms per document is stored for further analysis.

2.2.1.3. Data analysis. For data analysis, particular methods from statistics and data mining such as classification and clustering are applied (Han et al., 2012; Manning et al., 2009) and a wide range of software solutions exist to assist this process (e.g. *R*, *Weka*, *SPSS Modeler*, *Leximancer*). However, for a clear documentation of the research process, an individualized analysis software is more flexible and can be adapted to specific requirements and data sources. The single process steps can thus be traced (see Kayser and Shala, 2016 for further discussion).

2.2.1.4. Interpretation. Finally, interpreting the results is crucial because each dataset is subject to biases and limitations (e.g. completeness, representability). However, data is not self-explanatory and cannot speak for itself. Hence, methodological and domain knowledge is required and further skills and expertise are necessary (see, e.g., Kitchin, 2014). In addition, the results have to be embedded in the context of the foresight process for which they are intended. Text mining is an iterative process in which the results raise further questions that require additional searches, data or follow-up research (e.g. interviews, workshops) to validate the results.

2.2.2. Constraints of text mining

There are several constraints when implementing text mining. First, text mining analyses textual data primarily based on word frequencies and word relations. This ignores images and figures as well as irony, sarcasm, and everything *between the lines*. This implies that text mining is not suitable for some research questions or should not be used alone (see also Grimmer and Stewart, 2013). Algorithms handle data in a different manner compared to reading and deliver a surface analysis. Therefore, some research questions still require qualitative and manual analysis, such as the analysis of Delphi statements. As Kayser and Shala (2016) show, text mining delivers a rapid overview of the field, but a more detailed level (e.g. statistics, facts, numbers) needs manual work. The result of text mining is, most often, a summary that needs to be adjusted in an iterative process. Second, each data source has its own strength and limitations. For example, measuring R&D activities requires them to be patented or published (e.g., Cunningham et al., 2006; Bonino et al., 2010). This has to be taken into account when using text mining, especially when used together with textual data wherein the strength and weaknesses of a quantitative examination are less explored, e.g. for social media data or websites. Third, technical constraints for the application of text mining relate to designing the technical process (combination of appropriate algorithms). To process another type of textual data, adaption of the interfaces is necessary because the structure of each type of dataset varies. Therefore, a personal adaptable framework meets the requirements best (e.g. adaptable to further data sources, modular design, extendable). Thus, for this work, the authors decided to build a customized text mining-framework based on Python and SQL.

3. Using text mining for foresight

This section describes different textual data sources that are or that might be analyzed with text mining (Section 3.1) and outlines different applications that link text mining and foresight at a methodological level (Section 3.2).

3.1. Text as data

This section introduces text sources that are or could be used in foresight and can principally be analyzed with text mining.

3.1.1. Scientific publications, patents, and standards

Patents, scientific publications, and standards are used as indicators of technical change. Data is extracted from specialized databases that are quality-assured and updated on a regular basis (e.g. Web of Science). These sources are frequently used in foresight to examine science and technology developments.

In the field of publication analysis, text mining is used and applied on data fields like title, abstract, keywords, and even full texts (Cunningham et al., 2006; Kostoff, 2012). Different approaches for term extraction are applied, such as stopword removal-based approaches (Glenisson et al., 2005; Delen and Crossland, 2008) or approaches based on the grammatical instance such as *PoS*-extraction (van Eck et al., 2010). With regard to methodology, classification or cluster analysis (Glenisson et al., 2005; Delen and Crossland, 2008), topic modeling (e.g., Yau et al., 2014), and network and mapping approaches (e.g., van Eck and Waltman, 2011) are frequently applied. In recent years, there has been a growing interest in applying text mining for patent analysis to access unstructured text fields like abstracts, claims, or descriptions (Masiakowski and Wang, 2013; Tseng et al., 2007; Abbas et al., 2014). Text mining has a major advantage compared to manual approaches; it aggregates large quantities of patents, generates further information in the form of statistics or maps, and supports decision-making (see, e.g., Wang et al., 2010). The current applications consider different areas such as patent infringement detection (Lee et al., 2013; Park et al., 2012), monitoring the R&D landscape (e.g., Yoon et al., 2013), technology transfer (Park et al., 2013b), general technology planning (Park et al., 2013a; Choi et al., 2012; Wang et al., 2010), and automatic patent classification (Bonino et al., 2010; Cong and Loh, 2010).

In brief, many research activities are conducted and patents, standards, and scientific publications are analyzed with text mining in different applications. This article does not aim to further develop these approaches but to link these forms of data analysis to foresight applications and analyze other data apart from patents and publications, as illustrated in Section 3.2.3.

3.1.2. News articles

News articles keep the society informed and contribute to public opinion-making (e.g., Burkart, 2002). Their analysis may focus on public concerns, beliefs, and reservations. As for patents and scientific publications, the texts are edited and clearly written. For example, Yoon (2012) examines web news for weak signals in the field of solar cells and

Please cite this article as: Kayser, V., Blind, K., Extending the knowledge base of foresight: The contribution of text mining, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.017

appraises that web news are a more refined and reliable source than blogs or web pages for his case. As far as the knowledge of the authors goes, the systematic and automated analysis of news articles with text mining is rarely applied in foresight but principally enables the addressing of society-related issues.

3.1.3. Social media data

Social media are web-based applications such as YouTube, Facebook, or Twitter (e.g., Kaplan and Haenlein, 2010; Kietzmann et al., 2011). These platforms are relevant for data gathering and also for participatory aspects (see e.g., Kayser and Bierwisch, 2016). Principally, user-generated content such as blogs or Twitter may contribute with insights on societal discourses. For example, Cachia et al. (2007) examine the potential of online social networks for foresight and trend recognition. They conclude their principal discussion of applications and platforms with the observation that social networks indicate changes and trends in sentiment and social behavior and moreover foster creativity and collective intelligence. Pang (2010) describes an approach to scan Web 2.0 contents produced by *futurists* (people with expertise in foresight or futures studies) on different web channels. *Social scanning* may deliver a very precise summary of what is discussed and what attracts *futurists* attention, but Pang's work misses a practical realization. Amanatidou et al. (2012) describe how they analyzed Twitter and other publicly available web sources using text mining in the context of weak signal identification and horizon scanning. Albert et al. (2015) analyze blogs with reference to technology-maturity models while Glassey (2012) examines folksonomies (the tagging of web content with meta information) for their potential in early trend detection. In brief, the first ideas and applications use different social media platforms for collecting information and studying user interaction, especially for the purpose of trend detection and the assessment of technology maturity. For text mining tasks, there is a wide range of applications to access and analyze this data.

Due to the huge range of topics covered in social media data, the option to retrieve this data in a structure manner using APIs, worldwide many-to-many communication, and real-time access to public debates provide many options for quantitative analysis. While a number of works exist, there are still plenty of options to combine social media and foresight (e.g. for use during scenario workshops, exploration of social discourse on a topic, etc.).

3.1.4. Websites

A lot of information is publicly available on websites. This semi-structured data could be analyzed through text mining. At the moment, a number of applications use web data as innovation indicators. For example, company websites are accessed for reports on innovations (Gök et al., 2015). Youtie et al. (2012) examine websites of small and medium-sized enterprises in the field of nanotechnology with regard to technology transition from discovery to commercialization. Approaches for the general retrieval of topic-related websites are not used at present.

3.1.5. (Scientific) Reports and foresight studies

Foresight studies are a commonly used information source in foresight exercises. They are manually screened for future statements. To automate this time-consuming task at least partially, text mining would be of great value, as attempted by Amanatidou et al. (2012) for example. However, as they noticed, due to the length of the reports, the most frequent terms are not the most interesting. Therefore, cleaning and filtering are necessary for weak signal detection. Kayser and Shala (2016) analyze reports using text mining to summarize the topic and deliver an initial starting point for the following scenario development.

3.2. Text mining for foresight methods

Text Mining can be seen as methodical building block for extending and improving foresight methods and addressing new research questions. As the overview of possible data sources in Section 3.1 shows, most existing applications using text mining focus on scientific and technological data (e.g. patents, publications). However, topics other than science and technology can be addressed based on textual data. Especially for mapping societal aspects (technology diffusion, user acceptance, etc.), a wide range of options exists that should be explored in future research. Furthermore, the automatic comparison of textual data sets has not yet been addressed in the context of foresight and web data is rarely used or combined with foresight methods. Building on these observations, the following section illustrates three examples of combining foresight and text mining in a new manner.

3.2.1. Technology roadmapping

The following example combines roadmapping and text mining into a two-layered process model that integrates external and internal views. This model is designed to balance internal views by providing an objective overview of external trends and developments.

Roadmaps are an instrument for strategic future planning (Barker and Smith, 1995; Möhrle et al., 2013; Phaal et al., 2010). Some preliminary work related to roadmapping and text mining exists (e.g., Choi et al., 2013; Yoon et al., 2008; Lee et al., 2008; Huang et al., 2014). So far, different text mining techniques (e.g. topic modeling, text summarization, clustering) have been applied on different textual data sources (e.g. patents, product manuals). Together, these studies indicate that text mining is merely done initially to get a thematic overview. Text mining and roadmapping are not conducted in parallel; the core roadmapping is done exclusively by experts.

In the framework of Kayser et al. (2014), roadmapping is used for the internal strategy development while text mining is applied for the analysis of external data and changes. Supportive text mining tasks are assigned for each process step of roadmapping (see Fig. 3-1). Each process step is supported by specific analysis. Continuous feedback loops between the two layers—text mining and roadmapping—enrich the strategy process and serve as an objective base for balancing the internal views of the experts. First, data sources need to be identified based on the project scope and topic. In the illustrative case described in Kayser et al. (2014), scientific publication data on *cloud computing* is used. Based on the retrieved data, the analysis using text mining supports the initial exploration and identification of relevant terms (see Step 1) that can be used to refine the defined scope, e.g. by recognizing missing aspects. In the second step, the data is analyzed for trends to support the chronological ordering of the roadmapping objects. In the third step, network and association analyses indicate links between the objects of the roadmap to draw dependencies on the roadmap. The results of the previous steps serve as orientation for the final validation of the roadmap in Step 4.

The contribution of this approach is the structured integration of external data in internal roadmapping processes. Text mining adds a second (external) perspective to the (internal) strategic considerations of roadmapping. This provides better information to the participants of roadmapping when thinking about the future. Fig. 3-1 illustrates one way of combining roadmapping and text mining. To provide a starting point and illustrate the process, publication data was used in the case of Kayser et al. (2014). In future works other data such as social media, reports, or news articles can be used. For example, to reinforce the customer perspective in roadmapping, users' views can be integrated at different stages of the process by using data which is closer to user's interests, such as newspapers or social media.

3.2.2. Comparing public and scientific discourse

As stated in the introduction, foresight also addresses societal change, but most of the used indicators rely on scientific and technological developments (e.g. patents, standards). Thus, new indicators are required to map other areas. With regard to text-based indicators, one option is the examination of news reporting, as described in the following example of mapping the state-of-the-art in society. News is a

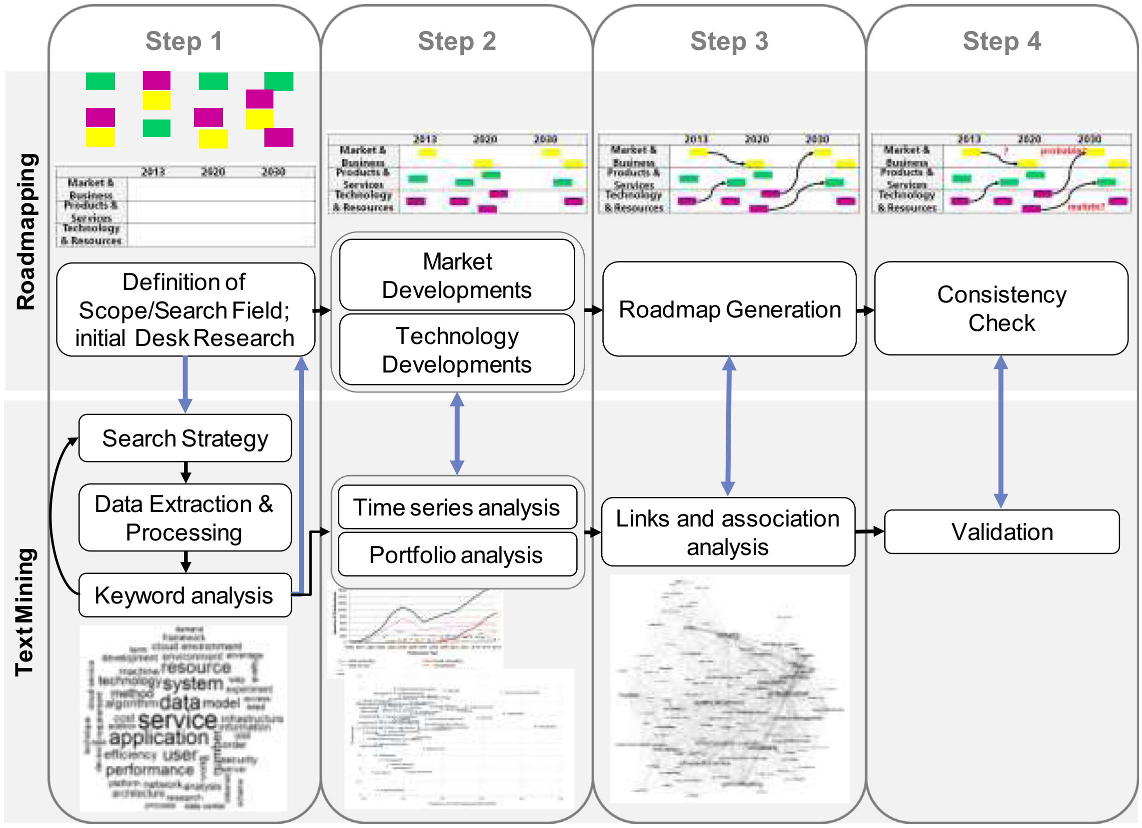

Fig. 3-1. Process model: roadmapping and text mining (adapted from Kayser et al., 2014).

recognized channel for innovation diffusion and plays an important role in keeping the society informed. The following example explores news articles and outlines with an approach that automatically compares scientific and public discourse (Kayser, 2016). To contrast the changes and developments in science and society, the link between the two is specifically addressed by comparing the content of news articles and scientific publications.

Compared to content analysis—the prevailing method for analyzing news (Krippendorff, 2013)—text mining has been used for the analysis to process the volume of textual data according to a common analytical scheme, reduce the manual effort in analyzing textual data, and conduct an automatic comparison of the texts.

Fig. 3-2 illustrates the results for the exemplary case of *vegan diet* (Kayser, 2016). This topic has attracted increasing societal interest in

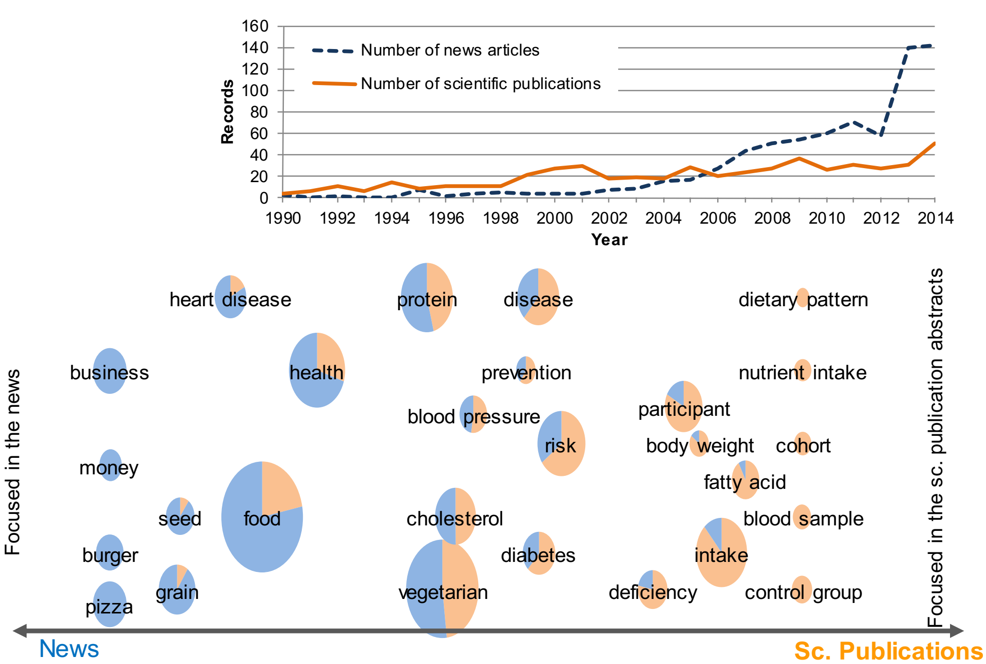

Fig. 3-2. Comparison of news reporting and scientific publications (topic: vegan diet) (Kayser, 2016).

Please cite this article as: Kayser, V., Blind, K., Extending the knowledge base of foresight: The contribution of text mining, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.017

the last 10 years. In contrast, the topic has been more constantly addressed in scientific discourse. The text analysis contrasts the terms extracted from the news and publication abstracts. While the size of the bubble relates to the average frequency of a term, the horizontal position and the shares indicate where it is predominantly addressed. In this example, science and the common public are talking about different things. While the news reports focus on lifestyle and cooking (e.g. food, burger, grain), the abstracts mostly cover medical and health issues (e.g. blood sample). However, there are also commonalities such as blood pressure or protein.

This case is more of an investigation for building experience in mapping society and automatically contrasting discourses than a fixed scheme of analysis. Numbers give an idea about the volume of text, but neglect the content. For foresight, the relevance of this approach is in the fast overview of thematic focus and parallels, the intensity of the debate, and an approximation of the current state of society. This enables the researcher to make inferences on technology acceptance and diffusion. In this case, it appears that science constantly addresses the health impact of vegan nutrition while the lifestyle aspect of vegan diet is addressed in the news, apart from medical issues especially. Therefore, the insights are a starting point for further analysis such as interviews or detailed desk research.

3.2.3. Scenario development

The following example describes an approach that combines scenario development and web-based information retrieval. This example addresses the increasing effort needed to observe evolving topics and introduces an approach that automates desk research to deliver a summary of a topic for the initial discussion in scenario development. Scenarios illustrate different future—each formulated as one scenario story. These scenarios serve as a framework to think about future challenges and developments that influence today's decisions (Reibnitz, 1991; van der Heijden, 2005). Among the many scenario approaches available at present, according to our knowledge not a single approach uses text mining or seeks more efficient ways to explore the scenario field, e.g. by automatic desk research. Principally, scenario development starts with desk research and the literature analysis for a comprehensive understanding of the topic, from which influence areas and future projections are derived to describe the scenario field. These are combined to create different scenario stories.

As described in Kayser and Shala (2017), first, a hashtag-based search on Twitter is used to delimit a thematic field (in this case *#quantifiedself*). After this, web mining is used to retrieve the websites mentioned in the collected tweets. By aggregating the content of the websites, this form of data retrieval briefly describes the scenario field and serves as a starting point for discussing possible futures. The results of text mining yield a comprehensive overview of the topic and summarize the scenario field. This is the starting point for an iterative process of discussing the text mining results and further desk research (e.g. figures, more details). Of course, a strategic alignment is necessary for adapting the results to the unit conducting the scenario process.

A key advantage of this method is that more content and data can be analyzed as compared to traditional literature analysis. Practically, the methods introduced here reduce the reading effort and thereby the time needed for desk research and literature analysis at the beginning of the scenario process. For example, more than 1000 websites are processed with web-based scenario development—much more that the maximum number that can be processed manually (Kayser and Shala, 2017).

4. Relevance of text mining for foresight

The starting point of this article was the observation that recent foresight methods are often based on literature analysis, patent and publication data, or expert opinions, but make little use of other data sources. In times of big data, many other options exist, in particular with respect to web content. A large volume of textual data is not considered systematically in foresight activities, analyzed in an automatic or comprehensive manner, or used together with other foresight methods such as roadmapping or scenario development. Therefore, this article shows several options for integrating and processing textual data in the context of foresight, and illustrates three examples.

4.1. Implications for the process of foresight

Concerning the process model of foresight introduced in Section 2, text mining contributes to all three stages of the process. As shown by the examples of roadmapping and future scenarios in Section 3, text mining can be a part of foresight or can be used in the process of foresight, as illustrated in Fig. 4-1.

The use of new data sources enables the researcher to address new aspects and research questions (see e.g., Boyd and Crawford, 2012; Kitchin, 2014). "*As such, the epistemological strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify potential questions* (*hypotheses*) *worthy of further examination and testing*" (Kitchin, 2014). This is also reflected in foresight practice, because different futures are constructed on the basis of assumptions derived from data about the state-of-the-art. In the case of explorative foresight approaches, it is essential to have a profound understanding of present developments when thinking about complex futures. A broad and comprehensive data basis summarizing the state-of-the-art acts as a starting point for further steps in the foresight process. Afterward, plausible future paths can be drawn, a solid base for decisions is provided, and more robust strategies can be derived.

Text mining broadens the information base as a starting point for foresight activities and helps in exploiting the steadily rising volume of (textual) data (e.g. Twitter, news articles, web mining). For example, news can be processed by content analysis (for an overview of content analysis, see Krippendorff, 2013) and Twitter data can be manually gathered. However, this takes more time and smaller quantities can be processed. Thereby, a larger number of opinions and statements can be analyzed. For instance, the views of more people are analyzed by using Twitter data than by conducting a small number of workshops. In addition, the automatic gathering of content with such a degree of variety and breadth is not possible with *classic* methods (e.g. interviews, workshops). Processing more and *new* data enables the integration of more views and stakeholder positions, and thereby extends the knowledge base of foresight.

Exploring and identifying relevant aspects in an objective manner is facilitated by text mining, as the example of roadmapping illustrates (Section 3.2.1). This structured summary of the state-of-the-art solves the problem getting input in other forms, such as workshops. The latter might be dominated by individual opinions, group dynamics, or people who want to push the discussion in a particular direction. In contrast, results from text mining are traceable and repeatable. However, they can have biases as well (e.g. very active interest groups on Twitter). Moreover, automated desk research reduces the effort and time needed for summarizing the considered field. Therefore, more content, opinions, and views can be processed and considered in the foresight exercise.

New techniques for data analysis, such as concept maps, topic models, the automatic comparison of datasets, and association analysis, complement discussions. For exploring the future, text mining highlights recent trends, contributes an external perspective, and looks for reflections. One of the main contributions of text mining to foresight is that foresight exercises can be built more precisely on the state-of-the-art, e.g., due to techniques such as web mining, for instance. This serves as a starting point for discussing possible futures to promote a creative discourse, in particular by hinting towards formerly disregarded aspects. In addition, text mining results may reflect or validate intermediate results of the ongoing foresight activity and contribute to the generation of future knowledge. New or further question can be derived from the results of the initial data analysis, as shown by the examples in Sections 3.2.2 and 3.2.3.

Please cite this article as: Kayser, V., Blind, K., Extending the knowledge base of foresight: The contribution of text mining, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.017

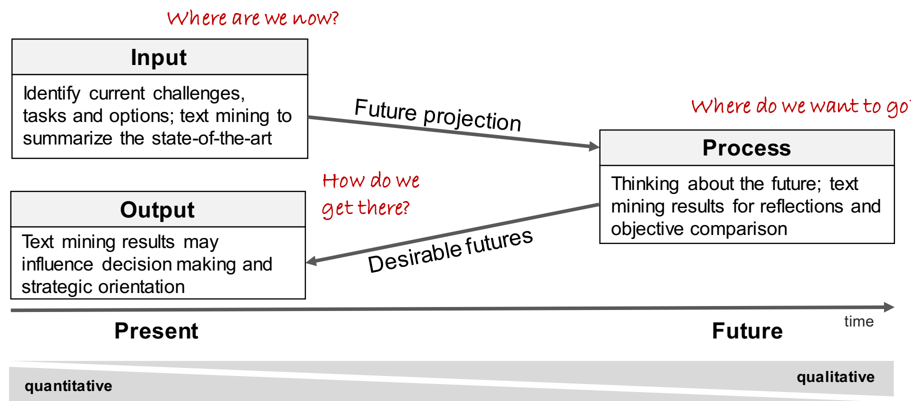

Fig. 4-1. Explorative foresight using text mining: balance of qualitative and quantitative thinking.

Finally, results from text mining are valuable for the quantification and underlining of statements and strategic choices to support decision-making.

However, also the limitations of using text mining together with foresight should be clear. Apart from the general restrictions of text mining as discussed in Section 2.2, the strength and limitations should be clarified if *new* data is used in foresight, for example in the case of Twitter data (Kayser and Bierwisch, 2016). Some results require further examination or further research effort. This is especially highlighted by the example in Section 3.2.2. The results deliver an initial idea and give an overview. The manual analysis and objective comparison of such a high number of texts would not have been possible without text mining. However, these results trigger a discussion and can generate hypotheses that need to be proved and validated through further investigation.

An elaboration of the role of data in foresight shows that foresight only functions as a combination of qualitative and quantitative thinking. The future cannot be derived from only data, because the farther we reach into the future, the less accurate this data gets (Amer et al., 2013; van der Heijden, 2005). Interpretation and alignment with the specific requirements is necessary because "*the quantitatively accessible nature of things declines steadily as we gaze further into the future*" (Pillkahn, 2008). This also relates to the accuracy and precision of models and simulations, which depend on the assumptions made today. Anyway, foresight is less about precision and accuracy and more about considering the different futures and the options we have today. This is accomplished by qualitative methods for the long-term view, which additionally strengthens the collaborative character of foresight. Ultimately, this means that the farther we reach into the future, the more qualitative the nature of foresight becomes, as highlighted in Fig. 4-1. Finally, qualitative and quantitative approaches complement each other and therefore should be combined (see e.g., Amer et al., 2013).

4.2. Future research directions

For designing foresight applications and methods, the added value of text mining is that data sources that have not been used so far can be accessed and more data can be processed and better analyzed. Foresight has a broad spectrum ranging from the micro to the macro perspective, and, as this article shows, questions at different levels of foresight can be addressed with text mining. This extends from examining systemic links and the function of innovation systems to enhancing the dynamic capabilities of firms (Kayser et al., 2014). Therefore, different combinations of data sources, text mining approaches, foresight process requirements (e.g. scope, time frame, topic), and foresight methods yield a wide range of options for future work. These four building blocks can be interchangeably combined to extend the spectrum of foresight methods. Thereby, the modular character of foresight encompasses further components and many new applications at the corporate, organizational, societal, or political levels can be designed in future work.

This article provides, in particular, a conceptual outline of how to improve foresight by using text mining. The examples outlined in this article are proof of concepts on the use of text mining in foresight and working with new data sources. A final assessment of the contribution of text mining to foresight is beyond the scope of this work, especially with regard to data sources such as web mining or social media data. So many options exist for future research.

Evaluation of the methods as proposed in this article as well as the data sources should be among the next steps. In particular, these methods should be implemented in (larger) foresight projects and exercises. In future work, foresight methods, textual data sources, and text mining approaches can be used for designing new applications such as real-time foresight applications or web-based roadmapping.

Solutions such as text mining that deal with increasing volumes of data are necessary and very pertinent in our present time for exploiting these information sources and gaining relevant insights. *Big data* offers potentials like data source and for the research process per se (see e.g., Boyd and Crawford, 2012). Approaches like text mining provide access to this breadth and variety of content from different sources. In future work, platforms such as YouTube, Instagram, or Pinterest could be examined for their contribution to foresight, especially when used together with trend analysis and weak signal detection. However, the strength and weaknesses of this data should be evaluated first.

Another issue is getting more experience in processing web data and social media. Techniques such as web mining enable quick access to data, in contrast to previous applications in the field of foresight. Of course, appropriate mechanisms are necessary for selecting and processing data. The real-time access to data offers many new options, especially in the context of trend recognition and early detection of change. One question remains: In light of the fast-changing world and disruptive change and innovation, is there a need for real-time data in foresight?

Apart from technological change, foresight addresses social changes. However, data sources like social media or user-generated content are rarely considered in foresight for capturing societal points of view. Using text mining, a larger set of views and opinions can be analyzed for generating a quick overview of the issues discussed, as shown in the example of Twitter data (Kayser and Bierwisch, 2016) or news articles (Kayser, 2016). This enables new forms of stakeholder integration and engagement in foresight processes. Therefore, more research effort should be spent here.

This article illustrates different ways of accessing and aggregating today's volumes of data and how to make use of it in foresight. In future work, text mining should be related to further foresight methods to test the benefits and weigh the qualitative and quantitative methods. Finally, text mining is most valuable for extending the information and knowledge base of foresight; it expands the range of foresight methods and improves the mix of methods applied in foresight. Thereby, new options for gaining knowledge from data are evolving. The topic is very relevant for different areas and, principally, strategies to generate value from data are very much the need of the hour.

References

Abbas, A., Zhang, L., Khan, S.U., 2014. A literature review on the state-of-the-art in patent analysis. World Patent Inf. 37, 3–13.
Albert, T., Möhrle, M.G., Meyer, S., 2015. Technology maturity assessment based on blog analysis. Technol. Forecast. Soc. Chang. 92, 196–209.
Amanatidou, E., Butter, M., Carabias, V., Könnölä, T., et al., 2012. On concepts and methods in horizon scanning: lessons from initiating policy dialogues on emerging issues. Sci. Public Policy 39 (2), 208–221.
Amer, M., Daim, T.U., Jetter, A., 2013. A review of scenario planning. Futures 46, 23–40.
Barker, D., Smith, D.J.H., 1995. Technology foresight using roadmaps. Long Range Plan. 28 (2), 21–28.
Berry, M.W., 2004. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, New York.
Bonino, D., Ciaramella, A., Corno, F., et al., 2010. World Patent Inf. 32 (1), 30–38.
Bornmann, L., Mutz, R., 2014. Growth rates of modern science: a bibliometric analysis. *CoRR* abs/1402.4578.
Boyd, D., Crawford, K., 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15 (5), 662–679.
Burkart, R., 2002. Kommunikationswissenschaft: Grundlagen und Problemfelder. 4th ed. Böhlau, Wien [et al.].
Cachia, R., Compañó, R., Da Costa, O., 2007. Grasping the potential of online social networks for foresight. Technol. Forecast. Soc. Chang. 74 (8), 1179–1203.
Choi, S., Park, H., Kang, D., Lee, J.Y., Kim, K., et al., 2012. Expert Syst. Appl. 39 (13), 11443–11455.
Choi, S., Kim, H., Yoon, J., Kim, K., Lee, J.Y., et al., 2013. R&D Manag. 43 (1), 52–74.
Cong, H., Loh, H.T., 2010. Pattern-oriented associative rule-based patent classification. Expert Syst. Appl. 37 (3), 2395–2404.
Cuhls, K., 2003. From forecasting to foresight processes—new participative foresight activities in Germany. J. Forecast. 22 (2–3), 93–111.
Cunningham, S.W., Porter, A.L., Newman, N.C., 2006. Special issue on tech mining: tech mining: exploiting science and technology information resources. Technol. Forecast. Soc. Chang. 73 (8), 915–922.
Da Costa, O., Warnke, P., Cagnin, C., Scapolo, F., 2008. The impact of foresight on policy-making: insights from the FORLEARN mutual learning process. Tech. Anal. Strat. Manag. 20 (3), 369–387.
de Miranda Santo, M., Coelho, G.M., dos Santos, D.M., Filho, L.F., 2006. Text mining as a valuable tool in foresight exercises: a study on nanotechnology. Technol. Forecast. Soc. Chang. 73 (8), 1013–1027.
de Smedt, P., 2013. Interactions between foresight and decision-making. In: Borch, K., Dingli, S.M., Jørgensen, M.S. (Eds.), Participation and Interaction in Foresight: Dialogue, Dissemination and Visions. Edward Elgar Publishing, Incorporated.
Delen, D., Crossland, M.D., 2008. Seeding the survey and analysis of research literature with text mining. Expert Syst. Appl. 34 (3), 1707–1720.
Eerola, A., Miles, I., 2011. Methods and tools contributing to FTA: a knowledge-based perspective. Futures 43 (3), 265–278.
Feldman, R., Sanger, J., 2008. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge, New York.
Glassey, O., 2012. Folksonomies: spontaneous crowd sourcing with online early detection potential? Futures 44 (3), 257–264.
Glenisson, P., Glänzel, W., Janssens, F., de Moor, B., 2005. Combining full text and bibliometric information in mapping scientific disciplines. Inf. Process. Manag. 41 (6), 1548–1572.
Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102 (1), 653–671.
Grimmer, J., Stewart, B.M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 1–31.
Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques. 3rd ed. Elsevier/Morgan Kaufmann, Amsterdam, Boston.
Havas, A., Schartinger, D., Weber, M., 2010. The impact of foresight on innovation policy-making: recent experiences and future perspectives. Res. Eval. 19 (2), 91–104.
Horton, A., 1999. A simple guide to successful foresight. Foresight 1 (1), 5–9.
Huang, L., Zhang, Y., Guo, Y., Zhu, D., Porter, A.L., 2014. Four dimensional science and technology planning: a new approach based on bibliometrics and technology roadmapping. Technol. Forecast. Soc. Chang. 81 (0), 39–48.
Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of social media. Bus. Horiz. 53 (1), 59–68.
Kayser, V., 2016. Comparing public and scientific discourse in the context of innovation systems. Technol. Forecast. Soc. Chang.
Kayser, V., Bierwisch, A., 2016. Using Twitter for Foresight: An Opportunity? Futures.
Kayser, V., Shala, E., 2016. Generating futures from text—scenario development using text mining. In: Daim, T.U., Chiavetta, D., Porter, A.L., Saritas, O. (Eds.), Anticipating Future Innovation Pathways Through Large Data Analysis. Springer International Publishing, Cham, s.l., pp. 229–245.
Kayser, V., Shala, E., 2017. Web-based Scenario Development: Process Improvements. (forthcoming).
Kayser, V., Goluchowicz, K., Bierwisch, A., 2014. Text mining for technology roadmapping: the strategic value of information. Int. J. Innov. Manag. 18 (03), 1440004.
Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S., 2011. Social media?: Get serious! Understanding the functional building blocks of social media. Bus. Horiz. 54 (3), 241–251.
Kitchin, R., 2014. Big data, new epistemologies and paradigm shifts. Big Data Soc. 1 (1), 1–12.
Könnölä, T., Scapolo, F., Desruelle, P., Mu, R., 2011. Foresight tackling societal challenges: impacts and implications on policy-making. Futures 43 (3), 252–264.
Kostoff, R.N., 2012. Text mining for science and technology - a review part I – characterization/scientometrics. Scientometrics 1 (1), 11–21.
Krippendorff, K., 2013. Content Analysis: An Introduction to Its Methodology. 3rd ed. SAGE, Los Angeles, London.
Lee, S., Lee, S., Seol, H., Park, Y., 2008. Using patent information for designing new product and technology: keyword based technology roadmapping. R&D Manag. 38 (2), 169–188.
Lee, C., Song, B., Park, Y., 2013. How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships. Tech. Anal. Strat. Manag. 25 (1), 23–38.
Manning, C.D., Raghavan, P., Schütze, H., 2009. An Introduction to Information Retrieval. Cambridge University Press, New York.
Martin, B.R., 1995. Foresight in science and technology. Tech. Anal. Strat. Manag. 7 (2), 139–168.
Masiakowski, P., Wang, S., 2013. Integration of software tools in patent analysis. World Patent Inf. 35 (2), 97–104.
Möhrle, M.G., Isenmann, R., Phaal, R. (Eds.), 2013. Technology Roadmapping for Strategy and Innovation: Charting the Route to Success. Springer, Berlin [et al.].
O'Brien, F.A., Meadows, M., 2013. Scenario orientation and use to support strategy development. Scenario Method Curr. Dev. Theory Pract. 80 (4), 643–656.
Ortner, H., Pfurtscheller, D., Rizzolli, M., Wiesinger, A., 2014. Zur Einführung – Datenflut und Informationskanäle. In: Ortner, H., Pfurtscheller, D., Rizzolli, M., Wiesinger, A. (Eds.), Datenflut und Informationskanäle, 1st ed. Innsbruck Univ. Press, Innsbruck.
Pang, A.S.-K., 2010. Social scanning: improving futures through Web 2.0; or, finally a use for twitter. Glob. Mindset Chang. 42 (10), 1222–1230.
Park, H., Yoon, J., Kim, K., 2012. Identifying patent infringement using SAO based semantic technological similarities. Scientometrics 90 (2), 515–529.
Park, H., Kim, K., Choi, S., Yoon, J., 2013a. A patent intelligence system for strategic technology planning. Expert Syst. Appl. 40 (7), 2373–2390.
Park, H., Ree, J.J., Kim, K., 2013b. Identification of promising patents for technology transfers using TRIZ evolution trends. Expert Syst. Appl. 40 (2), 736–743.
Phaal, R., Farrukh, C., Probert, D.R., 2010. Roadmapping for Strategy and Innovation: Aligning Technology and Markets in a Dynamic World. Cambridge. University of Cambridge, Institute for Manufacturing.
Pillkahn, U., 2008. Using Trends and Scenarios as Tools for Strategy Development. Wiley-VCH, Hoboken.
Popper, R., Butter, M., 2008. How are foresight methods selected? Foresight 10 (6), 62–89.
Reibnitz, U., 1991. Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung. Gabler, Wiesbaden.
Slaughter, R., 1995. The foresight Principle: Cultural Recovery in the 21st Century. Adamantine Press; Adamantine, London, England.
Tseng, Y.-H., Lin, C.-J., Lin, Y.-I., 2007. Text mining techniques for patent analysis. Inf. Process. Manag. 43 (5), 1216–1247.
van der Heijden, K., 2005. Scenarios: The Art of Strategic Conversation. 2nd ed. John Wiley & Sons, Chichester, West Sussex, Hoboken, N.J.
van Eck, N.J., Waltman, L., 2011. Text mining and visualization using VOSviewer. arXiv Preprint arXiv:1109.2058.
van Eck, N.J., Waltman, L., Noyons, E.C.M., Buter, R.K., 2010. Automatic term identification for bibliometric mapping. Scientometrics 82 (3), 581–596.
Voros, J., 2003. A generic foresight process framework. Foresight 5 (3), 10–21.
Wang, M.-Y., Chang, D.-S., Kao, C.-H., 2010. Identifying technology trends for R&D planning using TRIZ and text mining. R&D Manag. 40 (5), 491–509.
Weiss, S.M., 2010. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer, New York.
Yau, C.-K., Porter, A., Newman, N.C., Suominen, A., Newman, N., 2014. Clustering scientific documents with topic modeling. Scientometrics 100 (3), 767–786.
Yoon, J., 2012. Detecting weak signals for long-term business opportunities using text mining of Web news. Expert Syst. Appl. 39 (16), 12543–12550.
Yoon, B., Phaal, R., Probert, D.R., 2008. Structuring technological information for technology roadmapping: data mining approach. Proceedings of the 7th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases: World Scientific and Engineering Academy and Society. WSEAS Press, pp. 417–422.
Yoon, J., Park, H., Kim, K., 2013. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. Scientometrics 94 (1), 313–331.
Youtie, J., Hicks, D., Shapira, P., Horsley, T., 2012. Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. Tech. Anal. Strat. Manag. 24 (10), 981–995.

Please cite this article as: Kayser, V., Blind, K., Extending the knowledge base of foresight: The contribution of text mining, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.017